ABSTRACT

Summary: Endeavour is a tool that detects the most promising genes within large lists of candidates with respect to a biological process of interest and by combining several genomic data sources. We have benchmarked Endeavour using 450 pathway maps and 826 disease marker sets from MetaCore™ of GeneGo, Inc. containing a total of 9911 and 12 432 genes, respectively. We obtained an area under the receiver operating characteristic curves of 0.97 for pathway and of 0.91 for disease gene sets. These results indicate that Endeavour can be used to efficiently prioritize candidate genes for pathways and diseases.

Availability: Endeavour is available at http://www.esat.kuleuven.be/endeavour

Contact: sven.schuierer@novartis.com; leon-charles.tranchevent@esat.kuleuven.be

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on December 1, 2009; revised on May 14, 2010; accepted on June 4, 2010

1 INTRODUCTION

Identifying disease causing genes is a key challenge in human genetics. In the process of identifying such disease genes, researchers are often confronted with large lists of candidate genes among which only one or a few are actually causal. The validation of each candidate is often too costly and time consuming, so that only a few candidates are further experimentally validated. A related problem arises when trying to identify new members of a biological pathway. The selection of a small subset of optimal candidates for validation is called gene prioritization. Since going manually through all possible sources of information is a slow and tedious process, several bioinformatics methods have been developed to tackle this problem (Oti and Brunner, 2007; Zhu and Zhao, 2007).

We previously developed Endeavour (Aerts et al., 2006; Tranchevent et al., 2008) whose key feature is that it uses multiple genomic data sources (e.g. sequence, expression, literature and annotation) to estimate how promising a candidate gene is by measuring its similarity with a set of training genes. The training genes are known to be involved in the process of interest, and, on the other hand, the candidate genes to prioritize (at the bottom, in grey and orange). The inputs are, on the one hand, the training genes (on top, in red), known to be involved in the process of interest, and, on the other hand, the candidate genes to prioritize (at the bottom, in grey and orange). Data are collected for these genes: e.g. expression profiles, functional annotations and protein–protein interactions. Candidate genes are prioritized, i.e. ranked according to their similarities to the training genes.

For example, the gene in orange is the most promising candidate (i.e. it ranks first in position) because (i) its expression profile is similar to the red ones, (ii) it also shares several functional annotations and (iii) it is interacting with several training proteins.

3 bio-molecular pathways and 29 genetic diseases, representing around 700 prioritizations in total (Aerts et al., 2006). In the current study, we briefly report on the largest benchmark to date for a gene prioritization method using 1276 pathways and diseases from MetaCore and prioritizing a total of 22 343 genes.

2 METHODS

We used the MetaCore™ Pathway Maps and Disease Marker Sets as provided by GeneGo, Inc. in October 2008. This resulted in 450 pathway maps containing a total of 9911 genes, and 826 disease marker sets containing a total of 12 432 genes (see Supplementary Material). In addition, the OMIM and Gene Ontology based benchmarks were built as described in Aerts et al. (2006), see also Supplementary Material. The Endeavour prioritization platform was accessed remotely using a secured connection from a command line interface allowing the automatic processing of thousands of prioritizations.

3 RESULTS

The cross-validation procedure measures the ability of the program to capture the information of the known genes and to correctly use this information to prioritize the left-out gene. To assess the ability...