Large-scale benchmark of Endeavour using MetaCore maps

Sven Schuierer[1,*], Léon-Charles Tranchevent[2,*], Uwe Dengler [1,*], Yves Moreau[2§]

[1]Novartis

[2]Katholieke Universiteit Leuven, ESAT-SCD & SymBioSys Center for Computational Systems Biology, Leuven, Belgium.

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

SS: Sven.Schuierer@novartis.com

LT: Leon-Charles.Tranchevent@esat.kuleuven.be

UD: Uwe.Dengler@novartis.com

YM: Yves.Moreau@esat.kuleuven.be

Description:

SS is …

LT is PhD student in bioinformatics at the K.U. Leuven.

UD is …

YM is an associate professor at the department of Electrical Engineering at the K.U. Leuven.

Background

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 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 and several bioinformatics methods have been developed to tackle this problem [1,2] because going manually through all possible sources of information is a slow and tedious process. We have previously developed Endeavour [3,4] whose key feature of Endeavour is that it uses multiple genomic data sources (sequence, expression, literature, annotation, etc.) to measure the similarity between any candidate gene and the training genes to estimate how promising that candidate gene is. The training genes are the genes already known to play a role in the biological process under study. The underlying assumption is that the most promising candidates are the ones that exhibit many similarities with the training genes. A schematic view of the algorithm is shown on figure 1. Originally, Endeavour was benchmarked by leave-one-out cross-validations on 32 gene sets corresponding to 3 bio-molecular pathways and 29 genetic diseases, representing around 600 prioritizations in total [3]. In the current study, we briefly report on the largest benchmark to date for a gene prioritization method by benchmarking Endeavour using 1287 pathways and diseases from MetaCore and prioritizing a total of 22,752 genes.

Results and discussion

The cross-validation procedure measures the ability of the program to capture the information from the known genes and to correctly use that information to recall the

left-out gene. To assess the ability of Endeavour to capture the information of known pathway and disease-related gene sets, we used the Pathways Maps and Disease Pathways from MetaCore. Since the gene sets in MetaCore are manually curated, they are a reliable representation of the current knowledge of the functional contexts in which the genes are active. We have benchmarked Endeavour using 454 pathway maps and 833 disease maps containing respectively a total of 10,053 pathway members and 12,699 disease genes. For each prioritization run, the position of the left-out gene among the 100 candidates is recorded, Receiver Operating Curves (ROC) are eventually built (see Figure 2) and the area under the curve (AUC) is used as a measure of the performance. We obtained an AUC of 0.XX for the pathways. Moreover, 64% of the prioritizations have the left-out gene being ranked in the first position. The AUC value obtained for the diseases is 0.XX and 33% of the prioritizations have the left-out gene being ranked in the first position. Altogether, the results indicate that Endeavour efficiently prioritizes candidate genes for both pathways and diseases. As observed and discussed in our previous work [3], the performance of gene prioritization is higher for pathway than for diseases, because diseases often implicate a complex set of cascades making their profiling more challenging. Assessing the performance of gene prioritization methods, a novel type of bioinformatics tool, is of crucial importance. Our large-scale benchmark adds an important element to the demonstration of the effectiveness of gene prioritization methods. Interestingly, S. Schuirer and U. Dengler carried out the initial steps of evaluation independently from the core Endeavour team and without extensive prior knowledge of the Endeavour platform, so that biases caused by excessive fine-tuning (which are frequent in self-reported performance evaluation) have been avoided. We are aware of the many pitfalls of benchmarking gene prioritization and function

prediction methods [MYERS ET AL 2006], so that the performance observed in cross-validation studies is likely to be higher than that observed in prospective studies. We have recently conducted such a prospective validation in Drosophila [AERTS ET AL 2009], which also confirmed further the effectiveness of our strategy.

Methods

We used the MetaCore Pathway Maps and Disease Pathways as provided by GeneGo in October 2008. The first step was to map the EntrezGene identifiers used by MetaCore to EnsEMBL gene identifiers used by Endeavour. The pathway maps contain 3,328 distinct EntrezGene genes of which we could map 3,299 (99%) to EnsEMBL ids. The disease maps contain 8,231 distinct EntrezGene genes of which we could map 7,198 (87%) to EnsEMBL ids. Mapping from EntrezGene to EnsEMBL identifiers was obtained from BioMart [5] on December 11, 2008. We furthermore restricted the gene sets to sizes between 5 and 50 because too small or too large maps do not represent the most suitable sets for cross-validation. This resulted in 454 pathway maps containing a total of 10,053 genes, and 833 disease pathways containing a total of 12,699 genes. The Endeavour program was used remotely using a secured connection and from command line allowing the automatic run of thousands of prioritizations. The ROC curves were built using TIBCO.

Funding

This work was supported by the 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, MLDM, G.0733.09, G.082409, GBOU-McKnow-E, GBOU-ANA, TAD-BioScope-IT, Silicos, SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-IOTA3, O&O-Dsquare]; the Belgian Federal Science Policy Office [IUAP P6/25]; and the European

Research Network on System Identification (ERNSI) [FP6-NoE, FP6-IP, FP6-MC-EST, FP6-STREP, FP7-HEALTH].

Acknowledgements

(Text for this section if needed).

Key

1. Endeavour is a tool that detects the most promising genes within large list of candidates with respect to a biological process of interest and by combining several genomic data sources.

2. We have benchmarked Endeavour using 454 pathway maps and 833 disease maps from MetaCore containing, respectively, a total of 10,053 and 12,699 genes. The results show that an AUC of 0.XX and 0.XX can be obtained respectively for pathways and diseases.

3. The results indicate that Endeavour can be used efficiently to prioritize candidate genes for pathways and diseases.

Figures

Figure 1 - The Endeavour algorithm

A. 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). B. Data are collected for these genes: e.g., expression profiles, functional annotations, and protein-protein interactions. C. 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 in first 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.

Figure 2 – Performance of the validation

Results of the large-scale validation of Endeavour on the 454 pathways and 833 diseases from MetaCore. The disease receiver operating curve (ROC), in dark red, indicates an AUC of 0.XX and the pathway ROC, in light blue, indicates even a greater performance of 0.XX.

Reference List

1. Zhu M, Zhao S: **Candidate gene identification approach: progress and challenges.** *Int J Biol Sci* 2007, **3:** 420-427.

2. Oti M, Brunner HG: **The modular nature of genetic diseases.** *Clin Genet* 2007, **71:** 1-11.

3. Aerts S, Lambrechts D, Maity S *et al.*: **Gene prioritization through genomic data fusion.** *Nat Biotechnol* 2006, **24:** 537-544.

4. Tranchevent LC, Barriot R, Yu S *et al.*: **ENDEAVOUR update: a web resource for gene prioritization in multiple species.** *Nucleic Acids Res* 2008, **36:** W377-W384.

5. Haider S, Ballester B, Smedley D *et al.*: **BioMart Central Portal--unified access to biological data.** *Nucleic Acids Res* 2009, **37:** W23-W27.

1: Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. Finding
function: evaluation methods for functional genomic data. BMC Genomics. 2006 Jul
25;7:187.

1: Aerts S, Vilain S, Hu S, Tranchevent LC, Barriot R, Yan J, Moreau Y, Hassan
BA, Quan XJ. Integrating computational biology and forward genetics in
Drosophila. PLoS Genet. 2009 Jan;5(1):e1000351. Epub