Current and planned research

Giorgio Valentini
DI, Università degli Studi di Milano, Italy
e-mail: valentini@di.unimi.it

Current research

My main research activity is framed within machine learning, data mining and their application to relevant computational biology problems ranging from basic molecular biology to medicine. In particular in the three last years I focused on three main research lines:

1. Semi-supervised learning methods for the analysis of biomolecular networks.
2. Hierarchical ensemble methods for the automated function prediction of proteins
3. Methods to integrate complex and heterogeneous data.

1. Semi-supervised learning methods for the analysis of biomolecular networks. Most of the methods proposed for the analysis of graphs adopt local or global learning strategies to rank node labels or predict edges. On the contrary, we explored learning strategies based on the integration of both local and global learning strategies through kernelized score functions able to efficiently exploit the global topology of the network by using carefully designed graph kernels [5, 8, 10]. Another issue in node label ranking and classification of graphs is represented by the unbalance between positive and negative examples (this is particularly relevant for protein function prediction problems). To this end we designed

a novel class of parametrized Hopfield networks able to learn from the data the network parameters (neuron states and thresholds) taking into account the unbalance between positive and negative examples [7, 28, 30]. Methods based on kernelized score functions and parametrized Hopfield networks have been applied to the analysis of complex biomolecular networks for gene function prediction problems [8, 12, 13], gene disease prioritization [6,17,20] and drug repositioning (discovery of novel therapeutic indications for drugs originally designed for different pathologies) [5, 25].

2. Hierarchical ensemble methods for the automated function prediction (AFP) of proteins. Another relevant issue is posed by the hierarchical nature of the functional classes of proteins structured according to trees and directed acyclic graphs (DAG). To solve these complex classification problems I designed two ensemble methods able to exploit the hierarchical nature of AFP problems. The first one is based on the "True Path Rule" that governs hierarchical ontologies in biology and other application domains. Through carefully designed heuristics the method bottom-up propagates "positive predictions" from the leaves to the root of the tree-structured ontology, and top-down form root to leaves propagates "negative predictions", resulting in a two-way flow of information [14]. The second one is based on a classifier "bayesian-optimal" with respect to a hierarchical loss function. This approach allows to exchange information between classifiers associated to the nodes of the hierarchy according to a "message passing" strategy that simulates the optimal Bayesian classifier for this type of hierarchical problems [16]. I showed also that the synergy between hierarchical approaches, cost-sensitive and data integration methods is a key issue to solve complex AFP problems [11].

3. Methods to integrate structured data. In several domains, including bioinformatics, the integration of multiple and structured sources of data plays a key role. To this end we proposed several approaches for the construction and integration of networked data that have been applied to the automated function prediction of proteins, gene disease prioritiza-

tion and drug ranking and repositioning [5, 10, 19, 20, 23].

Other research lines lead to the development of supervised ensemble methods for the integration of multiple sources of biomolecular data in the context of cellular location [29] and gene function prediction problems [17], and mathematical models and clustering algorithms for other applications in computational biology [6, 9, 15, 31].

Planned research

Ongoing and future research can be divided into 5 main lines. The first two are strictly related to my current research and the last three represent novel research lines.

1. *Ensembles of semi-supervised models for automated function prediction*. In the context of the ongoing international challenge CAFA2 (Critical Assessment of Function Annotation experiments) we developed novel methods for AFP based on kernelized score functions, by experimenting test-and select ensemble methods to exploit different graph kernels and random walk kernels with different number of steps to better fit the specific functional characteristics of each GO term to be predicted.

2. *Design, analysis and efficient software implementation of hierarchical ensemble methods for DAG structured ontologies*. Another ongoing research line is represented by the design, analysis and efficient software implementation of hierarchical ensemble methods for DAG structured ontologies, with relevant applications to the prediction of protein functions in the Gene Ontology or human phenotypes in the context of the Human Phenotype Ontology.

3. *Network-based phenotype and outcome prediction and biomarker selection*. In the context of phenotype and outcome prediction problems using a set of selected biomarkers (e.g. gene expression signatures or allelic configurations of SNPs), supervised inductive models do not

explicitly take into account the functional or the genetic relationships between individuals. To overcome this problem, I plan to explore a novel "Network Medicine" based approach by modeling networks in the "patient space" instead of the "biomarker space", by constructing networks of patients on the basis of their functional or genetic similarities (e.g. by considering their expression profiles or their allelic configurations of SNPs), and then by applying semi-supervised transductive method to predict the phenotype or the clinical outcome of patients, and to select statistically relevant biomarkers.

4. *Big-data analysis in Computational Biology using secondary memory-based technologies*. Very recently I started a novel research line for "Big Data" analysis based on novel algorithmic approaches and on the exploitation of novel technologies for relevant applications in computational biology and medicine. In particular I am studying algorithms and technologies based on computer secondary memory to process huge graphs that cannot be loaded in primary memory, with the objective of analyzing huge "omics" data stored in large public databases with relevant application in basic molecular biology and medicine. Preliminary results yielded promising results in the context of multi-species protein function prediction problems [1, 21].

5. *Multi-task learning in Computational Biology*. As several classification tasks in Computational Biology are inherently related, Multitask Learning (MTL) is a natural approach for this domain. I plan to extend state-of-the-art MTL methods in three directions. First, to embed evolutionary relationships between species and hierarchies over functional classes in the context of automated function prediction methods. Second, to learn phenotypic abnormalities in human diseases, with a significant potential impact in molecular biology and medicine. Third, to extend transductive relational learning models in order to perform network-based multi-learning tasks for relevant problems in computational biology.

References

[1] M. Mesiti, M. Re, and G. Valentini. Think globally and solve locally: secondary memory-based network learning for automated multi-species function prediction. *GigaScience*, 3:5, 2014.

[2] G. Valentini, A. Paccanaro, H. Caniza, A. Romero, and M. Re. An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artificial Intelligence in Medicine*, 2014. (in press).

[3] H. Caniza, A. Romero, S. Heron, H. Yang, A. Devoto, M. Frasca, M. Mesiti, G. Valentini, and A. Paccanaro. GOssTo: a user-friendly stand-alone and web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics*, 2014. (in press).

[4] G. Valentini. Hierarchical Ensemble Methods for Protein Function Prediction. *ISRN Bioinformatics*, 2014(Article ID 901419):34 pages, 2014.

[5] M. Re and G. Valentini. Network-based Drug Ranking and Repositioning with respect to DrugBank Therapeutic Categories. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(6):1359–1371, 2013.

[6] I. Cattinelli, G. Valentini, E. Paulesu, and A. Borghese. A Novel Approach to the Problem of Non-uniqueness of the Solution in Hierarchical Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 24(7):1166–1173, 2013.

[7] M. Frasca, A. Bertoni, M. Re, and G. Valentini. A neural network algorithm for semi-supervised node label learning from unbalanced data. *Neural Networks*, 43:84–98, 2013.

[8] 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.

[9] A. Beghini, F. Corlazzoli, L. Del Giacco, M. Re, F. Lazzaroni, M. Brioschi, G. Valentini, F. Ferrazzi, A. Ghilardi, M. Righi, M. Turrini, M. Mignardi, C. Cesana, V. Bronte, M. Nilsson, Morra E., and R. Cairoli. Regeneration-associated Wnt signaling is activated in long-term reconstituting AC133bright acute myeloid leukemia cells. *Neoplasia*, 14(12):1236–1248, 2012.

[10] M. Re and G. Valentini. Cancer module genes ranking using kernelized score functions. *BMC Bioinformatics*, 13(Suppl 14/S3), 2012.

[11] N. Cesa-Bianchi, M. Re, and G. Valentini. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Machine Learning*, 88(1):209–241, 2012.

[12] M. Re, M. Mesiti, and G. Valentini. Drug repositioning through pharmacological spaces integration based on networks projection. *EMBnet.journal*, 18(Supplement A):30–31, 2012.

[13] M. Frasca, A. Bertoni, and G. Valentini. Regularized network-based algorithm for predicting gene functions with high-imbalanced data. *EMBnet.journal*, 18(Supplement A):41–42, 2012.

[14] G. Valentini. True Path Rule hierarchical ensembles for genome-wide gene function prediction. *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 8(3):832–847, 2011.

[15] M. Muselli, A. Bertoni, M. Frasca, A. Beghini, F. Ruffino, and G. Valentini. A mathematical model for the validation of gene selection methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:1385–1392, 2011.

[16] N. Cesa-Bianchi and G. Valentini. Hierarchical cost-sensitive algorithms for genome-wide gene function prediction. *Journal of Machine Learning Research, W&C Proceedings, Machine Learning in Systems Biology*, 8:14–29, 2010.

[17] M. Re and G. Valentini. Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. *Journal of Machine Learning Research, W&C Proceedings, Machine Learning in Systems Biology*, 8:98–111, 2010.

[18] O. Okun, G. Valentini, and M. Re. *Ensembles in Machine Learning Applications*, volume 373 of *Studies in Computational Intelligence*. Springer, Berlin, 2011.

[19] M. Frasca, A. Bertoni, and G. Valentini. An unbalance-aware network integration method for gene function prediction. In *MLSB 2013 - Machine Learning for Systems Biology - Berlin*, Berlin, Germany, 2013.

[20] G. Valentini, A. Paccanaro, H.C. Vierci, A.E. Romero, and M. Re. Network integration boosts disease gene prioritization. In *Network Biology SIG 2013 ISMB 2013*, Berlin, Germany, 2013.

[21] M. Mesiti, M. Re, and G. Valentini. Scalable Network-based Learning Methods for Automated Function Prediction based on the Neo4j Graph-database. In *Automated Function Prediction SIG 2013 - ISMB 2013*, Berlin, Germany, 2013.

[22] H.C. Vierci, A.E. Romero, S. Heron, H. Yang, M. Frasca, M. Mesiti, G. Valentini, and A. Paccanaro. GOssTo & GOssToWeb: user-friendly tools for calculating semantic similarities on the Gene Ontology. In *Bio-Ontologies SIG 2013 - ISMB 2013*, Berlin, Germany, 2013.

[23] M. Re, M. Mesiti, and G. Valentini. Comparison of early and late omics data integration for cancer modules gene ranking. In *NETTAB 2012 Workshop on Integrated Bio-Search*, Como, Italy, 2012.

[24] M. Re and G. Valentini. Random walking on functional interaction networks to rank genes involved in cancer. In *2nd Artificial Intelligence Applications in Biomedicine Workshop, in: AIAI 2012 - Artificial Intelligence Applications and Innovations*, IFIP AICT Series, pages 66–75. Springer, 2012.

[25] M. Re and G. Valentini. Large scale ranking and repositioning of drugs with respect to drugbank therapeutic categories. In *International Symposium on Bioinformatics Research and Applications (ISBRA 2012)*, volume 7292 of *Lecture Notes in Computer Science*, pages 225–236. Springer, 2012.

[26] M. Re and G. Valentini. Ensemble methods: a review. In *Advances in Machine Learning and Data Mining for Astronomy*, Data Mining and Knowledge Discovery, pages 563–594. Chapman & Hall, 2012.

[27] M. Re and G. Valentini. Genes prioritization with respect to Cancer Gene Modules using functional interaction network data. In *NETTAB 2011 Workshop on Clinical Bioinformatics*, Pavia, Italy, 2011.

[28] A. Bertoni, M. Frasca, and G. Valentini. Cosnet: a cost sensitive neural network for semi-supervised learning in graphs. In *European Conference on Machine Learning, ECML PKDD 2011*, volume 6911 of *Lecture Notes on Artificial Intelligence*, pages 219–234. Springer, 2011.

[29] A. Rozza, G. Lombardi, M. Re, E. Casiraghi, G. Valentini, and P. Campadelli. A novel ensemble technique for protein subcellular location prediction. In *Ensembles in Machine Learning Applications*, volume 373 of *Studies in Computational Intelligence*, pages 151–167. Springer, 2011.

[30] M. Frasca, A. Bertoni, and G. Valentini. A cost-sensitive neural algorithm to predict gene functions using large biological networks. In *Network Biology SIG: On the Analysis and Visualization of Networks in Biology, ISMB 2011*, Wien, 2011.

[31] A. Bertoni, M. Re, F. Sacca, and G. Valentini. Identification of promoter regions in genomic sequences by 1-dimensional constraint clustering. In *Neural Nets WIRN11 - Proceedings*, volume 234 of *Frontiers in Artificial Intelligence and Applications*, pages 162–169, 2011.