Poster D-5

Gene prioritization via genomic data fusion

Authors:
Peter Van Loo (*Human Genome Laboratory, Department of Human Genetics, Flanders Interuniversity Institute for Biotec*)
Stein Aerts (*Laboratory of Neurogenetics, Department of Human Genetics, Flanders Interuniversity Institute for Bi*)
Diether Lambrechts (*The Center for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotech*)
Sunit Maity (*The Center for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotech*)
Bert Coessens (*Bioinformatics group, K.U.Leuven ESAT-SCD*)
Frederik De Smet (*The Center for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotech*)
Leon-Charles Tranchevent (*Bioinformatics group, K.U.Leuven ESAT-SCD*)
Bart De Moor (*Bioinformatics group, K.U.Leuven ESAT-SCD*)
Peter Marynen (*Human Genome Laboratory, Department of Human Genetics, Flanders Interuniversity Institute for Biotec*)
Bassem Hassan (*Laboratory of Neurogenetics, Department of Human Genetics, Flanders Interuniversity Institute for Bi*)
Peter Carmeliet (*The Center for Transgene Technology and Gene Therapy, Flanders Interuniversity Institute for Biotech*)
Yves Moreau (*Bioinformatics group, K.U.Leuven ESAT-SCD*)

Short Abstract: We developed a novel bioinformatics method, ENDEAVOUR, to prioritize candidate genes underlying pathways or diseases, based on similarity to genes known to be involved in these processes. ENDEAVOUR can fuse information from multiple heterogeneous data sources. We successfully validated ENDEAVOUR computationally, as well as in vitro and in vivo.

Long Abstract:

The identification of genes involved in health and disease remains a formidable challenge. Here, we describe a novel bioinformatics method to prioritize candidate genes underlying pathways or diseases, based on their similarity to genes known to be involved in these processes. It is freely accessible as an interactive software tool, ENDEAVOUR, at http://www.esat.kuleuven.be/endeavour. Unlike previous methods, ENDEAVOUR generates distinct prioritizations from multiple heterogeneous data sources, which are then integrated, or fused, into one global ranking using order statistics. The data sources include sequence information (Blast, Interpro, regulatory motifs and cis-regulatory modules, disease probability), annotation (Gene Ontology, KEGG, text mining), expression data (EST libraries, microarray Gene Atlas), and protein-protein interaction. In addition, ENDEAVOUR offers the flexibility of including external data sources, such as in-house microarray data.
ENDEAVOUR prioritizes candidate genes in a three-step process. First, information about a disease or pathway is gathered from a set of known “training” genes by consulting multiple

data sources. Next, the candidate genes are ranked based on similarity with the training properties obtained in the first step, resulting in one prioritized list for each data source. Finally, ENDEAVOUR fuses each of these rankings into a single ranking, providing an overall prioritization of the candidate genes.

We validated ENDEAVOUR by a large-scale leave-one-out cross-validation. In each validation run, one gene was removed from a set of training genes and added to 99 random genes. We then determined the ranking of this left-out gene. We used 627 known disease genes from 29 different diseases and 76 known pathway genes from 3 receptor signaling pathways. The median rank of the left-out disease genes was 3 out of 100 and of the pathway genes 2 out of 100. Thus, ENDEAVOUR can efficiently prioritize both disease and pathway genes.

To assess whether ENDEAVOUR can also identify novel monogenic and polygenic disease genes, we performed 16 prioritizations of recently identified disease genes, each time using literature information only up to one year prior to their identification. For monogenic disease, 50% of disease genes were prioritized within the top 2% of candidate genes. For polygenic diseases, 50% were prioritized within the top 15%.

Furthermore, in a study of the myeloid differentiation pathway, we prioritized genes that had been linked to this pathway by microarray and cis-regulation analysis. In vitro validation showed that prioritization resulted in a significant increase in the number of true regulatory targets.

Finally, as a most stringent test, we validated ENDEAVOUR in an animal model in vivo. The DiGeorge syndrome (DGS) is a common congenital disorder, in which craniofacial dysmorphism and other defects result from abnormal development of the pharyngeal arches. ENDEAVOUR prioritization of 58 candidate genes in a 2 Mb region, involved in atypical cases of DGS, identified YPEL1 as a novel putative DiGeorge syndrome gene. In vivo validation in zebrafish by morpholino knockdown revealed that this gene is indeed involved in pharyngeal arch development.

As a conclusion, ENDEAVOUR offers novel opportunities for gene discovery.