Current research activity

Giorgio Valentini
AnacletoLab - Computational Biology and Bioinformatics
DI, Dipartimento di Informatica
Università degli Studi di Milano, Italy
e-mail: valentini@di.unimi.it

Machine Learning and Computational Biology represent my main research area. My main current research lines can be schematically summarized as follows:

1. Machine Learning for Personalized Genomic Medicine
2. Machine Learning for biomolecular network analysis and Network Medicine
3. Hierarchical ensemble methods for the prediction of biomolecular function and property of proteins and genes using structured ontologies
4. Big-data analysis in Computational Biology using parallel and secondary memory-based technologies
5. Machine Learning for the integration of complex biomolecular data

Machine Learning for Personalized Genomic Medicine

The identification of genetic variants associated with human diseases represents one of the core challenges in precision medicine, and requires the design and application of a new generation of machine learning-based

prediction methods able to prioritize potential "deleterious′ variants (i.e. causative or otherwise linked with disease risk) among the huge amount of neutral variants that represent natural genetic variation present in individuals.

Most of state-of-the-art ML-based methods do not adopt specific imbalance-aware learning techniques to deal with imbalanced data that naturally arise in several genome-wide variant scoring problems, thus resulting in a significant reduction of sensitivity and precision. We developed *hyperSMURF* (hyper-ensemble of SMOTE under-sampled random forests), a novel method that adopts imbalance-aware learning strategies based on resampling techniques and a hyper-ensemble approach to deal with highly imbalanced genomic data [1, 2]. This machine learning approach has been successfully applied as part of Genomiser, a software tool that uses both genotypic and phenotypic information, to discover variants in both coding and non coding regulatory regions associated with specific genetic Mendelian diseases [3].

Fine tuning of learning parameters of hyperSMURF may lead to significantly better results [4], and we are developing a High Performance Computing parallel version of this hyper-ensemble method in the context of the LISA project HyperGeV - Detection of Deleterious Genetic Variation through Hyper-ensemble Methods.

Machine Learning for biomolecular network analysis and Network Medicine

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, 6, 7, 8]. 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 [9, 10, 11, 12, 13].

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 [7, 14, 15], gene disease prioritization and drug repositioning (discovery of novel therapeutic indications for drugs originally designed for different pathologies) [6, 16].

Another ongoing research line is represented by the application of Game Theory concepts to structured node label ranking problems in computational biology. For instance, we were able to learn the set of GO terms associated with a protein as a whole , by representing protein function prediction as a graph-transduction game, where both the functional and structural similarities between proteins and the similarities between GO terms are taken into account and embedded in the same semi-supervised learning problem [17].

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 am exploring 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 [18].

Another research line is represented by novel approaches for the web visualization and analysis of biomolecular networks [19].

Hierarchical ensemble methods for the prediction of biomolecular function and property of proteins and genes using structured ontologies

The characterization of protein and gene functions is based upon specific ontologies, such as the Gene Ontology or the Human Phenotype Ontology [20].

The prediction of the protein function is a problem characterized by several issues [21], but one of the main is represented by the hierarchical structure of the taxonomies currently used to classify proteins and genes. To solve these complex classification problems I designed 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 from root to leaves propagates "negative predictions", resulting in a two-way flow of information [22]. 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 [23]. I showed also that the synergy between hierarchical approaches, cost-sensitive and data integration methods is a key issue to solve complex AFP problems [24].

Another significant application of hierarchical ensemble methods is represented by the prediction of abnormal human phenotypes using the Human Phenotype Ontology. In this context we proposed novel heirarchical ensemble methods for DAG-structured ontologies that achieved state-of-the-art results in this field [25, 26, 27, 28]. Preliminary results of the application of the the same methods to the GO ontology showed very promising results [29].

Big-data analysis in Computational Biology using parallel and secondary memory-based technologies

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 represents anothee research line in this context. 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. Experimental results yielded promising results in the context of multi-species protein function prediction problems [30, 31, 32]. Another ongoing research line uses GPU technology to speed-up computation and improve results in big biomolecular networks [33] with application to GO terms prediction using STRING networks constructed by putting together multiple species [34].

Machine Learning for the integration of complex biomolecular 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 prioritization and drug ranking and repositioning [35, 36, 37, 6, 8, 38, 39, 40].

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 [41] and gene function prediction problems [42, 43, 44], and mathematical and semantic models and clustering and networks integration and visualization algorithms for other applications in computational biology [45, 46, 47, 48, 49, 50].

References

[1] M. Schubach, M. Re, PN Robinson, and G. Valentini. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Scientific Reports*, 7(2959), 2017.

[2] M. Schubach, M. Re, PN Robinson, and G. Valentini. "variant relevance prediction in extremely imbalanced training sets". *F1000Research*, 6(1392), 2017.

[3] D. Smedley, M. Schubach, J. Jacobsen, S. Khler, T. Zemojtel, M. Spielmann, M. Jager, H. Hochheiser, N. Washington, J. McMurry, M. Haendel, C. Mungall, S. Lewis, T. Groza, G. Valentini, and P. Robinson. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *The American Journal of Human Genetics*, 99(3):595–606, 2016.

[4] A Petrini, M Schubach, M. Re, M Frasca, M Mesiti, G Grossi, T Castrignano, PN Robinson, and Valentini G. "parameters tuning boosts hypersmurf predictions of rare deleterious non-coding genetic variants". *PeerJ Preprints*, 5:e3185v1, 2017.

[5] 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.

[6] 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.

[7] 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.

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

[9] M. Frasca, S. Bassis, and G. Valentini. Learning node labels with multi-category hopfield networks. *Neural Computing and Applications*, 27(6):1677–1692, 2016.

[10] M. Frasca and G. Valentini. COSNet: An R package for label prediction in unbalanced biological networks. *Neurocomputing*, 237:397 – 400, 2017.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] S. Vascon, M. Frasca, R. Tripodi, G. Valentini, and M. Pelillo. Protein Function Prediction as a Graph-Transduction Game. *Pattern Recognition Letters*, 2018. (in press).

[18] J. Gliozzo, M. Notaro, A. Petrini, P. Perlasca, M. Mesiti, E. Casiraghi, , G. Grossi, M. Frasca, M. Re, A. Paccanaro, and G. Valentini. Modeling biomolecular profiles in a graph-structured sample space for clinical outcome prediction with melanoma and ovarian cancer patients. In *BITS 2017, Bioinformatics Italian Society Meeting*, Cagliari Italy, 2017.

[19] P Perlasca, G Valentini, M Frasca, and M Mesiti. Multi-species protein function prediction: Towards web-based visual analytics. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, iiWAS '16, pages 489–493, New York, NY, USA, 2016. ACM.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] Marco Notaro, Max Schubach, Peter Robinson, and Giorgio Valentini. "prediction of human phenotype ontology terms by means of hierarchical ensemble methods". *BMC Bioinformatics*, 18(1):449, Oct 2017.

[26] M. Notaro, M. Schubach, P.N. Robinson, and G. Valentini. Ensembling Descendant Term Classifiers to Improve Gene - Abnormal Phenotype Predictions. In *CIBB 2017, The 14th International Conference on Bioinformatics and Biostatistics*, Cagliari Italy, 2017.

[27] G. Valentini, S. Kohler, M. Re, M. Notaro, and P.N. Robinson. Prediction of human gene - phenotype asociations by exploiting the hierarchical structure of the human phenotype ontology. In *IWBBIO 2015 (3rd International Work-Conference on Bioinformatics and Biomedical Engineering)*, volume 9043 of *Lecture Notes in Bioinformatics*, pages 66–77. Springer, 2015.

[28] M. Notaro, M. Schubach, P. Robinson, and G. Valentini. Predicting new relationships between genes and human phenotype ontology terms. In *26th International Conference on Intelligent Systems for Molecular Biology (ISMB), Chicago (USA)*, 2018.

[29] P.N. Robinson, M. Frasca, S. Kohler, M. Notaro, M. Re, and G. Valentini. A hierarchical ensemble method for DAG-structured taxonomies. In *Multiple Classifier Systems - MCS 2015*, volume 9132 of *Lecture Notes in Computer Science*, pages 15–36. Springer, 2015.

[30] J. Lin, M. Mesiti, M. Re, and G. Valentini. Within network learning on big graphs using secondary memory-based random walk kernels. In *Complex Networks & Their Applications V*, volume 693 of *Studies in Computational Intelligence*, pages 235–245. Springer, 2017.

[31] 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.

[32] 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.

[33] A Petrini, M. Notaro, J. Gliozzo, M. Perlasca, M Mesiti, G. Valentini, G Grossi, and M Frasca. A GPU-based algorithm for fast node label learning in large and unbalanced biomolecular networks. *BMC Bioinformatics*, 2018. (accepted).

[34] A. Petrini, M. Notaro, J. Gliozzo, G. Valentini, G. Grossi, and M. Frasca. Speeding up node label learning in unbalanced biomolecular networks through a parallel and sparse GPU-based Hopfield model. In *BITS 2017, Bioinformatics Italian Society Meeting*, Cagliari Italy, 2017.

[35] M. Frasca, J.F. Fontaine, G. Valentini, M. Mesiti, M. Notaro, D. Malchiodi, and Andrade-Navarro M.A. Disease Genes must Guide Data Source Integration in the Gene Prioritization Process. In *CIBB 2017, The 14th International Conference on Bioinformatics and Biostatistics*, Cagliari Italy, 2017.

[36] M. Frasca, A. Bertoni, and G. Valentini. Unipred: Unbalance-aware network integration and prediction of protein functions. *Journal of Computational Biology*, 22(12):1057–1074, 2015.

[37] 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*, 61(2):63–78, 2014.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] M. Re and G. Valentini. Integration of heterogeneous data sources for gene function prediction using Decision Templates and ensembles of learning machines. *Neurocomputing*, 73(7-9):1533–37, 2010.

[44] M. Re and G. Valentini. Noise tolerance of Multiple Classifier Systems in data integration-based gene function prediction. *Journal of Integrative Bioinformatics*, 7(3:139), 2010.

[45] 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*, 30(15), 2014.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] C.T. Ba, E. Casiraghi, M. Frasca, J. Gliozzo, G. Grossi, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini, M Re, and G. Valentini. A Graphical Tool for the Exploration and Visual Analysis of Biomolecular Networks. In *CIBB 2018 - The 15th International Conference on Bioinformatics and Biostatistics*, Lisboa Portugal, 2018.