# CrossEndeavour: Gene Prioritization using Multiple Species

Léon-Charles Tranchevent[†], Sonia M. Leach[†], Shi Yu and Yves Moreau

September 16, 2009

**Abstract**

As more and more genomic-scale information becomes available, there becomes a growing need to integrate these sources, not only in a single species but across species. In particular, gene prioritization tasks can benefit greatly from integrating prioritizations within multiple organisms to support the results from a single species. In this paper, we present CrossEndeavour, an extension to Endeavour, the current state of the art prioritization software, which performs prioritization of genes in a target organism through integration of results on homologous genes in other organisms.

## 1 Introduction

Genomic-scale assays provide excellent hypothesis-generating tools for biologists, yet follow-up studies of the lists generated by these assays still require considerable cost and effort. The goal then becomes to prioritize the candidates based on prior information. Among the best state of the art programs for gene prioritization is the Endeavour tool [1], which uses genomic data fusion to build knowledge models of a training set of genes and prioritizes candidate genes based on their relationship to the training set according to the data models. Currently this prioritization is performed in a single species using data available for that organism, which may be sparse for a given data source or in general, for a given organism. The situation is particularly critical in newly-sequenced organisms though the problem exists even for well-studied organisms. For example, popular databases have very little protein-interaction data for mouse relative to that for yeast, worm or fly.

An increasingly popular technique to overcome the sparsity of data for a given organism is to transfer information from one organism by homology, where

*Dept. of Electrical Engineering (ESAT/SCD/BioI), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven Belgium. {leon-charles.tranchevent, sonia.leach, shi.yu, yves.moreau}@esat.kuleuven.be

[†]These authors contributed equally.

a score representing the reliability of the information often takes into account the degree of sequence homology. Here we describe CrossEndeavour, an extension of the Endeavour tool which allows integration of information from a variety of organisms using a homology-based technique.

# 2 Methods

## 2.1 Extending Endeavour to use Multiple Organisms

Given a training set of genes known to be involved in a biological process of interest, the Endeavour tool [1] for a single organism prioritizes a set of candidate genes by (i) inferring several models (based on various genomic data sources), (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. The model types span a range of available genomic data sources: ontologies and annotations, proteinprotein interactions, cis-regulatory information, gene expression data sets, sequence information and text-mining data. Currently, Endeavour is available for the following model organisms: Homo sapiens (human), Rattus norvegicus (rat), Caenorhabditis elegans (worm), Danio rerio (zebrafish), Drosophila melanogaster (fruit fly) and Mus musculus (mouse).

CrossEndeavour overcomes the fact that many model types for a given organism can have sparse coverage by simultaneously performing prioritization in multiple organisms using orthologous genes and augmenting the single organism results. In the single organism setting, Endeavour computes a distance value for each candidate gene against the training set profile according to a given model. Figure 1a depicts an example where a candidate gene is tested against a training set in human. CrossEndeavour adjusts the value computed in the single-species setting by integrating prioritization values from multiple species. CrossEndeavour generates an orthologous training set and orthologous candidate gene for each additional organism using pairwise homology scores by creating a one-to-one correspondence between each gene in the target organism to the gene in another organism with the maximum homology score (see description below of the orthology mapping). Figure 1b shows the orthologous test genes and training sets in rat and mouse with the corresponding homology scores. Figure 1c details how the distance value of the candidate gene in the target organism is adjusted by the distance values in the other organisms computed using the equivalent model type. These adjusted values are then used in place of the original single-species values to create a ranking for the model, which is then later merged with rankings from other model types, as done in the original Endeavour setting.

The cross-species adjustment to the test gene value in the target organism is comprised of three components. An obvious component from each alternate organism is the value computed using the single-species version of Endeavour to test the orthologous test gene against the orthologous training set. For example,

the value $v_8$ in Figure 1b is the value computed by single-species Endeavour for the rat test gene $R8$ against the training set $\{R4, R5\}$. Note that an orthologous training set can be smaller than the original human training set $\{H1, H2, H3\}$ either because multiple human training genes map to a single homolog (albeit with different homology scores), as in the example for rat, or because a human training gene fails to map to a homolog, as in the example for mouse. Thus, the adjustment includes a component $Z$ to discount for differences in the size of the training set in the alternate organism relative to that in the target organism. For the example Figure 1b, $Z = 2/3$ for both rat and mouse. The third component $H$ of the adjustment incorporates the strength of homology between genes in the target organism and those in the alternate organism. A summary of the homology between training sets is captured by taking the median homology, while taking the minimum of the test gene homology with the median training set homology provides a conservative estimate of the overall strength of the orthology mappings.

Using the quantities $v$, $H$ and $Z$ for each alternate organism (each of which is assumed to lie in the range $[0, 1]$, the adjusted value in the target organism is computed using a noisy-or model: $Y = 1 - \prod_i (1 - X_i)$ (see [2] for a useful exposition). The noisy-or function has the useful property that the value $Y$ is high with at least one high $X_i$, yet increases with additional support. Since in the Endeavour setting, *low* values are desirable, the formula for the adjustment uses the fact that $v, H$ and $Z$ are in the range $[0, 1]$ and subtracts from 1.0 to yield high or low values as necessary. Thus, given the single-species values $v_i$, training set size discounts $Z_i$, and orthology strength discounts $H_i$ for each alternate organism $i$, together with the single-species value $v$ for a test gene in the target organism, the cross-species value $v'$ is computed as:

$$v' = v * adj \tag{1}$$

where $adj = \prod_{\text{organism } i} (1 - H_i Z_i (1 - v_i))$. For a given model type in Endeavour, these cross-species adjusted values are used in place of the single-species values when computing the combined ranking over all model types.

There are a number of practical considerations given this formula. First, the raw values returned by Endeavour for each model type in each organism must be scaled to be compatible with those returned by other model types and to have a distribution over the range $[0, 1]$ that is reasonable with respect to the $Z$ and $H$ discounts. Figure 2 shows some typical distributions of $H$ and raw values for various Endeavour models, where for example in the right panels, the text-mining based model 'Text' value distribution is spread over a large range while values from the Gene Ontology model 'GO' is highly clustered in the near-zero range. In contrast, the $H$ values in the left panels of Figure 2 occupy the full range of values. The raw values returned by a given model are scaled to the range $[0 + \epsilon, 1 - \epsilon]$ using the formula

$$v = (1 - 2\epsilon) \frac{(v_{\text{raw}} - v_{\text{min}})}{v_{\text{max}} - v_{\text{min}}} + \epsilon.$$

The values $v_{\min}$ and $v_{\max}$ refer to the minimum and maximum, respectively, of the set of raw values $v_{\text{raw}}$. A minimum value of $\epsilon$ ensures that the final adjusted value takes into account any homolog information, which would be otherwise ignored in the product in Equation 1 if the minimum value was set exactly to 0.0. The maximum value of $1 - \epsilon$ ensures that the product takes into account the gene with the lowest ranked value under the Endeavour model.

The second important consideration concerns genes in the target organism which have no homologs in the alternate organisms or no homologs with information under the particular Endeavour model used. Since the noisy-or function favors genes with more and more support in additional organisms, it is possible for a gene which originally ranked highly in the target organism to drop dramatically in rank simply because it had no homologs or no information in the model. To prevent a bias against those genes, we replace $adj$ in Equation 1 by

$$adj^* = \begin{cases} \prod_{\text{organism } i}(1 - H_i Z_i(1 - v_i)) & \text{if } g_i \in \mathcal{H} \\ \text{median}(\{adj^* | g_j \in \mathcal{H}\}) & \text{if } g_i \notin \mathcal{H} \end{cases} \quad (2)$$

where $\mathcal{H}$ is the set of genes with at least one homolog with information under the given Endeavour model.

## 2.2 Orthology Mappings

Genes from one organism were mapped to their orthologs in another organism using homology scores specified by HomoloGene [4] or InParanoid [3]. As standard gene identifiers, CrossEndeavour uses Ensembl (`www.ensembl.org`) gene identifiers for human (ENSG), rat (ENSRNOG) and mouse (ENSMUSG), and the model organism database gene identifiers for worm (WormBase WBGene, `www.wormbase.org`), zebrafish (ZFIN ZDG-Gene, `zfin.org`) and fly (FlyBase FBgn, `flybase.org`) as shown in Table 1.

HomoloGene is a resource from NCBI [4] which provides homology scores between pairs of orthologous genes in fully sequenced organisms (Figure 3). The HomoloGene (Release 63) file was downloaded from `ftp.ncbi.nih.gov/pub/HomoloGene/current/homologene.xml.gz`. Gene identifier mappings were downloaded from Ensembl (Release 53, `www.ensembl.org/biomart`) using the biomart service (Table 1). When an identifier used within HomoloGene mapped to several identifiers of the type used in Endeavour, the information attached to the HomoloGene identifier was repeated for each corresponding Endeavour identifier.

InParanoid [3] identifies homology groups whereby each homolog from an organism is assigned a confidence value representing how closely it is related to the seed homolog which identifies the group (Figure 3). A single homology score for each pair of orthologous genes was generated by taking the minimum among the two confidence values for those two genes to the corresponding cluster. SQL tables for each pair of organisms of interest were downloaded from InParanoid (Release 6.1, `inparanoid.sbc.su.se/download/current/sqltables`). Gene identifier mappings were downloaded from Ensembl (Release 53) using the biomart

service (Table 1). Within biomart, the MGI identifiers were translated to ENSMUSG identifiers using the 'mgi_id' field which gave a one-to-one mapping. Mappings from FBpp to FBgn were made available from the curators of FlyBase (Release FB2009_03, `ftp.flybase.net/releases`).

| **Organism** | **Cross Endeavour Id Type** | **HomoloGene Id Type** | **InParanoid Id Type** |
|---|---|---|---|
| H. sapiens | ENSG | Entrez GeneId | ENSP |
| R. norvegicus | ENSRNOG | Entrez GeneId | ENSRNOP |
| C. elegans | WBGene | Entrez GeneId | WBGene |
| D. rerio | ZDG-Gene | Entrez GeneId | ZDG-Gene |
| D. melanogaster | FBgn | Entrez GeneId | FBpp |
| M. musculus | ENSMUSG | Entrez GeneId | MGI |

Table 1: Identifier Mappings

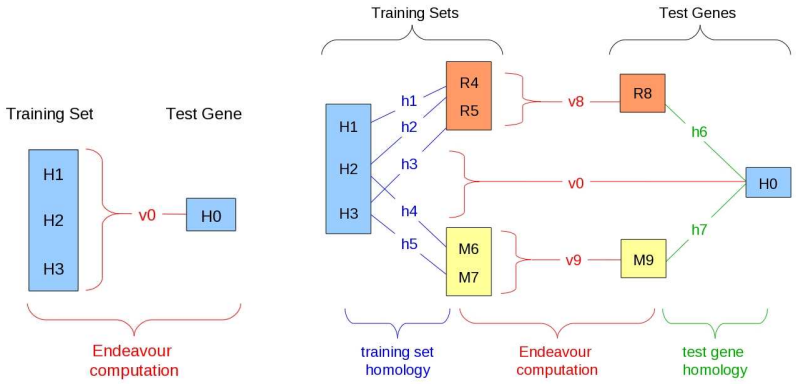
# References


[1] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., Carmeliet. P., Moreau, Y. (2006) Gene prioritization through genomic data fusion. *Nature Biotechnology*, May;24(5):537-44.

[2] Cozman, F. (2004) Axiomatizing Noisy-OR, Technical Report from Escola Politecnica da USP. BT/PMR/0409.

[3] O'Brien K.P., Remm M, Sonnhammer E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, Jan 1;33(Database issue):D476-80.

[4] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar. R., Federhen. S., Geer. L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott. D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L., Yaschenko, E. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 34(Database issue):D5-12.




(a) Endeavour (b) CrossEndeavour

**Incorporating Values from Alternate Organisms**

Training Set Size Discount

$Z$ = [*size of training set in alternate organism*]/[*size of training set in target organism*]

$Z_R$ = |{R4 R5}| / |{H1, H2, H3}| = 2/3

$Z_M$ = |{M6 M7}| / |{H1, H2, H3}| = 2/3

Orthology Strength Discount

$H$ = min[ test gene homology, median( training set homologies )]

$H_R$ = min[ h6, median( h1, h2, h3 )]

$H_M$ = min[ h7, median( h4, h5 )]

Combined Discount

$adj = \prod_{organism} [1 - Z*H*(1-v)]$

$adj = [1 - Z_R*H_R*(1-v8)] * [1 - Z_M*H_M*(1-v9)]$

**Adjusted Value in Target Organism**

$v' = v*adj$

$v0' = v0*adj$

(c) Value Adjustment

Figure 1: Model Similarity Value Calculation a) Endeavour: a value is computed for each test gene with respect to the training set. b) CrossEndeavour: training sets of orthologous genes are created for each alternate organism where pairwise homology scores are given between a training gene from the target organism and its maximum scoring ortholog in the alternate organism, a test gene is identified as the maximum scoring ortholog together with its specified homology score, and Endeavour calculates the analogous values for the orthologous test genes with respect to their corresponding orthologous training sets. c) Values in the alternate organisms are first discounted by a measure of homology and training set size and then values in the target organism are adjusted by a factor combining the alternative organism discounted values.

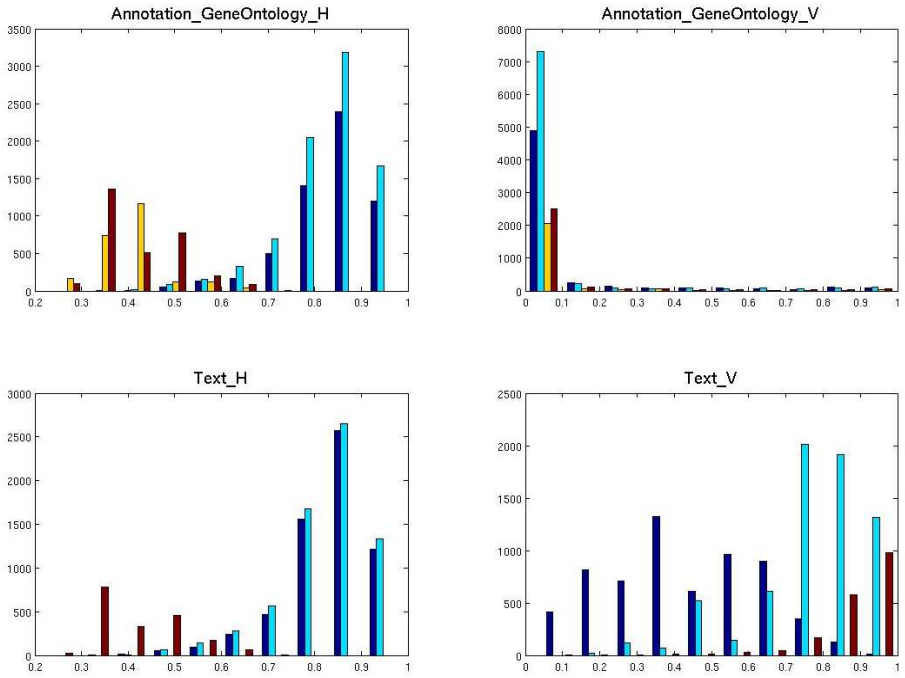



Figure 2: Representative distributions of homology strength discount $H$ and the model values $V$ for several Endeavour models. Using human as the target organism, set of bars in a histogram represents the values for the different alternate organisms, the number of which varies depending on the model.

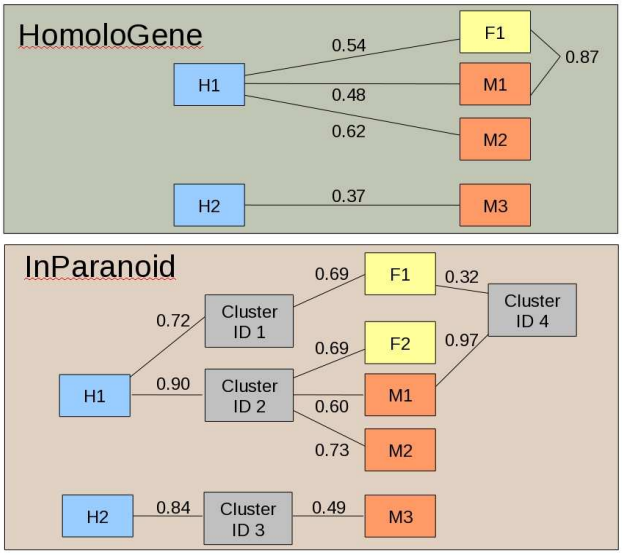



Figure 3: Diagram of Orthology Assignment Schemes