Large-scale benchmark of Endeavour using MetaCore maps
Sven Schuierer, Léon-Charles Tranchevent, Uwe Dengler and Yves Moreau
1Novartis Pharma AG, Postfach, CH-4002 Basel, Switzerland, 2Department of Electrical Engineering ESAT-SCD and 3SymBioSys Center for Computational Systems Biology, Katholieke Universiteit Leuven, Leuven, Belgium

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 MetaCoreTM 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

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 genes which are already known to play a role in the biological process under study. The underlying assumption is that the most promising candidate genes are the ones that exhibit many similarities with the training genes. A schematic view of the algorithm is shown in Figure 1. Originally, Endeavour was benchmarked by leave-one-out cross-validations on 32 gene sets corresponding to 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...
Endeavour using MetaCore maps

Fig. 2. Results of the large-scale validation of Endeavour on the 450 pathways and 826 disease marker sets from MetaCore. The disease ROC curve, in green, results in an AUC of 91.65% and the pathway ROC, in red, indicates an even better performance with an AUC of 97.72%. The dotted curves represent the performance for the OMIM diseases (dotted green, 94.12%) and the GO pathways (dotted red, 93.37%). The black curve serves as a control (49.86%). The optimal control experiment would consist of shuffled gene sets but randomly selected gene sets were used as an approximation. AUCs for diseases and pathways are significantly larger than the control AUC (Wilcoxon rank sum < 1e-6).


Funding: Research Council KUL (GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA); the Flemish Government (G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, ICCoS, ANMMM, MLDLM, G.0733.09, G.0824.09, GBOU-McKnow-E, GBOU-ANA, TAD-BioScope-IT, Silicos, SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-IOTA3, O&O-Dquare); the Belgian Federal Science Policy Office (IUA7 Po625); the European Research Network on System Identification (ERNSI; FP6-NoE, FP6-IP, FP6-MC-EST, FP6-STREP, FP7-HEALTH).

Conflict of Interest: none declared.

REFERENCES


