BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 22 2005, pages 4190–4191
doi:10.1093/bioinformatics/bti678

Sequence analysis

BADASP: predicting functional specificity in protein families using ancestral sequences

Richard J. Edwards* and Denis C. Shields

Clinical Pharmacology, The Royal College of Surgeons in Ireland, 123 St Stephen's Green, Dublin 2, Ireland

Received on August 2, 2005; revised on September 6, 2005; accepted on September 12, 2005
Advance Access publication September 13, 2005

ABSTRACT

Summary: Burst After Duplication with Ancestral Sequence Predictions (BADASP) is a software package for identifying sites that may confer subfamily-specific biological functions in protein families following functional divergence of duplicated proteins. A given protein phylogeny is grouped into subfamilies based on orthology/paralogy relationships and/or user definitions. Ancestral sequences are then predicted from the sequence alignment and the functional specificity is calculated using variants of the Burst After Duplication method, which tests for radical amino acid substitutions following gene duplications that are subsequently conserved. Statistics are output along with subfamily groupings and ancestral sequences for an easy analysis with other packages.

Availability: BADASP is freely available from http://www.bioinformatics.rcsi.ie/~redwards/badasp/

Contact: redwards@rcsi.ie

Supplementary information: A manual with further details can be downloaded from http://www.bioinformatics.rcsi.ie/~redwards/badasp/

INTRODUCTION

The divergence of proteins following gene duplication has long been recognized as an important process in the evolution of new or specific protein functions. Functional divergence is proposed to occur through some combination of neofunctionalization—the evolution of novel gene function—and subfunctionalization—the partitioning of two or more existing gene functions between paralogues (genes related by duplication) (Zhang, 2003). Although no consensus has yet been reached as to which process is more important, the distinction is somewhat irrelevant for the bioinformatic prediction of sites important for differences in gene function between paralogues (though it is vitally important for the interpretation of results). Instead, it is more pertinent to consider the types of substitution that occur at these sites and the phylogenetic signal that they leave.

Sites of functional change following duplication can be broadly classified into two categories, which Gu has named Type I and Type II (Gu, 2001). Type I functional divergence shows a change in selective constraint on a site following duplication, either by relaxation of existing purifying selection or by gaining functional importance at a previously unimportant site. In contrast, sites experiencing Type II divergence remain important in both duplicates but a different amino acid is favored in each duplicate. Both Type I and Type II divergence can occur as the result of either neofunctionalization or subfunctionalization. For example, subfunctionalization may occur by partitioning domain functions, with different domains maintained in different paralogues (Type I divergence), or by each paralogue specializing for a given set of existing substrates (Type II divergence). Similarly, new gene function may arise at previously unimportant sites (Type I) or by recruiting existing functional sites to the new function (Type II), while the paralogue fulfills the previous roll of the ancestral protein.

Several methods now exist to detect either Type I or Type II divergence (Lichtarge *et al*., 1996; Caffrey *et al*., 2000; Hannenhalli and Russell, 2000; Johnson and Church, 2000; Gu, 2001; del Sol Mesa *et al*., 2003; Kalinina *et al*., 2004a; Abhiman and Sonnhammer, 2005a). Many of these, however, are complex methods that lack simple software implementations and/or rely on additional information, such as structural data, which is not always available. Although there are available tools for state of the art predictions for divergence of either Type I (Gu and Vander Velden, 2002) or Type II (Kalinina *et al*., 2004b) for single protein families, there is still the need for a simple analysis package that can be run from the command line for multiple families and can potentially detect both Type I and Type II divergence. BADASP provides software to implement the previously published Burst After Duplication (BAD) algorithm (Caffrey *et al*., 2000), plus two variants for identifying Type I and Type II divergence that have been used successfully in identifying functionally interesting sites in platelet signaling proteins (unpublished data).

ALGORITHM

BADASP implements three versions of the BAD algorithm (Caffrey *et al*., 2000). All three versions are built on the same underlying assumption that sites critical to changes in gene function between paralogues are marked by a burst of radical amino acid substitutions directly after duplication, which are subsequently conserved within orthologous groups. This is calculated by comparing the changes in physiochemical properties along the relevant branches for each site:

$$\mathrm{BAD} = \mathrm{RC} - \mathrm{AC}, \quad (1)$$

where AC is the 'Ancestral Conservation' score, calculated as the change in physiochemical properties between the duplication node and the ancestral node for the subfamily; RC is the 'Recent Conservation', calculated as the mean change in properties between the ancestor of the subfamily and each orthologous (leaf) sequence.

*To whom correspondence should be addressed.

(1) BADT explicitly analyses two subfamilies, related by a single duplication event, for Type II divergence and is simply the sum of the BAD scores (1) for the two subfamilies.

(2) BADX looks for Type I neofunctionalization in a specific Query subfamily by comparing it with to its duplicate only:

$$\mathrm{BADX} = \mathrm{RC} - \mathrm{AC_X}, \quad (2)$$

where $\mathrm{AC_X}$ is a modified AC score, calculated as the change in physiochemical properties between the two post-duplication nodes; RC is for the Query subfamily only. This method is more robust to incorrect ancestral sequence assignment at the duplication node.

(3) BADN looks for Type II divergence across multiple subfamilies:

$$\mathrm{BADN} = \frac{\mathrm{BAD}\Sigma}{(N-1)}, \quad (3)$$

where BADΣ is the sum of the BAD scores (1) for each subfamily; N is the number of subfamilies. AC values (1) are calculated using the ancestor of each subfamily and the root of the tree, which should be the most ancient gene duplication.

As with other methods, all three BAD algorithms are obviously sensitive to alignment and tree quality. In addition, incorrect ancestral sequence prediction will give erroneous results.

IMPLEMENTATION

BADASP has been implemented using a set of open source Python modules. By default, the amino acid property matrix of Livingstone and Barton (1993) is used. Ancestral sequences are calculated using GASP (Edwards and Shields, 2004). This algorithm was specifically designed with BAD in mind and will reconstruct ancestral sequences for gapped columns of the alignment, allowing the use of partial sequences and/or BADN calculations for sites that are gaps in one or more subfamilies. In addition to the GASP ancestral sequences and BAD statistics, BADASP will also calculate a number of additional specificity and sequence conservation statistics to assist the interpretation of results.

Main BADASP output falls into three primary categories: (1) statistics for a given residue; (2) statistics for a given window size across (a) the whole alignment, (b) the Query protein of interest (if given) and (c) the predicted ancestral sequence of each subfamily; and (3) predicted ancestral sequences at (a) the root and (b) the ancestor of each subfamily. This output is in a tab- or comma-delimited file for easy manipulation or viewing with other programs. A batch mode with the option to output results from multiple datasets as a MySQL database is planned. In addition to this flat file, the standard ancestral sequence output of GASP (Edwards and Shields, 2004) and a file containing subfamily grouping data is also produced.

A manual with full details of algorithms, acceptable input formats, sequence statistics, output files and parameters can be found at http://www.bioinformatics.rcsi.ie/~redwards/badasp/

DISCUSSION

Sophisticated methods are now available for predicting sites of functional divergence. Abhiman and Sonnhammer have recently performed a large-scale analysis of FunShift (Abhiman and Sonnhammer, 2005a) to test its ability to discriminate between functionally diverged and functionally conserved enzymatic activities in related subfamilies (Abhiman and Sonnhammer, 2005b). However, there is currently no good dataset of individual residues responsible for functional specificity, with which different methods can be compared. Complexity is not always beneficial, and what one gains in the swings of sensitivity, one can lose in the roundabouts of interpretation. A place still exists, therefore, for an open source implementation of a simple algorithm, the results of which may be easier to understand and analyze. BADASP provides such an implementation that is suitable for use in a high-throughput automated analysis of many families. Alternatively, BADASP could provide useful supplemental data for a more focused analysis on a given family when used in parallel with a more complex method, such as DIVERGE (Gu and Vander Velden, 2002) for Type I or SDPpred (Kalinina *et al*., 2004b) for Type II divergence, or Rate Shift analysis (Abhiman and Sonnhammer, 2005a). The open source Python implementation allows extra measures of specificity or conservation to be added with relative ease.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Kate Johnston for suggestions during the preparation of the manuscript. This work was funded by the Health Research Board, Science Foundation Ireland and the Programme for Research in Third Level Institutions administered by the Higher Education Authority.

Conflict of Interest: none declared.

REFERENCES

Abhiman,S. and Sonnhammer,E.L. (2005a) FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res*., **33**, D197–D200.

Abhiman,S. and Sonnhammer,E.L. (2005b) Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins*, **6**, 6.

Caffrey,D.R. *et al*. (2000) A method to predict residues conferring functional differences between related proteins: application to MAP kinase pathways. *Protein Sci*., **9**, 655–670.

del Sol Mesa,A. *et al*. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol*., **326**, 1289–1302.

Edwards,R.J. and Shields,D.C. (2004) GASP: Gapped Ancestral Sequence Prediction for proteins. *BMC Bioinformatics*, **5**, 123.

Gu,X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol*., **18**, 453–464.

Gu,X. and Vander Velden,K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics*, **18**, 500–501.

Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol*., **303**, 61–76.

Johnson,J.M. and Church,G.M. (2000) Predicting ligand-binding function in families of bacterial receptors. *Proc. Natl Acad Sci. USA*, **97**, 3965–3970.

Kalinina,O.V. *et al*. (2004a) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci*., **13**, 443–456.

Kalinina,O.V. *et al*. (2004b) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res*., **32**, W424–W428.

Lichtarge,O. *et al*. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol*., **257**, 342–358.

Livingstone,C.D. and Barton,G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci*., **9**, 745–756.

Zhang,J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol*., **18**, 292–298.