In Silico Biology 3 (2003) 313–319
IOS Press

Communication

BLASTing Proteomes, Yielding Phylogenies

Georg Fuellen[1,*], Michael Spitzer[1], Paul Cullen[2] and Stefan Lorkowski[3]

[1]*Integrated Functional Genomics, IZKF, University Hospital Münster (Hautklinik), Von-Esmach-Str. 56, 48149 Münster, Germany*
E-mail: fuellen@uni-muenster.de, spitzem@uni-muenster.de
[2]*Ogham GmbH, Mendelstr. 11, 48149 Münster, Germany*
E-mail: cullen@ogham.de
[3]*Institute of Arteriosclerosis Research, Domagkstr. 3, 48149 Münster, Germany*
E-mail: stefan.lorkowski@uni-muenster.de

Edited by E. Wingender; received 4 October 2002; revised 3 March 2003; accepted 7 May 2003; published 24 May 2003

ABSTRACT: We develop a procedure called RiPE (Retrieval-induced Phylogeny Environment) that automatically performs an evolutionary analysis of a protein (sub)family, (i) by retrieving the relevant sequences via a homology search, (ii) by using the search report to construct the alignment using only homologous subsequences (taking into account their neighborhood with a low chance of homology), (iii) by realigning, and (iv) by generating phylogenetic trees based on the alignment. In a first implementation of our scheme, we start with the available proteome data of model organisms, perform a PSI-BLAST search, use MView to convert hits into a multiple alignment, and perform realignment and tree building. As a test case, we have investigated the human ABC transporters of the subfamily G, starting with the five known human ABCG transporters. Our method retrieved homologous sequences not previously analyzed, generating a tree that is more plausible and better supported than previously published trees. The RiPE 0.1 prototype is available at the RiPE website, http://ifg-izkf.uni-muenster.de/fuellen/RiPE/ripe.html**.**

KEYWORDS: ABC transporters, phylogeny, evolution, multiple alignment, homology search

INTRODUCTION

Many papers have been published that describe structural and/or functional investigations of sets of related proteins. Often, these analyses include a phylogenetic tree of the amino acid sequences. Most papers describe how the alignment and tree were constructed and provide accession numbers of the proteins used. However, descriptions of how the sequences were sampled are rare, making it difficult to know why the authors used certain sequences but not others. Moreover, it is often hard to reproduce old analyses. The best procedures described in the literature are either data sampling approaches based on

*Corresponding author.

Electronic publication can be found in *In Silico Biol.* **3**, 0027 <http://www.bioinfo.de/isb/2003/03/0027/>, 24 May 2003.

sequence motifs that are thought to occur in the family under consideration (e. g., hidden markov models, textual patterns) [Dean *et al.*, 2001; Atchley *et al.*, 1999], or BLAST-based searches using a specific E-value as the cutoff [Sicheritz-Ponten and Andersson, 2001; Lee *et al.*, 2001; Kueltz, 1998; Robinson *et al.*, 2000], or a combination of both [Sanchez-Fernandez *et al.*, 2001].

To address this deficiency, we develop a tool called RiPE (Retrieval-induced Phylogeny Environment) that automates the evolutionary investigation of a large number of proteins in a protein family. We wish to improve the analysis by including only well-defined subsets of the data. An investigation of human ABCG transporters, for example, should include only those parts of related sequences that have a minimum chance of being homologous to human ABCG proteins. This requires that the retrieval of related sequences and the determination of their homologous components be combined into a single homology search (the term "homology" is used in the sense covered by standard homology search methods such as BLAST [Altschul *et al.*, 1997], and we make no attempt to distinguish orthologs from paralogs). Our pipeline for the comparative analysis of related proteins is based on the partial pairwise alignments (high-scoring segment pairs, HSPs, in BLAST terminology) of all relevant hits. Our method does not restrict the analysis to the characteristic domain of a protein family, but rather selects the appropriate subsequences automatically using the output of the homology search algorithm. In the current implementation of RiPE, PSI-BLAST [Altschul *et al.*, 1997] is used, but the standard parameters are modified so that HSPs are extended farther than usual. Consequently, our pipeline benefits from a multiple alignment step that in particular realigns regions with a low chance of homology.

We use a selected subfamily of ABC proteins as a test case. The members of the human ABC protein family are classified by a widely used scheme that is mainly based on the structural arrangement of the ABC protein-specific domains, the ATP-binding cassette (ABC) and the transmembrane region (TMD) [Dean *et al.*, 2001]. The resulting seven subfamilies are termed ABCA, ABCB, ..., ABCG. Single polypeptides encoding a "full-transporter" display a domain arrangement of ABC-TMD-ABC-TMD or TMD-ABC-TMD-ABC. Eukaryotic transporters assembled from multiple proteins are usually formed by two "half-transporters", either ABC-TMD or TMD-ABC. Compared to the 48 known human ABC proteins, the human ABC protein subfamily G [Annilo *et al.*, 2001; Lee *et al.*, 2001; Lorkowski *et al.*, 2001a; Lorkowski *et al.*, 2001b; Engel *et al.*, 2001; Lorkowski and Cullen, 2002] studied here is composed exclusively of ABC-TMD half-transporters. Using RiPE, ABC transporters with homology to the human ABCG proteins were identified in the proteomes of model organisms, and phylogenetic analysis was performed.

MATERIALS AND METHODS

General scheme

We will first present the general scheme of our pipeline (Fig. 1).

Step 1. The input of the pipeline is k sequences $S = \{s_1, ..., s_k\}$ for which the evolutionary history in the context of related database sequences is to be elucidated. The sequences S can be aligned directly to provide a profile for a homology search. Alternatively, one sequence s_1 is declared the reference sequence and a zeroeth iteration of our pipeline is run with this query sequence, generating a multiple alignment of the homologous regions of only those sequences S that are used to start the pipeline, filtering out all other hits. In this case, the sequences $S = \{s_1, ..., s_k\}$ must all be included in the database.

Step 2. A homology search is performed using the sequences S as the profile. The homology search determines the hits (partial pairwise alignments of the query profile with database sequences).

Step 3. The hits calculated by the homology search are converted into an alignment composed of the sequences *S* (these are already aligned) and the hits, which are added ("stacked") line by line, using the profile *S* as the reference. By design, this alignment includes only "related" sequences that show homology with the sequences *S* under investigation, and the sequences *S* themselves. The "related" sequences are constrained to regions that show homology to the sequences *S*. These are partial database sequences that ignore those regions that are not homologous to the sequences *S*. The homology constraint is the result of exploiting the homology search report, since the hits in the report corresponding to the database sequences are partial, *i. e.* restricted to regions with a specified chance of homology to the query profile *S*.
Step 4. A threshold is set on the chance of homology that is required for a (partial) hit sequence to be included in the alignment. In the best case, *E*-values [Karlin and Altschul, 1990; Altschul *et al.*, 1997] can be used, and these *E*-values experience a distinctive decay at some point. The hits with significant *E*-values comprise a subfamily under the assumption that the decay in *E*-value is triggered by the absence of distinctive amino acids within sequences outside the subfamily. In general, the homology threshold has to be specified in advance based on expert opinion.
Step 5. Finally, the sequences are realigned and the tree is inferred.

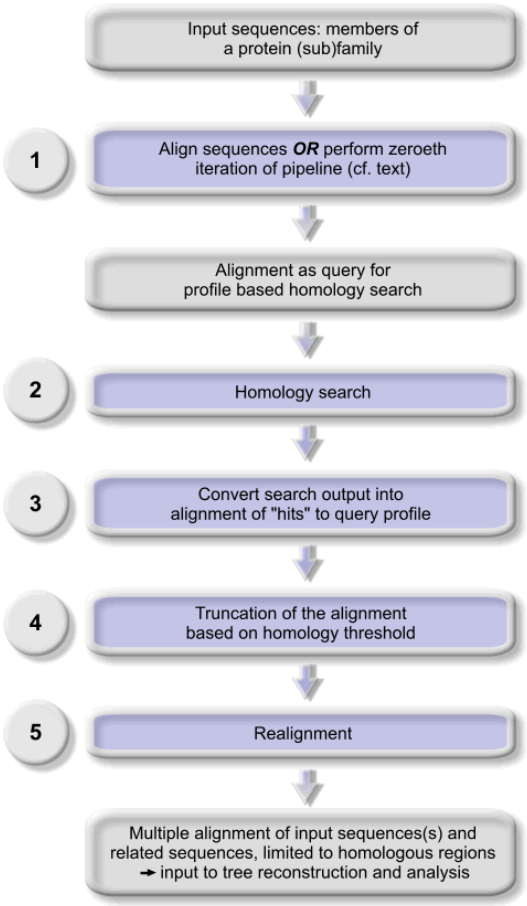

Fig.1. General scheme of the RiPE pipeline for evolutionary analysis.

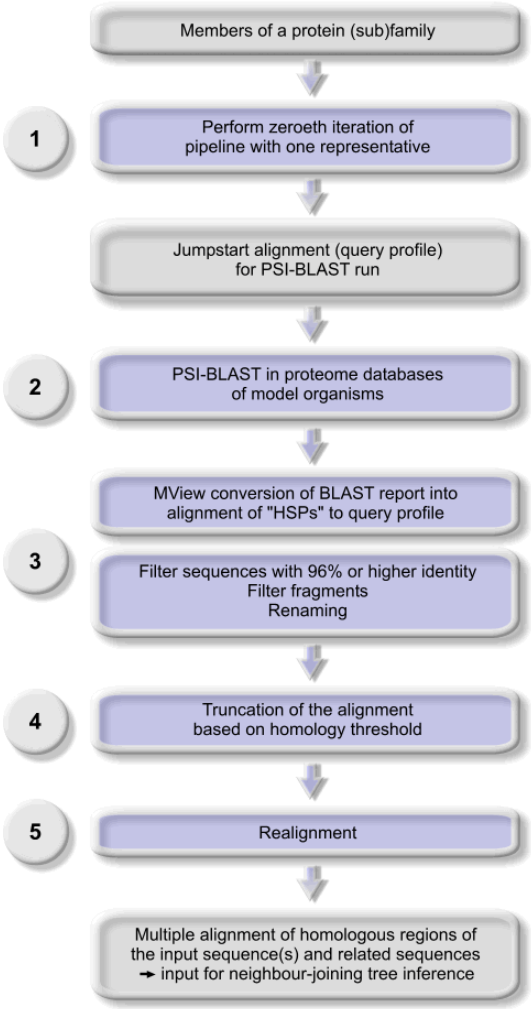

Fig.2. Implementation of the RiPE 0.1 pipeline for the evolutionary analysis of a protein (sub)family.

Current implementation - Data retrieval

In our current implementation of the general RiPE scheme (Fig. 2), RiPE 0.1, we searched the proteomes of *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Saccharomyces cerevisia* and *Escherichia coli*. Proteomes were retrieved from the EBI website on April 23, 2002 in FASTA format.

Zeroeth iteration *(step 1)*. The analysis was started using a sequence set $S = \{s_1, ..., s_k\}$ and a PSI-BLAST search (zeroeth iteration of RiPE) with s_1 as query, filtering out only the hits belonging to the other sequences in S. These partial sequences are the alignment of the sequences S.

Homology search *(step 2)*. The most important component of RiPE 0.1 is the PSI-BLAST search that is executed next. The alignment of the sequences S is used as the jumpstart alignment for the "blastpgp" executable implementing PSI-BLAST. For formal reasons, "blastpgp" also requires the query sequence to be specified; according to the *blastpgp* README this query sequence plays no role

for the execution of the search, since the jumpstart alignment is used for all positions along the query length. This behaviour is specified by using upper-case letters at all positions of the jumpstart alignment. Some parameters of the PSI-BLAST search are modified (see the supplementary data for a list) to yield high-scoring segment pairs (HSPs, partial pairwise alignments of the query profile to the database) of maximum possible length, including regions of questionable homology, if these are found in the neighborhood of regions with sufficient chance of homology.

Alignment generation *(step 3).* All HSPs are assembled into one alignment, utilizing MView [Brown *et al.*, 1998] in "discrete" stacking mode. This means that each HSP is allocated a separate row in the alignment. The step that is the hardest to automate is the filtering of redundant hits, and the renaming of proteome sequence IDs/descriptions into a human-readable standard format (see the RiPE website for details).

Homology threshold *(step 4).* If the PSI-BLAST report features a significant decay in *E*-values, there is presumably a transition from the last *S*-related sequence to the first sequence that is associated with another (sub)family. At this point, the alignment is truncated. Such a significant decay cannot be expected in all cases; it is also possible that the *E*-values level off smoothly, in particular if a subfamily does not feature one or more distinctive homologous region(s), absence of which triggers a decay in *E*-value. In such cases, it is necessary to set an *E*-value threshold that specifies the depth of the analysis.

Realignment and tree building *(step 5).* The resulting filtered BLAST-based multiple alignment of *S*-related sequences is realigned via ClustalW [Thompson *et al.*, 1994] or Dialign [Morgenstern, 1999] using default parameters, and for all alignments, the neighbor joining [Saitou and Nei, 1987] method built into ClustalW is employed to yield a tree, again using defaults.

RESULTS AND DISCUSSION

Input data and pipeline application. The five human ABCG transporters currently known are ABCG1, ABCG2, ABCG4, ABCG5 and ABCG8 [Lorkowski and Cullen, 2002]. The human proteome does not encode any other closely related proteins. To start RiPE 0.1 with all known human ABCG transporters, we performed a PSI-BLAST search using the ABCG1 protein sequence as our query, filtering only the hits of the other human ABCG transporters. Due to the high similarity between these, the result was essentially a multiple alignment of the human ABCG transporters. Execution of RiPE continues with the main PSI-BLAST search, using the alignment of the human sequences as query profile ("jumpstart alignment"; the PSI-BLAST report, the query sequence and the jumpstart alignment are available as supplementary data.) Next we assembled the multiple alignment implied by the BLAST HSPs, using MView in discrete mode. In our specific case, ABC full-transporter hits with domain structure ABC-TMD-ABC-TMD featured two separate HSPs to the jumpstart alignment (with the ABC-TMD structure typical of human half-transporters) providing a natural separation of the full-transporters into two pieces. We then performed filtering and renaming steps as discussed above. To delete redundancies in the alignment, we removed sequences that were more than 96% identical. To delete fragmentary sequences, we requested that the motifs typical of ABC proteins, the so-called Walker A, ABC signature and Walker B motifs as found in the query profile, have counterpart amino acids in the HSPs of the hits. Investigating the remaining hits, we observed a major decay of the *E*-value at sequence 81, the last sequence of this decay being sequence 83 (the resulting alignment is available as supplementary data, as is a chart of the *E*-values of the remaining hits, a table of the sequences close to the decay, and the list of the first 84 sequences that precede the decay in *E*-value). Coincidentally, a bacterial sequence (*Escherichia coli* MALK, part of the maltose transport complex) was found at the point of decay and could be used as outgroup to root trees. Apart from the distinctive decay in *E*-values, followed by hits from other ABC families, there was further evidence that we had retrieved all members of the ABCG subfamily. When we

examined the 25% of sequences following the point of decay, these were found to cluster into two subtrees. The first subtree included sequences from a variety of species, including human ABCA3 and ABCA10, but no ABCG homolog. The second subtree contained only *Escherichia coli* sequences (the tree is available as supplementary data).

ABCG tree based on RiPE. Finally, we used Dialign to realign and ClustalW software to calculate the phylogenetic tree by neighbor joining. We compared our RiPE-based tree with the tree published for the *Arabidopsis* inventory [Sanchez-Fernandez *et al.*, 2001] and with the tree published in [Annilo *et al.*, 2001]. We found that our tree was superior in many respects, as described on the RiPE website.

CONCLUSIONS

RiPE is an evolutionary analysis pipeline with a focus on homologous data only. Current limitations of our method are: 1) That it has only been tested using ABC transporters. 2) That removal of fragments and the renaming of sequences is not yet completely automated. We believe that RiPE can be developed further and be generalized to work for many multi-domain protein families.

ACKNOWLEDGEMENTS

Stefan Lorkowski and Paul Cullen are participants in the Fifth Framework project "Macrophage Function and Stability of the Atherosclerotic Plaque" (**MAFAPS,** QLG2-1999-01007) that is supported by the European Commission. We would like to thank Nigel Brown, Wayne Matten, Paul Kersey and Andrey Rzhetsky for valuable help and advice.

SUPPLEMENTARY DATA

Supplementary data can be found in the electronic publication in *In Silico Biology* at http://www.bioinfo.de/isb/2003/03/027/.

Parameters for PSI-BLAST that were modified
ABCG1 query sequence for zeroeth iteration
Jumpstart alignment of human ABCG sequences
PSI-BLAST report of homology search started with human ABCG sequences
MVIEW alignment derived from PSI-BLAST report
Chart of the *E*-values of the hits considered
Table of the sequences close to the decay
List of the first 84 sequences that precede the decay
RiPE-based tree employing Dialign without correction, including the next 25% of the sequences following the point of decay

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**, 403-410.

- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389-3402.
- Annilo, T., Tammur, J., Hutchinson, A. and Rzhetsky, A. (2001). Human and mouse orthologs of a new ATP-binding cassette gene, ABCG4. Cytogenet. Cell. Genet. **94**, 196-201.
- Atchley, W. R., Terhalle, W. and Dress, A. (1999). Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J. Mol. Evol. **48**, 501-516.
- Brown, N. P., Leroy, C. and Sander, C. (1998). MView: A Web compatible database search or multiple alignment viewer. Bioinformatics **14**, 380-381.
- Dean, M., Rhetzky, A. and Allikmets, R. (2001). The human ATP-binding cassette (ABC) transporter superfamily. Genome Res. **11**, 1156-1166.
- Engel, T., Lorkowski, S., Lueken, A., Rust, S., Schluter, B., Berger, G., Cullen, P. and Assmann, G. (2001). The human ABCG4 gene is regulated by oxysterols and retinoids in monocyte-derived macrophages. Biochem. Biophys. Res. Commun. **288**, 483-488.
- Karlin, S. and Altschul, S. F. (1990). Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA **90**, 5873-5877.
- Kueltz, D. (1998). Phylogenetic and functional classification of mitogen- and stress-activated protein kinases. J. Mol. Evol. **46**, 571-588.
- Lee, M. H., Lu, K., Hazard, S., Yu, H., Shulenin, S., Hidaka, H., Kojima, H., Allikmets, R., Sakuma, N., Pegoraro, R., Srivastava, A. K., Salen, G., Dean, M. and Patel, S. B. (2001). Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. Nat. Genet. **27**, 79-83.
- Lorkowski, S. and Cullen, P. (2002). ABCG subfamily of human ATP-binding cassette proteins. Pure Appl. Chem. **74**, 2057-2081.
- Lorkowski, S., Kratz, M., Wenner, C., Schmidt, R., Weitkamp, B., Fobker, M., Reinhardt, J., Rauterberg, J., Galinski, E.A. and Cullen, P. (2001a). Expression of the ATP-binding cassette transporter gene ABCG1 (ABC8) in Tangier disease. Biochem. Biophys. Res. Commun. **283**, 821-830.
- Lorkowski, S., Rust, S., Engel, T., Jung, E., Tegelkamp, K., Galinski, E. A., Assmann, G. and Cullen, P. (2001b). Genomic sequence and structure of the human ABCG1 (ABC8) gene. Biochem. Biophys. Res. Commun. **280**, 121-131.
- Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15**, 211-218.
- Robinson, D. R., Wu, Y. M. and Lin, S. F. (2000). The protein tyrosine kinase family of the human genome. Oncogene **19**, 5548-5557.
- Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**, 406-425.
- Sanchez-Fernandez, R., Davies, T. G., Coleman, J. O. and Rea, P. A. (2001). The Arabidopsis thaliana ABC protein superfamily, a complete inventory. J. Biol. Chem. **276**, 30231-30244.
- Sicheritz-Ponten, T. and Andersson, S. G. (2001). A phylogenomic approach to microbial evolution. Nucleic Acids Res. **29**, 545-552.
- Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673-4680.