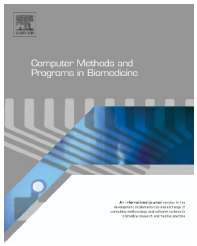
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 111 (2013) 711–714

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Pinda: A Web service for detection and analysis of intraspecies gene duplication events

*Dimitrios-Georgios Kontopoulos, Nicholas M. Glykos**

Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, 68100 Alexandroupolis, Greece

ARTICLE INFO

Article history:
Received 24 August 2012
Received in revised form
23 May 2013
Accepted 27 May 2013

Keywords:
Sequence analysis
Gene duplication
Phylogenetic analysis
Web service

ABSTRACT

We present Pinda, a Web service for the detection and analysis of possible duplications of a given protein or DNA sequence within a source species. Pinda fully automates the whole gene duplication detection procedure, from performing the initial similarity searches, to generating the multiple sequence alignments and the corresponding phylogenetic trees, to bootstrapping the trees and producing a Z-score-based list of duplication candidates for the input sequence. Pinda has been cross-validated using an extensive set of known and bibliographically characterized duplication events. The service facilitates the automatic and dependable identification of gene duplication events, using some of the most successful bioinformatics software to perform an extensive analysis protocol. Pinda will prove of use for the analysis of newly discovered genes and proteins, thus also assisting the study of recently sequenced genomes. The service's location is http://orion.mbg.duth.gr/Pinda. The source code is freely available via https://github.com/dgkontopoulos/Pinda/.

1. Introduction

With the continuous improvement of high-throughput sequencing technologies and the ever-increasing pace of sequencing whole genomes, a vast number of genomic loci has arisen, whose functions remain to be clarified. To some degree, it has been proven possible to infer the likely functions of an uncharacterized gene by elucidating its evolutionary background [1]. In this train of thought, recognizing the duplication events that led to the emergence of a given gene can be of substantial importance.

Gene duplications play a major evolutionary role, providing new genes that are usually free from selective pressure and therefore contribute to genomic plasticity [2]. Genes that originate from a duplication event gradually accumulate mutations that may cause nonfunctionalization or may bring upon phenomena of neofunctionalization or subfunctionalization [2,3]. In some specific cases, novel genes do not follow any of those models, retaining their original functions and thus increasing the dosage of a particularly important gene product [2,4].

To our knowledge, a fully automated procedure that – starting from a sequence – would facilitate the identification of specific gene duplications within a source organism is not available. Tools and programs to analyze pre-calculated trees are available [5,6] but these are not targeted for detection of intraspecies duplications and are not as straightforward to use when starting from a simple sequence. Performing this process manually requires the retrieval of relevant sequences, followed by a multiple sequence alignment and construction of a dendrogram. Furthermore, additional effort is needed in

* *Corresponding author*. Tel.: +30 25510 30620x77620; fax: +30 25510 30620.
E-mail addresses: glykos@mbg.duth.gr, nmglykos@gmail.com (N.M. Glykos).
URL: http://utopia.duth.gr/~glykos/ (N.M. Glykos).

0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.05.021

order (a) to analyze the tree topology and (b) to calculate a confidence level for the putative gene duplications that can be identified.

To fill this gap, we have developed Pinda. The service presented here runs on a GNU/Linux machine and has been written in Perl, JavaScript and R. All major browsers are supported. The service's location is http://orion.mbg.duth.gr/Pinda, whereas the source code is freely available under the Gnu Affero General Public License via https://github.com/dgkontopoulos/Pinda.

2. Methods

2.1. User interface

The user interface of Pinda is very straightforward, requiring as a bare minimum a protein (or DNA) sequence and an email address [see Supporting Figures 1–3]. The work flow adopted by the service is the following: At the starting page of the Web service, the user enters the sequence in a text box. For amino acid sequences, Pinda will then execute blastp [7] against the Swiss-Prot database and will generate a list of organisms whose hits were found in the blast report, sorted by E-value in descending order. The user may either select an organism from that list or explicitly specify an organism of their choice either by its name or via its Taxonomy ID. For nucleotide sequences, Pinda will prompt the user to explicitly define the organism of choice. The user may also choose the appropriate database for the analysis and/or select execution parameters.

Supplementary material related to this article can be found in the online version, at http://dx.doi.org/10.1016/j.cmpb.2013.05.021.

2.2. From sequences retrieval to dendrogram creation

In the second stage of the calculation, and if the input sequence is an amino acid one, PSI-BLAST [7] is run and its results are filtered by organism. The search is performed against a user-defined database (either Swiss-Prot or Uniprot). Alternatively, if the input sequence is a nucleotide one, blastn [8] is performed and its results are filtered as well. Short sequences whose length is smaller than 20% of the input sequence are excluded from the analysis. Additionally, and if the length of a hit exceeds the length of the input sequence by at least three times, only the sequence of the longest High Scoring Pair [7] is retained. Hits belonging to the organism of choice undergo multiple sequence alignment with Clustal Omega [9] for protein sequences or with Kalign2 [10] for DNA ones. Protein alignments with up to 150 sequences are by default evaluated by ZORRO [11] which directs to specific sites that are then masked in order to reduce the alignment uncertainty. DNA alignments or protein alignments containing more than 150 sequences skip this step. In the final stage, the alignment is used for the creation of a Neighbor-Joining tree using ClustalW [12]. The tree is bootstrapped [13] 1000 times and the bootstrapped version is passed to an R script which extracts bootstrap values and tip-to-tip lengths, using the R packages ape [14] and ade4 [15]. The NJ-tree is also plotted using the R script.

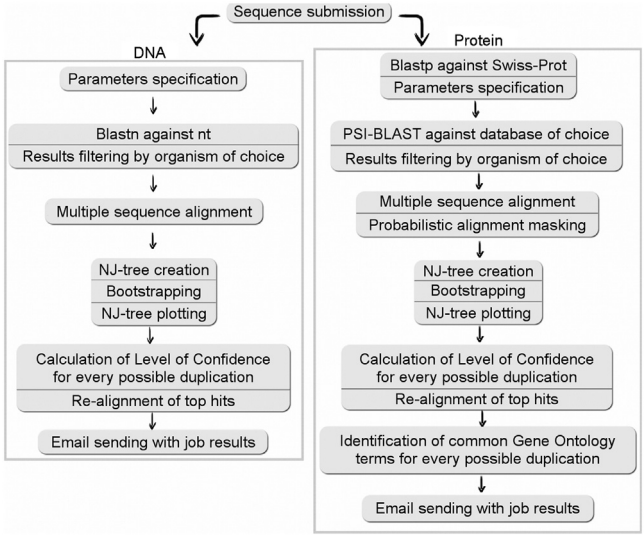


Fig. 1 – The Pinda pipeline flowchart. See text for details.

2.3. Calculation of duplication confidence

For each of the leaf nodes of the tree – except the input sequence –, a confidence value for a duplication event is calculated as the ratio of the product of bootstrap values (along the path connecting the input sequence with the leaf node) over the sum of distances from the input sequence to the leaf under examination. For the set of confidence values, a Z-test is computed and the resulting Z-values are used for the calculation of the level of confidence for each result. Hits with a level of confidence of 50% or higher are realigned with the input sequence and a new alignment is produced.

2.4. Gene Ontology terms comparison

Additionally, Gene Ontology terms [16] for every resulting sequence are compared with those of the input sequence, provided that the input sequence is a protein one. Finally, the user is notified of the results via email, which includes the results table, the NJ-tree and links to the alignments and the NJ-tree files that were produced [see Supporting Figure 4].

Supplementary material related to this article can be found in the online version, at http://dx.doi.org/10.1016/j.cmpb.2013.05.021.

The flowchart of the Web service is shown in Fig. 1.

3. Results and discussion

Pinda has been cross-validated using a wide range of proteins and genes and its results have been found to be in full agreement with literature. A representative subset of the validating test-set (including the corresponding references) is listed in Table 1.

To present a more detailed example of the application of Pinda illustrating its putative uses and applications, we discuss a test case based on the evolutionary relationship between the Type III Secretion System and bacterial flagella [23] noting that this test case has not been used during

Table 1 – A selection of test sets used for validating Pinda.

Genes/proteins	Accession numbers	Source organism	Reference
SIR2 and HST1	Swiss-Prot: P06700; P53685	*Saccharomyces cerevisiae*	[17]
LWS-1/2	Swiss-Prot: Q9W6A7; Q8AYN0	*Danio rerio*	[18]
Pseudo-COI and COI	GenBank: AB443574; AY098462	*Drosophila ananassae*	[19]
LMP7 and PSMB5	Swiss-Prot: P28063; O55234	*Mus musculus*	[20]
CLTC/CLTCL1	Swiss-Prot: Q00610; P53675	*Homo sapiens*	[21]
LSIII and NLNTP	GenBank: AB098531; AB098532	*Laticauda semifasciata*	[22]

program development. The input to the service was the FliN protein from *Pseudomonas syringae* (Uniprot code E7P3E4, see Supporting Information, Figure S1). The database selected was Uniprot and the run was performed using the defaults (Supporting Information, Figures S2, S3). The results (Supporting Information, Figure S4) are in excellent agreement with the literature: The top hit is the SPOA (Surface presentation of antigens, Uniprot Q4ZQU4) protein, followed by a series of FliN-related proteins from various pathovars, – and what is important – by a series of Type III Secretion System proteins related to the HrcQA protein, in good agreement with the established functional and structural relationship between the Type III Secretion System and bacterial flagella [23].

In general, running times vary, depending on the number of sequences similar to the query. Having said that, it takes only ∼2.5 min (on the currently available server hardware) for the detection of 29 possible duplications of the Red-sensitive opsin-1 sequence [18] in *Danio rerio* (Swiss-Prot: Q9W6A7), within the Swiss-Prot database. Running the pipeline for the DNA sequence that produces the aforementioned protein (GenBank: AB087803) against the nt database (GenBank, EMBL, DDBJ and refseq RNA entries) takes ∼13.5 min.

Given the extensive set of the calculations performed by Pinda, the service is quite efficient, with an average time-to-results time of approximately 30 min (for an otherwise idle server). We believe that with minimal effort (essentially a single copy–paste of a sequence), molecular biologists can acquire a list of possible duplication candidates of any sequence, allowing them to proceed with further analysis. Research groups lacking dedicated bioinformatics researchers (or dedicated hardware) are also offered the capability to identify specific duplication events, overcoming the bottlenecks previously mentioned. Finally, since Pinda is an open source project, an expansion of its underlying functions can be envisioned, in collaboration also with other investigators and other projects.

Conflicts of interest

There are no conflicts of interest to declare.

REFERENCES

[1] I. Friedberg, Automated protein function prediction – the genomic challenge, Briefings in Bioinformatics 7 (2006) 225–242.

[2] F.A. Kondrashov, I.B. Rogozin, Y.I. Wolf, E.V. Koonin, Selection in the evolution of gene duplications, Genome Biology 3 (2002), research0008.1–0008.9.

[3] M. Lynch, J.S. Conery, The evolutionary fate and consequences of duplicate genes, Science 290 (2000) 1151–1155.

[4] J. Zhang, Evolution by gene duplication: an update, Trends in Ecology and Evolution 18 (2003) 292–298.

[5] C.M. Zmasek, S.R. Eddy, A simple algorithm to infer gene duplication and speciation events on a gene tree, Bioinformatics 17 (2001) 821–828.

[6] C.M. Zmasek, S.R. Eddy, RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs, BMC Bioinformatics 3 (2002) 14.

[7] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research 25 (1997) 3389–3402.

[8] Z. Zhang, S. Schwartz, L. Wagner, W. Miller, A greedy algorithm for aligning DNA sequences, Journal of Computational Biology 7 (2000) 203–214.

[9] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Molecular Systems Biology 7 (2011) 539.

[10] T. Lassmann, O. Frings, E.L.L. Sonnhammer, Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features, Nucleic Acids Research 37 (2009) 858–865.

[11] M. Wu, S. Chatterji, J.A. Eisen, Accounting for alignment uncertainty in phylogenomics, PLoS ONE 7 (2012) e30288.

[12] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.

[13] J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap, Evolution 39 (1985) 783–791.

[14] E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language, Bioinformatics 20 (2004) 289–290.

[15] S. Dray, A.B. Dufour, The ade4 package: implementing the duality diagram for ecologists, Journal of Statistical Software 22 (2007) 1–20.

[16] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25–29.

[17] C.A. Froyd, L.N. Rusche, The duplicated deacetylases sir2 and hst1 subfunctionalized by acquiring complementary inactivating mutations, Molecular and Cellular Biology 31 (2011) 3351–3365.

[18] T. Tsujimura, T. Hosoya, S. Kawamura, A single enhancer regulating the differential expression of duplicated red-sensitive opsin genes in zebrafish, PLoS Genetics 6 (2010) e1001245.

[19] K. Sawamura, K. Koganebuchi, H. Sato, K. Kamiya, M. Matsuda, Y. Oguma, Potential gene flow in natural populations of the *Drosophila ananassae* species cluster inferred from a nuclear mitochondrial pseudogene, Molecular Phylogenetics and Evolution 48 (2008) 1087–1093.
[20] D.H. Bos, Natural selection during functional divergence to LMP7 and proteasome subunit X (PSMB5) following gene duplication, Journal of Molecular Evolution 60 (2005) 221–228.
[21] D.E. Wakeham, L. Abi-Rached, M.C. Towler, J.D. Wilbur, P. Parham, F.M. Brodsky, Clathrin heavy and light chain isoforms originated by independent mechanisms of gene duplication during chordate evolution, Proceedings of the National Academy of Sciences of the United States of America 102 (2005) 7209–7214.
[22] T.J. Fujimi, T. Nakajyo, E. Nishimura, E. Ogura, T. Tsuchiya, T. Tamiya, Molecular evolution and diversification of snake toxin genes, revealed by analysis of intron sequences, Gene 313 (2003) 111–118.
[23] A. Blocker, K. Komoriya, S.I. Aizawa, Type III secretion systems and bacterial flagella: insights into their function from structural similarities, Proceedings of the National Academy of Sciences of the United States of America 100 (2003) 3027–3030.