Bias and Benefit Induced by Intra-Species Paralogy in Guilt by Association Methods to Predict Protein Function

Laurent Brehelin, Olivier Gascuel

Bias and benefit induced by intra-species paralogy in guilt by association methods to predict protein function

Laurent Bréhélin Olivier Gascuel

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier,
Projet Méthodes et Algorithmes pour la Bioinformatique,
UMR 5506 CNRS - Université Montpellier II,
161 rue Ada, 34392 MONTPELLIER Cedex 5, France

Sequence homology is a widely used principle for the functional annotation of the genes of newly sequenced genomes. It has been used for years via methods as Blast [1] or Pfam [7]. However, depending of the organism, large portions of genes cannot be annotated this way, either because no homologous genes have been already characterized, or because standard tools fail to detect homology when the divergence between sequences is too large. For example, for *Plasmodium falciparum* (the main causative agent of Malaria) only ∼40% of the predicted genes can be annotated by homology, leaving ∼60% of orphan genes. Nonhomology methods are needed to obtain functional clues for those orphan genes. Recently, methods based on post-genomic data (mainly gene expression and protein interaction data) have been proposed. These are commonly referred as *Guilt by Association* (GBA) methods. Contrary to sequence homology which works in an inter-species way —*i.e.* genes characterized in other species are used to annotate the genes of the newly sequenced genome—, GBA approaches work in an intra-species way: the genes already characterized in the genome —*e.g.* by direct assay or using homology— help for the annotation of the others genes (the guilt by association principle). Gene expression data are often used, since genes with similar transcriptomic profile are likely to share common functional roles [5]. In the same way, protein interaction data are also used, since proteins that share common interactors are likely to share common functions [3].

Part of these new post-genomic methods work in a non-supervised way (*e.g.* [5, 6]): first a gene clustering algorithm is run on the post-genomic data to cluster the genes into several groups. Then, in each cluster and for each potential function, a statistical test is applied to compare the proportion of genes annotated with this function in the cluster, with that in the complete set of genes. Functions that appear over-represented in one cluster are used to annotate the uncharacterized genes that belong to this cluster. Another part of GBA methods work in a supervised way (*e.g.* [2]): first, based on the post-genomic data of the already characterized genes, a supervised learning algorithm is run to learn a predictor, *i.e.* a function that takes as input the post-genomic measurements of a given gene, and outputs one or several functional predictions for that gene. This predictor is then used to annotate the uncharacterized genes. Other GBA methods mix several types of post-genomic data with sequence-based information (*e.g.* presence/absence of a Pfam domain), but still proceed in a supervised way [4]. Performance of these GBA annotation methods is usually assessed by cross-validation on the already characterized genes.

x1 However, all these approaches do not distinguish between genes which have, or have not, intra-species homologues (hence paralogues). Homology is a powerful source of information to predict the (secondary and tertiary) structure of proteins. As predicting the structure of a protein that has an homologue of known structure is by far more easy than when no homologue is known, we usually distinguish between the two subproblems. Approaches specially designed for one or the other problem are separately developed and tested on appropriate benchmarks. This

has the advantage both to fully benefit from homology when it is available, and to avoid misevaluating the prediction accuracy. We argue here that such a distinction should also be applied to functional annotation methods, because not accounting for intra-species paralogy actually biases the estimate of the method performance. Indeed, we show that 1) functional similarity and paralogy are closely related, *i.e.* paralogous genes often share similar kind of functions; and 2) the proportion of characterized genes that possess a characterized paralogue is by far higher than that of the uncharacterized genes. As a result, the performance computed by cross-validation on the characterized genes is an optimistic estimate of what can be expected on the uncharacterized genes.

We propose and discuss a correction procedure accounting for paralogy, which should be used to properly estimate the performance of any GBA-based annotation method. We use this procedure to estimate the optimistic bias induced by paralogues on GBA predictors, based on the analysis of *Sacharomyces cerevisiae* and *Plasmodium falciparum* transcriptomic data. Next, just as with protein structure prediction, we propose a general scheme that distinguishes between the genes that have, or have not, a characterized paralogue. This general scheme clearly boosts the accuracy of the nearest-neighbor-based supervised method that we use to illustrate our purpose. Although this study is based on particular organisms, data, and methods, its conclusions should hold for any GBA methods.

References

[1] SF Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 5 1990.

[2] Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Sugnet, Terrence S. Furey, Jr. Manuel Ares, and David Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences*, 1:262–267, 2000.

[3] C Brun, F Chevenet, D Martin, J Wojcik, A Guenoche, and B Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1), 2003.

[4] Y Chen and D Xu. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Res*, 32(21):6414–6424, 2004.

[5] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95(25):14863–14868, Dec 8 1998.

[6] KG Le Roch, Y Zhou, PL Blair, M Grainger, JK Moch, JD Haynes, P De La Vega, AA Holder, S Batalov, DJ Carucci, and EA Winzeler. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639):1503–1508, Sep 12 2003.

[7] EL Sonnhammer, SR Eddy, E Birney, A Bateman, and R Durbin. Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Res*, 26(1):320–322, Jan 1 1998.