# Phylogenetic analysis workflow using the BioExtract Server

**Yosr Bouhlal[1], Douglas M Jennewein[1] and Carol Lushbough[1]**

[1]Computer Science Department, University of South Dakota, Vermillion SD ,USA

**Abstract -** *Several molecular and genetic tools, browsers and servers are currently available online for biologists to access, analyze and process data. The BioExtract Server represents a powerful web-based data integration application, combining the most common biological databases with the most used algorithms by scientists from all the biological fields. This Server allows researchers to extract data, execute local and web-accessible analytic tools and create customized workflows. Each workflow can be used for different similar queries and offers an easy access to the results of all the executed tools at once. This paper describes a BioExtract workflow providing a simple phylogenic analysis, as one of the numerous applications that the BioExtract Server offers to biologist researchers.*

**Keywords:** BioExtract Server, workflow, phylogenetic analysis

## 1 Introduction

The study of genome evolution involves a global comparative approach in which individual genetic events are considered and integrated in their evolutionary context, which in turn may be correlated to the population history, the environment and the different phonemes [1]. Many tools and techniques are currently used to study evolution and infer the evolutionary relationship between species and organisms. These techniques include morphology, anatomy, paleontology, physiology and molecular phylogeny [2].

Phylogeny based analysis provides an ideal framework for performing such investigations, by pinpointing when a genetic event occurred and by identifying the simultaneous occurrence of several events [1]. There are principally five stages in the molecular phylogenic analysis [2]. The first stage is the acquisition of the sequence which can be performed through many sources including Genbank or HomoloGene gene databases, Rfam for RNA, Pfam for proteins or ICTV for viruses. Once sequences are acquired, a multiple sequence alignment will be performed on homologous sequences. This stage is considered a critical step of phylogenetic analysis subject to many important considerations. The next stages will be the specification of a statistical model of nucleotide or amino acid evolution, the construction of the evolution tree, and finally the interpretation of the generated tree [2]. Among an important number of online tools and servers, the BioExtract Server (http://bioextract.org) is a powerful Web-based data integration application that can be used to help researchers accomplish all these phylogenetic analysis steps. The BioExtract Server was designed to help scientists consolidate, analyze, and serve data from heterogeneous bio-molecular databases [3]. It allows them to query multiple data sources, save query results as searchable data sets, execute local and Web-accessible analytic tools, and create computational customized workflows [3, 4].

We describe here a simple BioExtract Server workflow that can be used for standard phylogenetic analysis starting from a protein sequence query.

## 2 Methods

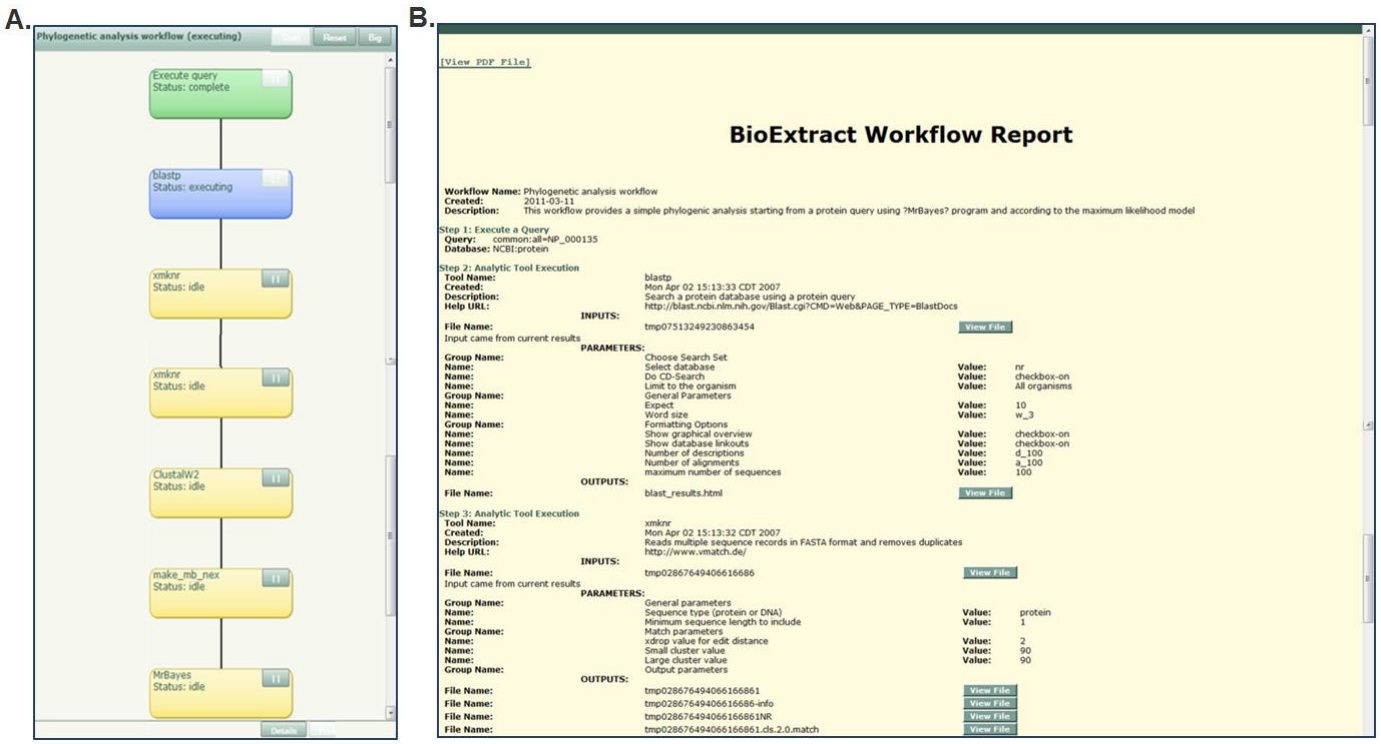
The BioExtract Server was used to create a workflow for comparing and aligning a number of nucleotide sequences to build a phylogenetic evolutionary tree (Figure 1A). The workflow covers five steps using five different molecular online tools (Table 1).

**Table1**. Tools used in the phylogenic workflow

| Tool | Common Link | Description | Ref. |
|---|---|---|---|
| Blastp | http://blast.ncbi.nlm.nih.gov/Blast.cgi | Search protein database using a protein query | [5] |
| xmknr | http://vmatch.de | Reads multiple sequence records in FASTA format and removes duplicates | [6] |
| ClustalW | http://clustal.org | Computes a multiple sequence alignment for Protein or DNA sequences | [7] |
| make_mb_nex | http://bioextract.org/ | Creates a MrBayes nexus file from a clustal alignment file | |
| MrBayes | http://mrbayes.csit.fsu.edu/ | Estimate phylogeny upon Bayesian inference which is based on the probability of a tree conditioned on the observations. | [8] |

The query sequence can be selected from the common or specific databases available through the BioExtract Server or simply uploaded from a private source. When executing the workflow, similar sequences will be extracted by the "Blastp" tool. Duplicate sequences will be then removed using "xmknr" tool, a simple shell script utilizing the Vmatch tool. Users can further refine the Blastp results by selecting sequences according to the length or the E score through the "extract page" before running the next step. Once selected, sequences will be aligned using the "ClustalW" multiple sequences alignment program.

**Figure 1.** (A) BioExtract Server workflow created for the phylogenic analysis (B) The first three steps of the workflow showed on the general report




In order to perform the phylogenic analysis for the remaining aligned sequences, we developed a new tool "make_mb_nex" and included it within the BioExtract Server tools page.

This tool will create a nexus file from an alignment file. The user can configure the created nexus file by specifying the appropriate evolutionary model and the MCMC (Markov chain Monte Carlo) algorithm parameters. For this study, Following parameter settings were used: set nst=6 rates=invgamma [according to the General Time Model GTM]; mcmc ngen=1000; samplefreq=10; sump burnin=25 and sumt burnin=25. Finally, the generated nexus file is executed on the "MrBayes" program according to the maximum likelihood model [9] and the evolutionary tree is drawn.

In order to test the feasibility and usefulness of this workflow, the human Frataxin protein sequence (variant 1: NP_000135) was used as the initial query.

# 3 Results

Human Frataxin protein is a mitochondrial protein encoded by the FXN gene and seems to be implicated in the iron-sulfur clusters. Reduced or modified frataxin causes Freidreich's ataxia, an autosomal recessive neurodegenerative disorder. Alternative splicing results in multiple transcripts variants [10]. The variant 1 [NP_000135] was used as an input query to run the BioExtract Server phylogenetic analysis workflow.

The execution of the workflow led to the extraction of 100 sequences homologous to the frataxin protein. After duplicates were excluded, multiple sequence alignments are performed for all the sequences. Once poorly aligned sequences are removed, the corresponding phylogenetic tree is estimated using a Bayesian method based on a general time reversible (GTR) model.

All the results are summarized in a general report (Figure 1B). Each output can be visualized or uploaded by clicking on "View File" corresponding to each tool. As an example of the workflow output, the resulting phylogenic tree and the credibility values are shown on Figure 2.

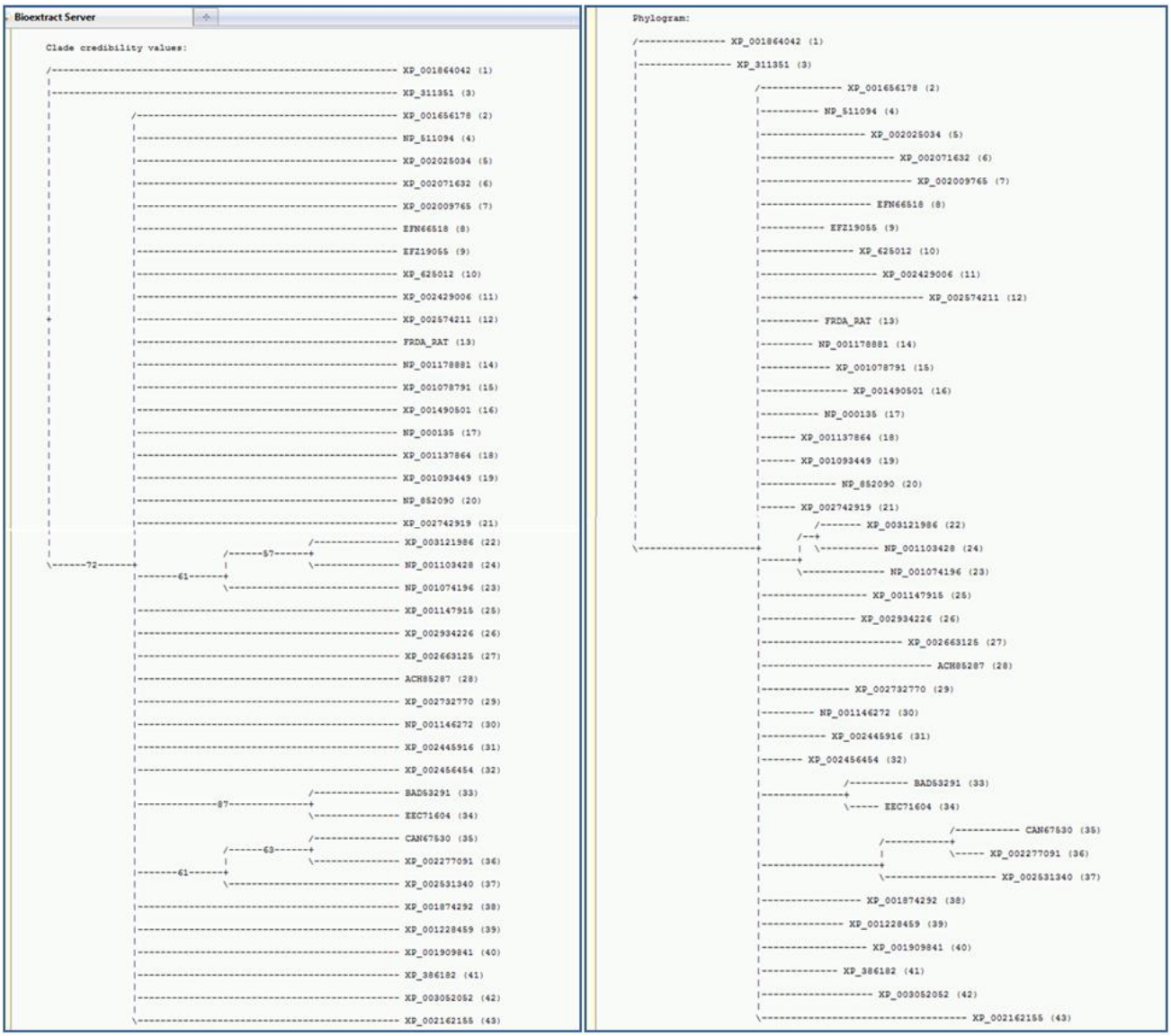
This workflow is shared on the "MyExperiment" portal and is accessible through the following link http://www.myexperiment.org/workflows/1941.html.

# 4 Discussion

The BioExtract Server is a Web-based system designed to aid researchers in the analysis of distributed genomic data by providing a platform to facilitate the creation of bioinformatic workflows [4]. The basic operations of the BioExtract server allow researchers via their Web browsers to: specify data sources; flexibly query data sources with a range of relational operators; apply analytic tools; download result sets; and store query results for later reuse. As the researcher works with the system, their "steps" are saved in the background. At any time these steps can be saved as a workflow simply by providing a name and description. Once saved, these workflows can be executed and/or modified [3]. The execution of any created workflow generates the running of all the tools at once, and provides access to all the results via the general workflow report. Consequently, the results are obtained in an extremely reduced time compared to conventional methods. In addition, the results are recorded in the workflow and can be easily retrieved from the server when needed.

The workflow presented in this paper provides a simple phylogenic analysis starting from a protein query. Users can

**Figure 2.** Clade credibility values and Phylogram




modify the query by simply changing the accession number on the workflow's query step. Similar workflows can be created to analyse DNA or RNA sequences by modifying the query database on the first step and replacing Blastp with Blastn in the second step of the workflow. The two first steps of the workflow can also be eliminated if the user needs to directly upload his or her own aligned sequences. Several other enhancements can be added to the phylogeny analysis workflow by adding additional phylogenic tools and packages available through the BioExtract Server such as "PAUP", "dnadist", "propars" and "MrModelist".

This workflow represents one of numerous applications that the BioExtract Server offers to biologist researchers. Containing a large cluster of tools and giving access to numerous databases, the BioExtract Server can be used for genomic and protein annotation, sequence mutation analysis, gene or protein function prediction and many other complex molecular and genetic analyses. Some of these applications are actually shared on the "MyExperiment" website [http://www.myexperiment.org/] where they can be easily launched and used.

Various enhancements to the BioExtract Server are under development that, when added, will broaden the spectrum of users by adding more tools and databases which could be used for many additional biological fields.

# 5 References


[1] Gouret P, Thompson JD, Pontarotti P. PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinformatics. 2009 Sep 19; 10:298.

[2] Jonathan Pevsner. Bioinformatics and Functional Genomics (second edition). Wiley-Blackwell 2009: 215-69.

[3] Lushbough C, Bergman MK, Lawrence CJ, Jennewein D, Brendel V. BioExtract server--an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data. IEEE/ACM Trans Comput Biol Bioinform. 2010 Jan-Mar; 7(1):12-24.

[4] Lushbough CM, Brendel VP. An overview of the BioExtract Server: a distributed, Web-based system for genomic analysis. Adv Exp Med Biol. 2010; 680:361-9.

[5] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol., 1990; 215(3):403-10.

[6] Abouelhoda MI, Kurtz S, Ohlebusch E. Replacing suffix trees with enhanced suffix arrays. J Discrete Algorithms 2004; 2: 53–86.

[7] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res., 1994; 22[22]:4673-80.

[8] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001 Aug; 17(8):754-5.

[9] Evolutionary trees from DNA sequences: a maximum likelihood approach. Felsenstein J. J Mol Evol. 1981; 17(6):368-76.

[10] Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F et al. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science. 1996 Mar 8; 271(5254):1423-7.