Key2Ann: a tool to process sequence sets by replacing database identifiers with a human-readable annotation

Andreas Pürzer[1], Felix Grassmann[2], Dietmar Birzer[2] and Rainer Merkl[2*]

[1]University of Applied Sciences, Department of Computer Science and Mathematics,
93025 Regensburg

[2]Institute of Biophysics and Physical Biochemistry, University of Regensburg,
93040 Regensburg

Summary

Deducing common properties or degrees of phylogenetic relationship by analyzing a grouping or clustering of sequence sets is a frequently used technique in computational biology. If interpreted by means of visual inspection, the conclusions depend for many of these applications on meaningful names for the input data. In accordance with the aim of the analysis, the sequences should be provided with names indicating the function of the genes or gene-products, the phylogenetic position or other properties characterizing the contributing species. However, sequences extracted from databases are most often annotated with identifiers which only implicitly contain the desired information. To solve this problem, we have designed and implemented a tool named Key2Ann, which replaces in multiple fasta files the database keys with short terms indicating the taxonomic position or other features like the gene name or the EC-number. In addition, properties like habitat, growth temperature or the degree of pathogenicity can be coded for microbial species. To allow for highest flexibility, the user can control the composition of the names by means of command line parameters. Key2Ann is written in Java and can be downloaded *via* http://www-bioinf.uni-regensburg.de/downl/Key2Ann.zip. We demonstrate the usage of Key2Ann by discussing three typical examples of phylogenetic analysis.

1 Introduction

Many algorithms of computational biology are based on concepts of machine learning. For example neural networks [1, 2] or support vector machines [3] have been used successfully. However, if training is not possible due to the experimental setup or the lack of data, other methods have to be utilized to group or classify objects. Prominent examples are cluster techniques [4] used for the analysis of gene expression levels or algorithms generating trees to study phylogenetic relationship [5, 6]. In these cases, an important task of each analysis is to deduce common features of those elements constituting subgroups or subtrees. Using information from a taxonomy browser [7] or from a classification scheme for gene or protein functions like COG [8], eggNOG [9], KEGG [10] or FunCat [11], as well as a controlled vocabulary like Gene Ontology [12] are ideal approaches for a detailed analysis *via* computational methods. However, these concepts are less adequate to label leaves in a phylogenetic tree or the elements of subclusters if the resulting graphs have to be interpreted by means of visual inspection, which is a common approach. This is especially true, if large data sets have to be analyzed and if for example the phylogenetic lineage of the samples as well as other parameters have to be considered and compared. In these cases, short names that indicate the origin or the function of the sequences or that characterize e. g. the habitat of the

[*] To whom correspondence should be addressed. Email: rainer.merkl@biologie.uni-r.de

doi:10.2390/biecoll-jib-2011-153

species contributing the sequences are more adequate. The usage of such labels containing customized information makes possible or at least alleviates the identification of common features in subgroups or subtrees.

Having this problem in mind, we designed and implemented the software tool Key2Ann. It replaces automatically database keys or identifiers with annotations, i.e. tailored names which can be interpreted easily.

2 Compiling sequence labels

The scope of Key2Ann is the processing of sequence sets by replacing identifiers with “telling” names. It was our aim to implement a tool allowing the annotation of a broad range of sets consisting of DNA or protein sequences. We decided to resolve GI-numbers and keys from the SwissProt [13] database. Thus, hits obtained by means of the BLAST server from NCBI databases [7] and entries deduced e. g. from the Pfam [14] or UniProt database [13] can be processed. Key2Ann reads input files in multiple fasta format and generates output files in the same format; their names have to be given after the options `–i` and `–o`. The composition of the labels to be generated is controlled by means of additional command line arguments. In the following, we describe these arguments and their effects; implementation and other details of usage are explained below in section Implementation. The following properties may be encoded in the labels:

(1) *Phylogenetic position:* to indicate the taxonomic position of a species, the arguments `superkingdom, phylum, class, order, family` and `genus` can be used. The first character of any label gives the superkingdom according to “A” (Archaea), “B” (Bacteria), “E” (Eukaryotes), “V” (Viruses/Viroids), “U” (Unclassified), or “O” (Other sequences). All further taxa are given by a combination of two unique characters.

(2) *Molecular function:* using the command line arguments `gene` or `ec_number` adds the gene name or the EC-number to the labels.

(3) *Features of microbial species:* when using the options `habitat`, `temprange`, `salinity`, `pathogenicity` or `gc_content`, short infixes indicating the respective features are added for genes/gene products of microbial species.

Thus, the program can e. g. be started with the options `–i input.fa –o output.fa superkingdom phylum habitat`. Per default, the parameters `superkingdom phylum` are used. In all cases, the species name is added at the end of each label using three characters. If two resulting labels are identical, numbers are appended to make them unique. Table 1 lists all command line options steering label composition and the tokens used to assemble names. Table 2 shows some examples of database keys and resulting labels. Using the options indicated in the first line of Table 2, these keys are converted to the labels given in the respective columns. The user can resolve the phylogenetic position of a species down to the granularity needed for the considered analysis. The usefulness of this concept can be studied by comparing rows # 5 – 8 of Table 2 which list the resulting names for sequences from closely and distantly related Firmicutes. The hierarchical mapping of taxonomical information onto the resulting names has a further useful effect: A lexicographical sorting groups the sequences according to their taxonomical position, which helps to organize and process the data sets. This is useful when e. g. working with an editor for multiple sequence alignments like Jalview [15] .

Tab. 1: A complete listing of command line options and infixes used to create Key2Ann labels. Column 1 lists the command line option. Column 2 gives the resulting infixes which are concatenated to the emerging name. Column 3 lists the meaning of the infixes. Note that the features habitat, salinity, temprange, pathogenicity and gc_content are processed for prokaryotic species only, due to the limited content of the respective data set.

Command line option	Token in new name	Meaning
`superkingdom`	A	Archaea
	B	Bacteria
	E	Eukaryota
	V	Viruses/Viroids
	U	Unclassified
	O	Other sequences
`phylum`	the first two characters of the phylum	
`class`	the first two characters of the class	
`order`	the first two characters of the order	
`family`	the first two characters of the family	
`genus`	the first two characters of the genus	
`gene`	the name of the gene the sequence is coding for	
`ec_number`	the EC-number if this sequence is an enzyme	
Additional annotation available for prokaryotes		
`habitat`	Ha	host-associated
	Sp	specialized
	Te	terrestrial
	Mu	multiple
	Aq	aquatic
`salinity`	Mh	moderate halophilic
	Eh	extreme halophilic
	Nh	non-halophilic
	Mh	mesophilic
`temprange`	Ht	hyperthermophilic
	Pt	psychrophilic
	Tt	thermophilic
	Ct	cryophilic
	Mt	mesophilic
`pathogenicity`	Hp	human pathogen
	Pa	pathogen
	Np	no pathogen
`gc_content`	GC-content in percent	

Tab. 2: Labels generated by Key2Ann. Column 1 lists the number of the respective key. The database keys are from DNA (#6) and from protein sequences (all others). Entry #2 is from the Pfam database, #4 and #10 are from SwissProt, all other ones are from NCBI databases. Some keys are abbreviated versions, indicated by “…”. The following columns give the output generated by different combinations of command line arguments. Arguments are abbreviated according to `sk = superkingdom`, `ph = phylum`, `cl = class`, `or = order`, `fa = family`; `tr = temprange` and `gn = gene` indicate two options adding the temperature range of the habitat or the gene name. The prefix for the phylogenetic lineage is supplemented by a three letter infix deduced from the name of the species. E. g. “Eco” is for *Escherichia coli* and “Hsa” for *Homo sapiens*. The infixes “Mt” and “Ht” indicate mesophilic and hyperthermophilic species. Properties that cannot be resolved are indicated by “__“.

#	**Database key**	sk ph	sk ph cl	sk ph cl or	sk ph fa tr	sk ph gn
1	gi|89108846|ref|AP_002626.1| [Escherichia coli str. K-12 substr. W3110]	BPr_Eco	BPrGa_Eco	BPrGaEn_Eco	BPrEn_Eco_Mt	BPr_Eco_hisF
2	HIS4_ECOSM/1-234	BPr_Eco_1	BPrGa_Eco_1	BPrGaEn_Eco_1	BPrEn_Eco_Mt_1	BPr_Eco_hisA
3	tr|B8D8Y0|B8D8Y0_BUCA5 OS=Buchnera aphidicola subsp. Acyrtho..	BPr_Bap	BPrGa_Bap	BPrGaEn_Bap	BPrEn_Bap_Mt	BPr_Bap_gyrA
4	sp|P32914|SPT4_YEAST OS=Saccharomyces cerevisiae	EAs_Sce	EAsSa_Sce	EAsSaSa_Sce	EAsSa_Sce___	EAs_Sce_SPT4
5	gi|16077181|ref|NP_387994.1| elongation factor Tu [Bacillus subtilis …]	BFi_Bsu	BFiBa_Bsu	BFiBaBa_Bsu	BFiBa_Bsu_Mt	BFi_Bsu_tufA
6	gi|21952760:89-1324 Bacillus circulans bglM gene complete cds	BFi_Bci	BFiBa_Bci	BFiBaBa_Bci	BFiBa_Bci___	BFi_Bci_bglM
7	gi|58255243|gb|AAV43480.1| elongation factor [Lactobacillus acidophilus]	BFi_Lac	BFiBa_Lac	BFiBaLa_Lac	BFiLa_Lac_Mt	BFi_Lac___
8	gi|160934405|ref|ZP_02081792.1| hypo. Prot CLOLEP_03278 [Clostr….]	BFi_Cle	BFiCl_Cle	BFiClCl_Cle	BFiCl_Cle_Mt	BFi_Cle___
9	gi|51094833|gb|EAL24079.1| euk. trans. elon. factor 1 [Homo sapiens]…	ECh_Hsa	EChMa_Hsa	EChMaPr_Hsa	EChHo_Hsa___	ECh_Hsa_LOC3932793
10	sp|Q9Y8T7|TRPC_AERPE Indole… OS=Aeropyrum pernix GN=trpC	ACr_Ape	ACrTh_Ape	ACrThDe_Ape	ACrDe_Ape_Ht	ACr_Ape_trpC

3 Exploring the distribution of orthologous proteins

In the following, we exemplify the usage of Key2Ann by means of three typical applications. The first study is related to the protein TrpB, the beta subunit of the tryptophan synthase [16], which catalyses the final reaction of tryptophan biosynthesis. To study sequence composition of homologs from Archaea and Bacteria, a phylogenetic tree has to be computed. If BLAST at the NCBI is used to collect protein sequences, the entries of the set used for tree construction are labeled with strings like “>gi|15897776|ref|NP_342381.1| tryptophan synthase subunit beta [Sulfolobus solfataricus P2]“. This line does not contain any information about the phylogenetic lineage of the species *Sulfolobus solfataricus*. Therefore, without a further processing of the identifiers it would be hard or perhaps impossible in larger data sets to deduce from the resulting tree patterns indicating e. g. a clustering of phylogenetic groups. As TrpB occurs in hyperthermophilic, thermophilic and mesophilic species, it might be that the growth temperature affects the distribution of the variants, too. This is why growth temperature of all species contributing sequences to the analysis has to be considered when analyzing the tree topology. Using Key2Ann, the user can create labels that indicate both the phylogenetic lineage and other parameters like growth temperature. Therefore, before computing the phylogenetic tree, the data set consisting of TrpB sequences was processed by means of Key2Ann using the options `superkingdom phylum class order temprange`. The resulting multiple fasta file was fed into Jalview [15] to create a multiple sequence alignment and loaded into SplitsTree4 [17] to generate the phylogenetic tree shown in Figure 1. The tree makes clear that TrpB sequences from Bacteria (labels starting with a “B”) occur in a distinct subtree (upper half of the tree). For Bacteria, growth temperature (indicated by a suffix “Ht”, “Tt”, or “Mt” for hyperthermophilic, thermophilic, or mesophilic species) does not strongly affect the position of the species in the tree. The two TrpB sequences from Eukaryotes (prefix “E”) cluster with bacterial ones. In contrast to Bacteria, the localization of archaeal species in the tree is quite inhomogeneous. Mesophilic Archaea (labels with prefix “A” and suffix “Mt”) group with bacterial species. However, TrpB sequences from hyperthermophilic and thermophilic Archaea (prefix “A” and suffices “Ht” or “Tt”) form a distinct subtree (lower half of the tree). A further analysis of smaller subtree elements indicates that Crenarchaeota and Euryarchaeota (prefixes “ACr” and “AEu”) do not form distinct groups. Frequently, species from these two phyla occur together in the same subtree. Even entries belonging to the same phylogenetic class (e. g. those with prefix “ACrTh”) are spread over different subtrees. These findings indicate a complex evolution of archeal TrpB; for details see [16].

This analysis highlights a typical application of Key2Ann: The resulting labels were used to analyze the composition of subtrees or subclusters and to deduce facts about phylogenetic relation or the impact of the habitat on the composition or distribution of genes and their products. Thus, a further application might e. g. be the study of horizontal gene transfer by analyzing phylogenetic trees (for a review see [18]). In this case, outliers are relevant: Entries that cluster with phylogenetically unrelated species have most plausibly been acquired by lateral transfer. Such a case will be shown by the second application.

Many pathogenic species possess the protein SpaP, which is involved in the surface presentation of antigens (see e. g. [19]). In order to study the distribution of SpaP among pathogenic and non-pathogenic species, the same procedure as above was used to collect sequences and to create a tree. In this case, the Key2Ann options `superkingdom phylum class order pathogenicity` were used. The phylogenetic tree depicted in Figure 2 indicates that the proteins are clustered due to phylogenetic relationship and not due to pathogenicity: Closely related pathogenic as well as non-pathogenic species possess proteins with highly similar sequences. Most interestingly, the entry BPrBeNe_Cvi_Hp from the human-pathogen beta-proteobacterium *Chromobacterium violaceum* clusters with enterobacteriales - which are gamma-proteobacteria - and lies isolated from other beta-proteobacterial proteins. This observation indicates that this gene of *C. violaceum* was most likely acquired *via* horizontal gene transfer.

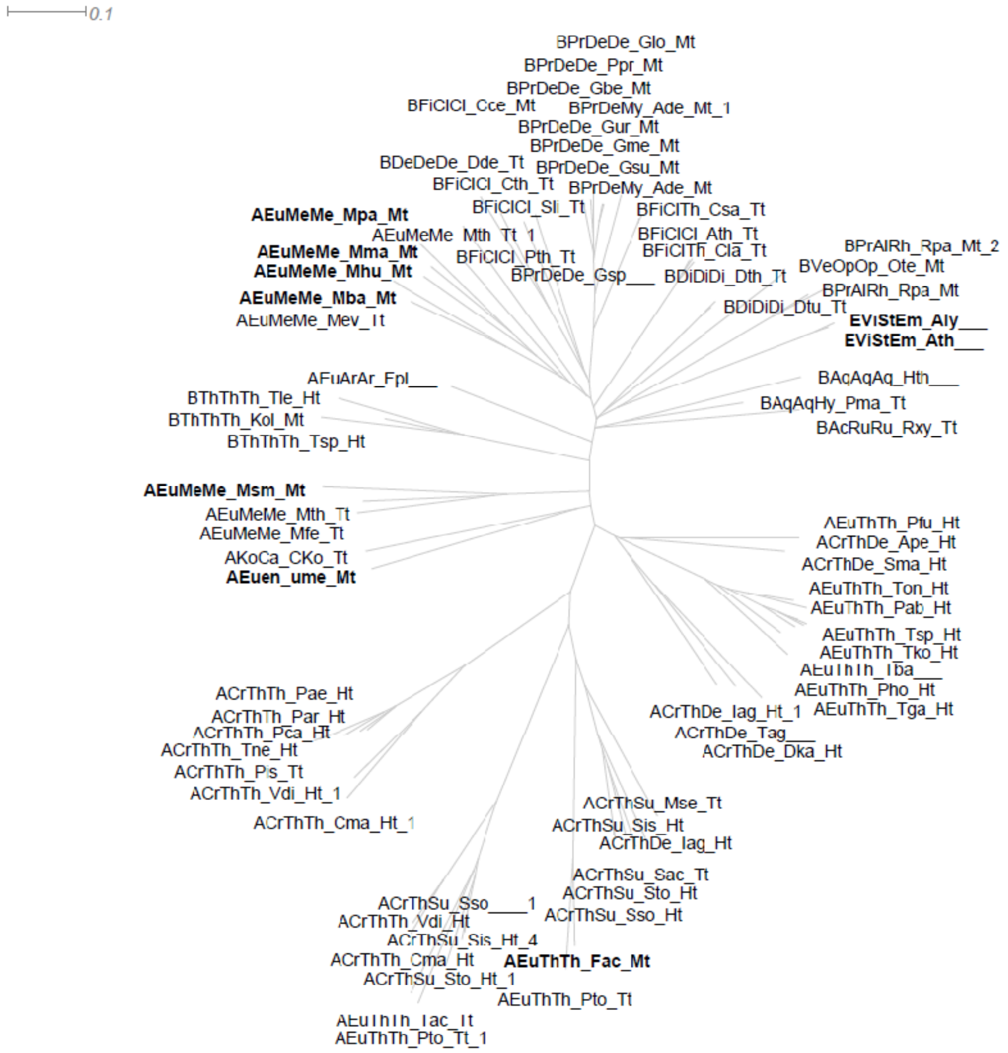

Fig. 1: A phylogeny of TrpB sequences. The sequences were collected by using BLAST and stored in a multiple fasta file. NCBI-labels were replaced by utilizing Key2Ann with the options `superkingdom phylum class order temprange`. The resulting file was processed by means of Jalview [15] and SplitsTree4 [17]. The first character of each label indicates the superkingdom ("E" Eukaryota, "B" Bacteria, "A" Archaea), the next two the phylum (e. g. "Cr" Crenarchaeota, "Eu" Euryarchaeota). The following infix gives the phylogenetic lineage (see Table 1) and the name of the species. E. g. the label AEuThTh_Pto_Tt indicates the lineage Archaea "A", Euryarchaeota "Eu", Thermoplasmata "Th", Thermoplasmatales "Th", and the species name *Picrophilus torridus* abbreviated as Pto. The ending "Tt" is one of the suffixes "Ht", "Tt", and "Mt" giving the growth temperature categorized according to hyperthermophilic, thermophilic, and mesophilic species. Missing data are indicated by a "__". A number is added if the same label was generated for two sequences (e. g. for paralogous copies or subspecies). Labels for sequences from eukaryotic species (starting with an "E") and of mesophilic Archaea (starting with an "A" and ending with an "Mt") are printed bold; see text for details related to the position of these entries in the phylogenetic tree.

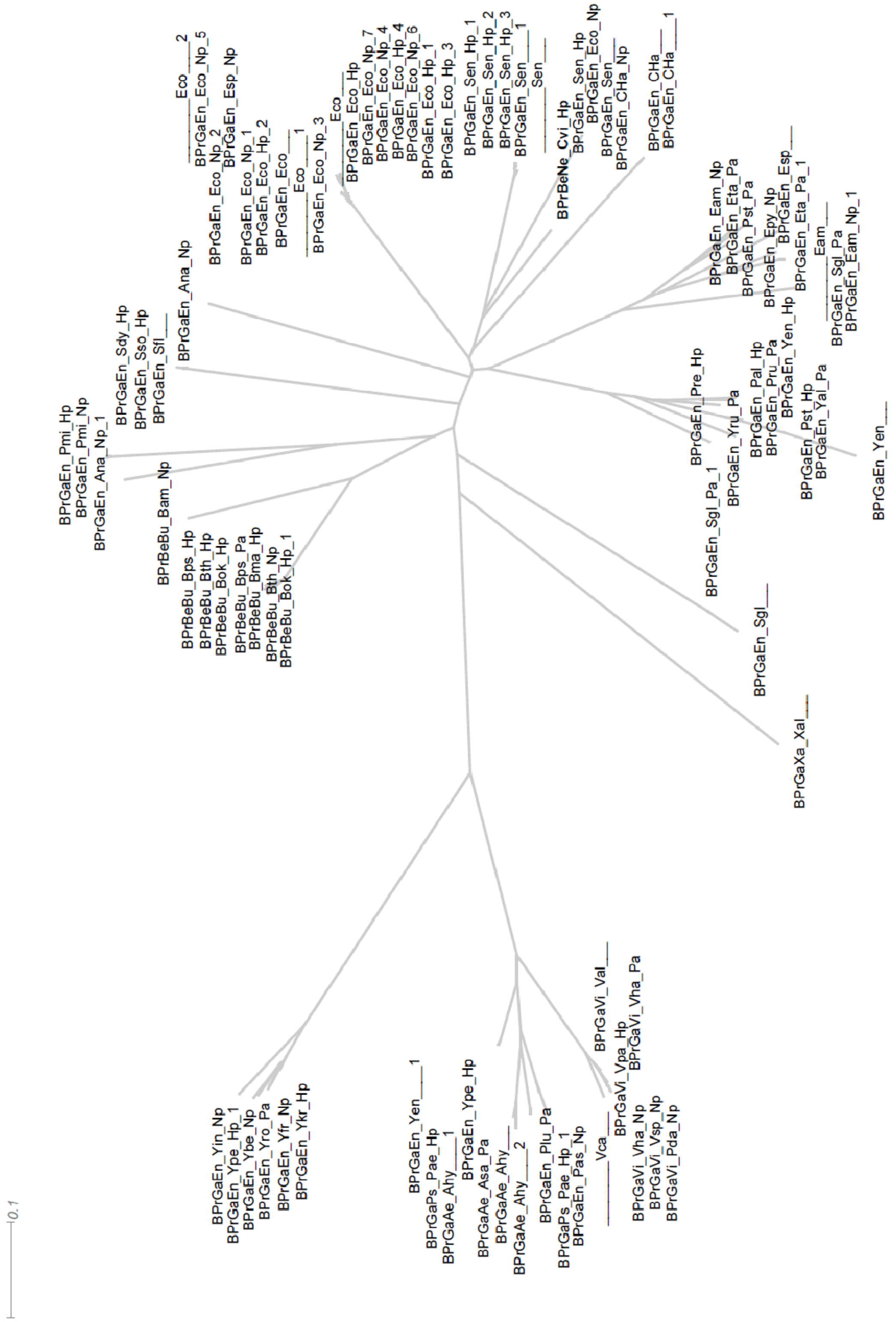

Fig. 2: A phylogeny for SpaP proteins. The sequences were collected by using BLAST and stored in a multiple fasta file. The resulting file was processed by means of Jalview [15] and SplitsTree4 [17]. Labels created by utilizing Key2Ann and the options `superkingdom phylum class order pathogenicity` replaced the NCBI-labels. The first characters indicate the phylogenetic lineage. For BPrBeNe_Cvi_Hp it is "B" Bacteria, "Pr" proteobacteria, "Be" beta-proteobacteria, "Ne" Neisseriales. "Cvi" is an abbreviation of the species name. Suffixes "Hp", "Pa", and "Np" indicate whether a species is known to be a human pathogen, a pathogen, or a non-pathogenic species. The tree makes clear that both pathogenic and non-pathogenic species possess this protein and that phylogenetic relation is the driving force for the grouping and not

the fact that a species is a pathogenic or a non-pathogenic one. The entry BPrBeNe_Cvi_Hp (printed in bold) related to the beta-proteobacterium *Chromobacterium violaceum* lies isolated from other beta-proteobacterial proteins (indicated by the prefix BPrBe) among enterobacteriales, which are gamma-proteobacteria (prefix BPrGaEn). Most plausibly, this gene was acquired via horizontal gene transfer. Missing data are indicated by a "__". A number is added if the same label was generated for two sequences (e. g. for paralogous copies).

The third example is related to the composition of Ribulose-1,5-bisphosphate carboxylase oxygenase (RubisCO), which is one of the most abundant enzymes on earth. Phylogenetic analyses have shown that there exist three classes of RubisCO proteins, catalyzing the same reactions. In addition, there exists a RubisCO-like protein; for details see e. g. [20]. In cyanobacteria as well as in phototropic and chemoautotropic proteobacteria, the enzyme consists of two subunits. As this enzyme is part of the Calvin cycle, it has an important role in carbon fixation. Bacteria possessing RubisCO have colonized many habitats which differ drastically in carbon dioxide concentration. In order to test whether the habitat has shaped in a noticeable manner the composition of the small subunit, a tree was constructed utilizing Key2Ann and the options `superkingdom phylum habitat.` As Figure 3 shows, for three bacterial phyla the proteins give rise to subtrees according to their phylogenetic lineage, irrespective of an aquatic or terrestrial habitat.

Quite frequently, questions as asked above have to be answered in many bioinformatical applications. Often, phylogenetic trees are used to study the clustering of a specific orthologe or of a specific phylogenetic group in a distinct subtree or subcluster. By querying databases to retrieve sequences, data sets can be generated easily. As demonstrated above, visual analysis of resulting trees is much easier if leaves were labeled by means of Key2Ann.

4 Integration into a Software Pipeline

Besides a visual inspection, clusters or trees can as well be analyzed by using computational methods. Due to the standardized composition of Key2Ann labels, their parsing is trivial. Thus, Key2Ann can also be integrated into a software pipeline to process automatically a large number of multiple fasta files. Whenever the pipeline consists of a script file starting a series of programs, Key2Ann can be incorporated quite easily by adding a further command line. One application might be an exhaustive search for horizontally acquired genes based on a phylogenetic approach as outlined above. Thus, one might assess each gene of a genome by means of a series of program calls. The first step of this software pipeline would be a BLASTing of the gene under study in order to collect homologous sequences. As a second step, Key2Ann can be used to create labels indicating the full phylogenetic lineage. The third step would be the construction of a phylogenetic tree. As a last step, a program exploiting the topology of the tree and the labels generated by Key2Ann could identify sequences lying isolated from closely related entries. A well-supported topological disagreement between a tree deduced for one gene family and that determined for another one is a strong indicator for horizontal gene transfer [21]. Such an approach would ideally complement methods that exploit codon usage [22] to identify alien genes.

5 Implementation

Key2Ann is written in Java. It utilizes the open source framework of BioJava [23] to access NCBI databases. To minimize execution time, the program infers taxonomical information from a local copy of the NCBI taxonomy tree. In order to deduce further properties of microbial species, we exploit a local copy of the file *ftp.ncbi.nlm.nih.gov/genomes/Bacteria/lproks_0.txt* containing key-features of more than 3600 microbial species. The command line option –`u` allows the automatic update of all locally stored data *via* download from the NCBI. If the option –`m` is used, an additional file containing a mapping that describes the correspondence of the input and the newly generated labels will be created. As explained, the names of the input and output files have to be given after the arguments `-i` and –`o`, respectively.

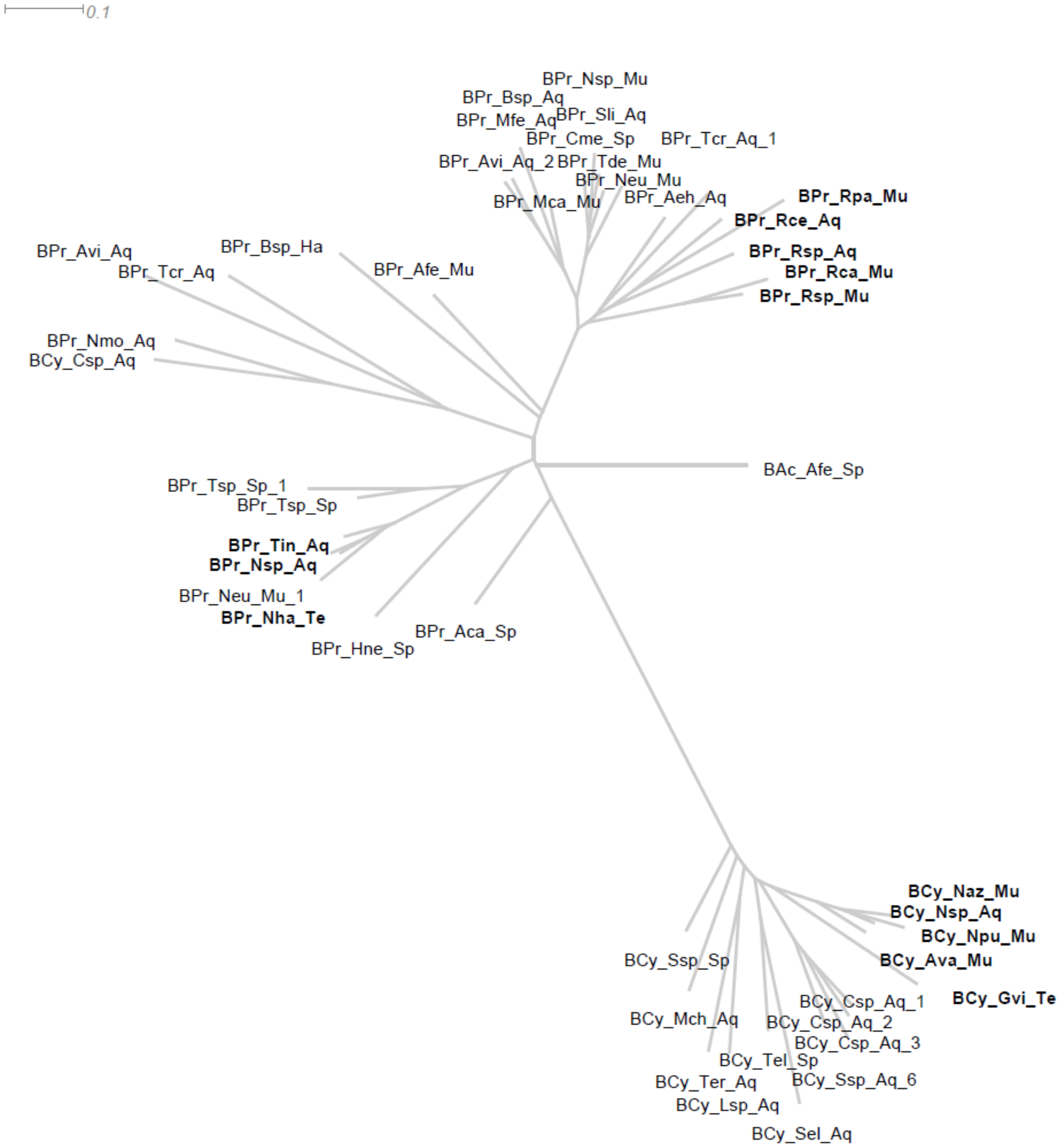

Fig. 3: A phylogeny for the small subunit of bacterial RubisCO. The sequences were collected by using BLAST and stored in a multiple fasta file. Labels created by utilizing Key2Ann and the options `superkingdom phylum habitat` replaced the NCBI-labels. The resulting file was processed by means of Jalview [15] and SplitsTree4 [17]. In the Key2Ann labels the first three characters indicate the phylogenetic lineage, i. e. "B" Bacteria, "Pr" proteobacteria; the following infix gives the name of the species. The infixes "Aq", "Te", "Mu", "Sp", "Ha" specify the habitat as aquatic, terrestrial, multiple, specialized, and host-associated. The tree shows that phylogenetic relationship and not the habitat determines the composition of the proteins. The entries printed in bold indicate that in three bacterial phyla the proteins from aquatic and terrestrial species (infixes "Aq", "Te" and "Mu") are grouped according to their phylogenetic lineage and not related to the habitat. A number is added if the same label was generated for two sequences (e. g. for paralogous copies).

As all the information needed to compile the desired labels is accessible *via* data integration frameworks, we tried during software development to collect data by means of HTTP-based application programming interfaces. We considered the NCBI C++ toolkit, the UniProtJAPI [24] and the BioJava library [23] and decided in favor of Java due to the easier mode of distributing and utilizing the application. However, benchmark tests demonstrated that when using the UniProtJAPI we were not able to assemble the data within an acceptable time interval. UniProtJAPI follows the concept of a distributed system, which implies a lot of HTTP serialization and overhead. For example, deducing the UniProtKB-key and the scientific name of a species took approximately 40 seconds for each entry due to the sequential querying of several remote servers. In contrast, the BioJava library uses a REST-based query mechanism [25], which requests the available data from all NCBI databases by means of a single query and extracts the desired information locally. As computational speed and network traffic are critical parameters of our application, we decided in favor of BioJava, which performs better than UniProtJAPI. In order to minimize the number of queries being processed *via* Internet, Key2Ann accesses NCBI databases to deduce the gene name and the EC number only, all other data are deduced from local copies of NCBI data sets. As explained above, these data sets can be updated quite easily.

We consider to extend Key2Ann in two, quite different directions: To further improve the usability and coverage, more databases should be accessible. For these, we have to find ways to retrieve data in a timely manner. A second feature that would greatly support phylogenetic analyses (see e. g. [26] as an example of sequence reconstruction) is an automatic retrieval of related gene (i. e. DNA) sequences for data sets consisting of protein sequences. In many cases, DNA sequences support better a phylogenetic analysis than protein sequences do. In contrast, finding homologs is more efficient with protein sequences. However, collecting by hand for a large set of protein sequences the related DNA entries is as tedious and error prone as the creation of labels. As this task has to be performed quite often, there is a need to develop such a tool for automatic retrieval.

6 Conclusion

We have implemented a tool focusing on a specific task to be performed quite often in computational biology. By replacing identifiers, Key2Ann creates “telling” labels which alleviate the interpretation of phylogenetic trees or clusters, especially if the data sets to be analyzed are large. As our tool is written in Java, it can be utilized on a large number of computer platforms. Due to command line arguments, the labels can be tailored according to specific needs. Integration of Key2Ann into protocols for the automatic processing of data sets by means of a series of program calls is achieved easily.

References

[1] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology, 232(2):584-599, 1993.

[2] O. Dor and Y. Zhou. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins, 66(4):838-845, 2007.

[3] J. R. Bradford and D. R. Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics, 21(8):1487-1494, 2005.

[4] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America, 95(25):14863-14868, 1998.

[5] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution, 17(6):368-376, 1981.

[6] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4):406-425, 1987.

[7] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 36(Database issue):D13-21, 2008.

[8] R. L. Tatusov, E. V. Koonin and D. J. Lipman. A genomic perspective on protein families. Science, 278(5338):631-637, 1997.

[9] L. J. Jensen, P. Julien, M. Kuhn, C. von Mering, J. Müller, T. Doerks and P. Bork. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res, 36(Database issue):D250-254, 2008.

[10] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research, 38(Database issue):D355-360, 2010.

[11] H. W. Mewes, S. Dietmann, D. Frishman, R. Gregory, G. Mannhaupt, K. F. Mayer, M. Münsterkötter, A. Ruepp, M. Spannagl, V. Stümpflen and T. Rattei. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res, 36(Database issue):D196-201, 2008.

[12] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried and R. White. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 32(Database issue):D258-261, 2004.

[13] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengut, C. J. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu and C. Yeats. New developments in the InterPro database. Nucleic Acids Research, 35(Database issue):D224-228, 2007.

[14] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats and S. R. Eddy. The Pfam protein families database. Nucleic Acids Research, 32(Database issue):D138-141, 2004.

[15] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp and G. J. Barton. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9):1189-1191, 2009.

[16] R. Merkl. Modelling the evolution of the archeal tryptophan synthase. BMC Evolutionary Biology, 7:59, 2007.

[17] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution, 23(2):254-267, 2006.

[18] R. G. Beiko and M. A. Ragan. Detecting lateral genetic transfer: a phylogenetic approach. Methods in Molecular Biology, 452:457-469, 2008.

[19] Z. Z. T. Wen, D. Yates, S. J. Ahn and R. A. Burne. Biofilm formation and virulence expression by *Streptococcus mutans* are altered when grown in dual-species model. BMC Microbiology, 10:111, 2010.

[20] F. R. Tabita, T. E. Hanson, S. Satagopan, B. H. Witte and N. E. Kreel. Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 363(1504):2629-2640, 2008.

[21] M. W. Smith, D. F. Feng and R. F. Doolittle. Evolution by acquisition: the case for horizontal gene transfers. Trends in Biochemical Sciences, 17(12):489-493, 1992.

[22] S. Waack, O. Keller, R. Asper, T. Brodag, C. Damm, W. Fricke, K. Surovcik, P. Meinicke and R. Merkl. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics, 7:142, 2006.

[23] R. C. Holland, T. A. Down, M. Pocock, A. Prlic, D. Huen, K. James, S. Foisy, A. Drager, A. Yates, M. Heuer and M. J. Schreiber. BioJava: an open-source framework for bioinformatics. Bioinformatics, 24(18):2096-2097, 2008.

[24] S. Patient, D. Wieser, M. Kleen, E. Kretschmann, M. Jesus Martin and R. Apweiler. UniProtJAPI: a remote API for accessing UniProt data. Bioinformatics, 24(10):1321-1322, 2008.

[25] Y. Kwon, Y. Shigemoto, Y. Kuwana and H. Sugawara. Web API for biology with a workflow navigation system. Nucleic Acids Research, 37(Web Server issue):W11-W16, 2009.

[26] M. Richter, M. Bosnali, L. Carstensen, T. Seitz, H. Durchschlag, S. Blanquart, R. Merkl and R. Sterner. Computational and experimental evidence for the evolution of a $(\beta\alpha)_8$-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. Journal of Molecular Biology, 398(5):763:773 2010.