ENDEAVOUR–Prioritising Disease Genes Using Multiple, Heterogeneous Information Sources

Bert Coessens*, Stein Aerts*, and Bart De Moor

Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Belgium
*These authors contributed equally to this work.

ABSTRACT

Summary: ENDEAVOUR is an application in Java designed towards the prioritisation of candidate disease genes. Different information sources are used to establish the similarity between the candidate genes and a disease model. These similarities are combined using order statistics to obtain an overall ranking with significance threshold.
Availability: ENDEAVOUR can be launched via Java Web Start at http://www.esat.kuleuven.ac.be/endeavour.
Contact: bert.coessens@esat.kuleuven.ac.be

INTRODUCTION

In the field of linkage analysis and association studies researchers are often confronted with large lists of candidate disease genes, especially when investigating complex multigenic diseases. Investigating all possible candidate genes is a tedious and expensive task that can be alleviated by selecting for analysis only the most salient genes.

In this context we developped a method for the computational prioritisation of a list of candidate disease genes based on multiple heterogeneous information sources [**?**]. It is a generic approach based on the assumption that a new candidate disease gene has 'similar properties' as a set of other genes that represents a certain biological case (for instance, a disease). The training set can be seen as a model for the biological case comprising multiple submodels, one for each information source. We built an application called ENDEAVOUR to allow automated retrieval and summarisation of these different information sources, and visualisation of the results. ENDEAVOUR includes ten default information submodels which can be added to the model at the user's discretion. Users can also create and add their own information submodels.

DESCRIPTION

ENDEAVOUR currently uses ten different information sources: GO annotation, textual descriptions from MEDLINE abstracts, microarray gene expression data, EST-based anatomical expression data, KEGG pathway membership, protein domain annotation of Interpro, transcription factor binding sites (TFBSs) and modules of TFBSs, BIND protein interaction data, and BLAST sequence similarity. The 'similarity' between the properties of two genes is measured either using correlation measures (for vector-based data) or using meta-analysis (for attribute-based data). The meta-analysis also takes genome-wide occurrences of attributes into account. All candidate genes are ranked according to the similarities of each property separately. The ten resulting rankings are combined into an overall rank and p-value using order statistics.

In the first step, the user creates a model `MODEL` by generating a training set `TRAIN` starting from a list of gene identifiers. These can be Ensembl, HUGO, or LocusLink IDs. HUGO or LocusLink IDs are automatically mapped to Ensembl IDs via the Ensembl Mart database, and the Ensembl ID is used further as the central reference. If one has gene lists with other types of identifiers, they can be mapped to Ensembl IDs using web-based tools

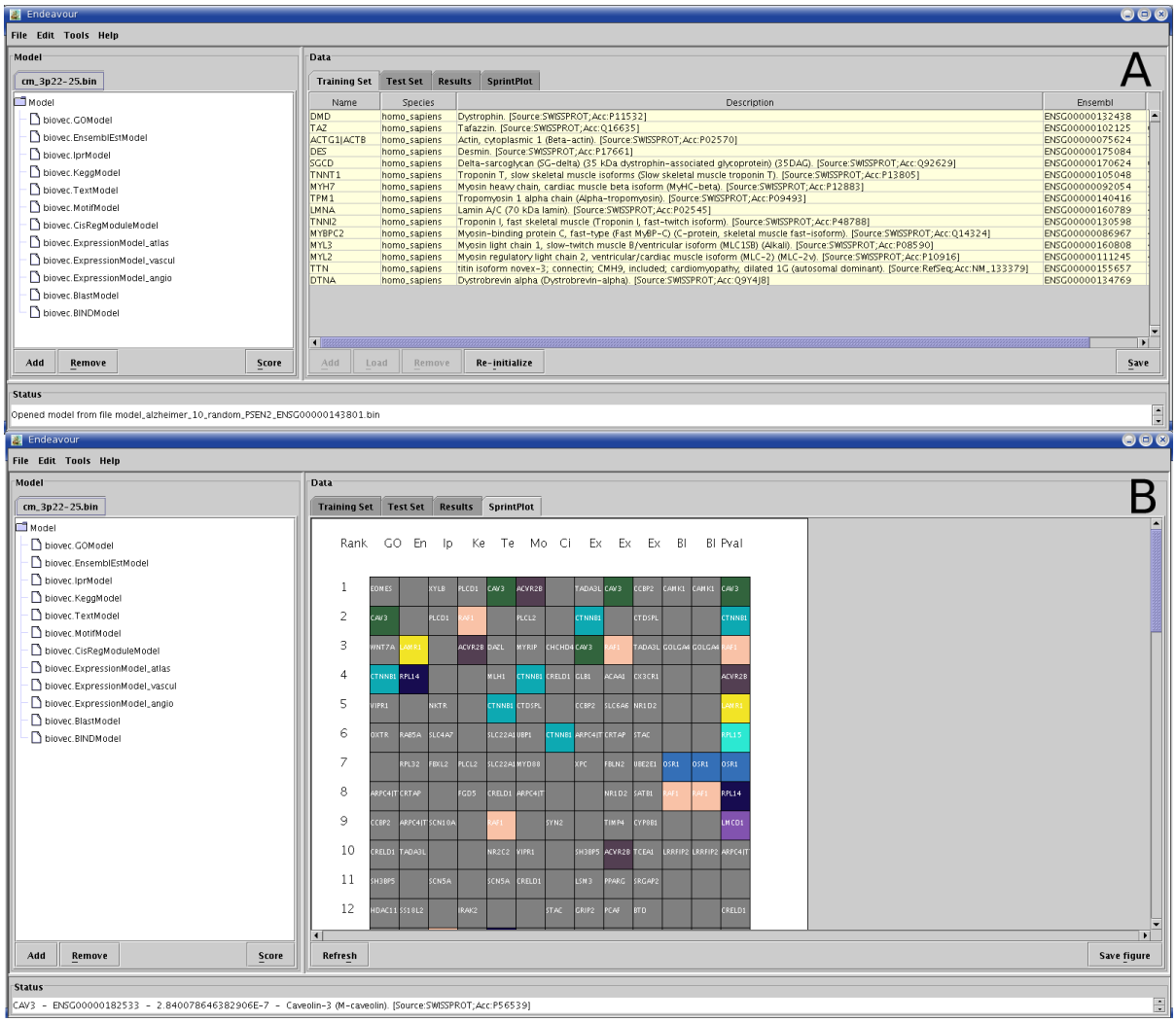


Figure 1: The ENDEAVOUR software tool. (A) A training set has been loaded (each yellow record represents a gene), and ten submodels have been trained on this set, as shown in the white box on the left. (B). A set of test genes has been scored with all submodels, resulting in a spreadsheet view of the scores (the 'Results' tab, not shown) and in a Sprintplot view shown here. Each gene can be given a color, and the location of a gene in a column represents its rank for this submodel. The last column is the overall ranking based on the combination (by order statistics) of the ten individual rankings. More screenshots with detailed descriptions can be found in the ENDEAVOUR manual at http://www.esat.kuleuven.ac.be/endeavour.

like EnsMart [1] or MatchMiner [2].

Then, the user can initialise `MODEL`, and by simply clicking the 'Add' button, he or she can choose which of the available submodels to train. The training is done automatically, and for each submodel, the user can consult the training results. These results are descriptions with characteristics that are specific for the training set. For example, for the text submodel, the 20 most descriptive terms are given.

Next (or meanwhile the submodels are being trained), a test set `TEST` can be constructed similarly as the construction of a training set. Then the genes in the test set can be scored with all trained submodels by pressing the 'Score' button. The results of the scoring are visualized in two different ways. The first is a spreadsheet where each column contains the scores of each used submodel, and the rows represent the test genes. The user can sort the test genes according to any of the columns. The second visualization is an interactive image that we call a 'Sprintplot'. It is a matrix of colored squares representing the test genes. In each column, the squares are sorted differently, that is, according to one of the used submodels. In the last column the squares are sorted according to the overall rank, based on order statistics. Using a color scheme, the user can quickly observe all the rankings for all the submodels at once, and can thereby judge which test gene scores high according to certain submodels. Training sets, test sets, results, and complete models with all information can be saved in binary format or flat files for further use. The speed of model training and scoring is only several to tens of seconds for all data types exept for the *cis*-regulatory module model and for the BLAST model which take several minutes due to the retrieval of upstream human-mouse conserved sequences and the combinatorial optimization step for the former and due to the retrieval and indexing of coding sequences for the latter.

DISCUSSION

We developped an application in Java called ENDEAVOUR that is able to prioritise large lists of genes. The prioritisation is realised by summarising heterogeneous information from different sources and comparing it with a set of training genes that comprises a model for a disease or biological process. The results of the different comparisons is combined using order statistics to obtain a p-value for each candidate gene that can be used to place a significance threshold.

To our knowledge this is the first time a bioinformatics method integrates this many data types of different origins, representing known information from the literature and curated databases, as well as new information that allows unknown genes to be involved in the prioritisation as well.

Future work comprises the assessment of weighting schemes ... or the automatic calculation of weights from a training set (if good negative examples are available).

ACKNOWLEDGEMENTS

We wish to thank all groups and consortia that made their data freely available: Ensembl, NCBI (LocusLink and MEDLINE), Gene Ontology, BIND, KEGG, Atlas, InterPro, BioBase (public release of TRANSFAC at www.gene-regulation.com), and the Disease Probabilities from Lopez-Bigas and Ouzounis [3]. We also thank Joke Allemeersch for her advice on the order statistics and Peter Van Loo for testing our software within a wet-lab setting.

References

[1] E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X.M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp, and T. Hubbard. Ensembl 2004. *Nucleic Acids Res*, 32(1):D468–70, 2004.

[2] K.J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay, and J.N. Weinstein.

MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol*, 4(4):R27, 2003.

[3] N. Lopez-Bigas and C.A. Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32(10):3108–14, 2004.