Bioinformatics Methods exam project: Automated function prediction by network-based protein ranking.

January 15, 2017

The goal of this project is the prediction of protein function by ranking proteins with respect to Gene Ontology (GO) terms [13], using network-based methods implemented in the *RANKS* R package [15], downloadable from *CRAN* (https://cran.r-project.org).

1 Data

Three networks representing the functional similarity between proteins are available:

1. The *DanXen* network encompasses *Danio rerio* (zebrafish) and *Xenopus laevis* (a small austral frog) proteins.
2. The *SacPomDis* network includes *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Dictyostelium discoideum* (unicellular eukaryotes).
3. The (*Dros*) network is reserved to *Drosophila melanogaster* (fruit-fly), a model organism for insects.

Each network is constructed by integrating 8 different sources of information from public databases (Table 1).

As class labels (groundtruth) for the proteins included in the integrated network the Gene Ontology BP, MF and CC experimental annotations extracted from the Swissprot database have been used (http://www.expasy.org/). The number of the terms (classes) in the three networks varies from 184 to 919 (CC - Cellular component), from 358 to 2195 (MF - Molecular Function), and from 2281 to 5037 (BP - Biological Processes). The number of nodes (proteins) in the *DanXen*, *Dros* and *SacPomDis* networks is respectively 6250, 3195 and 15836.

Availability of the data. All the data (networks and annotations) are downloadable from: `http://homes.di.unimi.it/DMB1617`. The directory `Nets` include the network files, where networks are represented as named symmetric

Database	Type of data
PRINTS [1]	Motif fingerprints
PROSITE [5]	Protein domains and families
Pfam [3]	Protein domain
SMART[8]	Simple Modular Architecture Research Tool (database annotations)
InterPro [9]	Integrated resource of protein families, domains and functional sites
Protein Superfamilies[4]	Structural and functional annotations
EggNOG [10]	Evolutionary genealogy of genes: Non-supervised Orthologous Groups
Swissprot [2]	Manually curated keywords describing the function of the proteins at different degrees of abstraction

Table 1: Data base and type of data used to construct the integrated protein similarity network

adjacency matrices, whose entries represent the weights of the underlying graph. Rows and columns have names corresponding to the protein identifiers. The directory `Annotations` include the annotations of the proteins. Rows are proteins (and correspond to the rows/columns of the networks in the Nets directory) and columns correspond to GO terms. If A is a matrix of annotations, $A[i, j] = 1$ if protein i is annotated with the GO term j, otherwise $A[i, j] = 0$. Note that some annotations files are numbered since they are split in parts: for instance. `danxen1.ann.BP.rda, danxen2.ann.BP.rda` are *DanXen* GO BP annotations in which GO terms are split in 2 separated sets.

2 Methods

Each group have to apply the following methods to the prediction of protein functions, all implemented in the R RANKS package:

1. The guilt-by-association method [14], implemented in the `do.GBA` function of the package.
2. The random walk algorithm [7], implemented in the `do.RW` function of the package.
3. The kernelized scored functions [11], implemented in the `do.RANKS` function of the package.

3 Experimental setup

To evaluate the generalization performance of the methods each group should apply a 5-fold cross-validation experimental setting.

To compare the results between methods the following metrics should be used: 1) "per class" ranking measures, i.e. the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision Recall

Curve (AUPRC); 2) a "per-example" metric, i.e. the multiple-label hierarchical F-score. If we indicate as $TP_j(t)$, $TN_j(t)$ and $FP_j(t)$ respectively the number of true positives, true negatives and false positives for the protein j at threshold t, we can define the "per-example" multiple-label precision $Prec(t)$ and recall $Rec(t)$ at a given threshold t as:

$$Prec(t) = \frac{1}{n}\sum_{j=1}^{n}\frac{TP_j(t)}{TP_j(t)+FP_j(t)} \qquad Rec(t) = \frac{1}{n}\sum_{j=1}^{n}\frac{TP_j(t)}{TP_j(t)+FN_j(t)} \quad (1)$$

where n is the number of examples (proteins). In other words $Prec(t)$ ($Rec(t)$) is the average multi-label precision (recall) across the examples. The F-score multi-label depends on t and according to CAFA2 experimental setting [6], the maximum achievable F-score ($Fmax$) should be adopted as the main multi-label "per-example" metric:

$$Fmax = max_t \frac{2Prec(t)Rec(t)}{Prec(t)+Rec(t)} \quad (2)$$

Note that AUROC is automatically computed by the `do.GBA`, `do.RW` and `do.RANKS` functions of the *RANKS* package (using the *PerfMeas* package from CRAN). To compute the AUPRC I suggest to use the *precrec* CRAN package. The F-score multi-label can be computed through the function `find.best.f` available in the source R file `F-hier.R`.

Note that the methods should be run only on classes having 5 or more annotations: the classes (GO terms) having less than 5 annotations should be removed.

4 Tasks to be performed by each group

One of the three networks listed in Section 1 and one the ontologies (BP, MF, CC) will be assigned to each group. Considering the large number of classes included in the BP ontology, the BP GO terms to be predicted have been split in two (*DanXen* and *Dros*) and four (*SacPomDis*) parts. Additionally also MF terms have been split in 4 parts with. *SacPomDis* Each group should predict the GO terms associated to each protein for only the assigned network and the assigned GO ontology or subpart of the GO ontology in case of BP predictions.

For instance the following data could be assigned to a group: the `sacpomdis.UA.net.filt.rda` network (*SacPomDis*) as functional network and `sacpomdis3.ann.BP.rda` for the corresponding BP annotations.

For the assigned network and GO ontology (or subpart of the GO ontology for BP), each group should provide:

1. The scores computed by each method (Section 2) for each protein and each GO term.

2. For each GO term the computed AUROC and the corresponding average values across classes (GO terms) for each of the considered methods.

3. For each protein the corresponding hierarchical F-score, as well as the average Hierarchical F-score, precision and recall across proteins.

4. For the random walk algorithm, consider at least the predictions obtained with 1 and 2 steps.

5. For the kernelized score functions (RANKS), consider at least the predictions obtained with 1 and 2 steps random walk kernels and both the average and the nearest-neighbour score functions.

6. A report including the experimental set-up, the results and a brief discussion of the obtained results. In particular use R graphics to present and analyze the results, and to compare the results achieved by the different methods.

7. The R code used for the experiments should be added as an appendix to the report.

5 Optional work

This part is not mandatory for the exam. The predictions for the *Homo sapiens* proteins can be computed in the same way as explained in the previous Section 4 using the Homo sapiens *STRING* functional network [12] (`homo.string10.rda`) and one of the corresponding GO term annotations for CC, MF or BP (`homo.ann.BP.rda, homo.ann.MF.rda, homo.ann.CC.rda`).

References

[1] Terri K. Attwood, Paul Bradley, Darren R. Flower, Anna Gaulton, Neil Maudling, AL Mitchell, G Moulton, A Nordle, K Paine, P Taylor, et al. Prints and its automatic supplement, preprints. *Nucleic acids research*, 31(1):400–402, 2003.

[2] The UniProt Consortium. Uniprot: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 2015.

[3] Robert D Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, et al. Pfam: clans, web tools and services. *Nucleic acids research*, 34(suppl 1):D247–D251, 2006.

[4] Julian Gough, Kevin Karplus, Richard Hughey, and Cyrus Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of molecular biology*, 313(4):903–919, 2001.

[5] Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian JA Sigrist. The prosite database. *Nucleic acids research*, 34(suppl 1):D227–D230, 2006.

[6] Y. Jiang et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(184), 2016.

[7] S. Kohler, S. Bauer, D. Horn, and P.N. Robinson. Walking the interactome for prioritization of candidate disease genes. *Am. J. Human Genetics*, 82(4):948–958, 2008.

[8] Ivica Letunic, Richard R Copley, Birgit Pils, Stefan Pinkert, Jörg Schultz, and Peer Bork. Smart 5: domains in the context of genomes and networks. *Nucleic acids research*, 34(suppl 1):D257–D260, 2006.

[9] Nicola J Mulder, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Virginie Buillard, Lorenzo Cerutti, Richard Copley, et al. New developments in the interpro database. *Nucleic acids research*, 35(suppl 1):D224–D228, 2007.

[10] Jean Muller, Damian Szklarczyk, Philippe Julien, Ivica Letunic, Alexander Roth, Michael Kuhn, Sean Powell, Christian von Mering, Tobias Doerks, Lars Juhl Jensen, et al. eggnog v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic acids research*, 38(suppl 1):D190–D195, 2010.

[11] M. Re, M. Mesiti, and G. Valentini. A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks. *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 9(6):1812–1818, 2012.

[12] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerte-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. String v10: protein2̆013protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 2014.

[13] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.

[14] W. Tian, L. Zhang, M. Tasan, F. Gibbons, O. King, J. Park, Z. Wunderlich, J.M. Cherry, and F.P. Roth. Combining guilt-by-association and guilt-by-profiling to predict Saccharomices cerevisiae gene function. *Genome Biology*, 9:S7, 2008.

[15] G. Valentini, Armano G., M. Frasca, J. Lin, M. Mesiti, and M. Re. RANKS: a flexible tool for node label ranking and classification in biological networks. *Bioinformatics*, 32:2872–2874, 2016.