

FunCat Functional Inference with Belief Propagation and Feature Integration

Dimitrij Surmeli,^{1,2} Oliver Ratmann,³ Hans-Werner Mewes,^{1,4} Igor V. Tetko^{1,*}

¹Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, Neuherberg, Germany

²BrainLAB, Kapellenstrasse 12, Feldkirchen, Germany

³Centre for Biostatistics, Imperial College, London, United Kingdom

⁴Chair f. Genome-oriented Bioinformatics, Technische Universität München, Life and Food Science Center Weihenstephan, Am Forum 1, D-85354 Freising-Weihenstephan, Germany

This is preprint of article:

Surmeli D, Ratmann O, Mewes H, Tetko IV. FunCat functional inference with belief propagation and feature integration. Comput Biol Chem. 2008 Oct ;32(5):375-377.

<http://dx.doi.org/10.1016/j.compbiolchem.2008.06.004>

Address for correspondence:

Dr. Igor V. Tetko

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Neuherberg, D-85764, Germany

Ingolstädter Landstraße 1,

85764 Neuherberg, Germany

Telephone: +49-89-3187-3575

Fax: +49-89-3187-3585

e-mail: itetko@vcclab.org

ABSTRACT

Pairwise comparison of sequence data is intensively used for automated functional protein annotation, while graphical models emerge as promising candidates for an integration of various heterogeneous features. We designed a model, termed heterogeneous Relational Markov Network (hRMN) that integrates different genomic features and implemented a variant of Belief Propagation for functional annotation transfer. hRMN allows the assignment of multiple functional categories while avoiding common problems in annotation transfer from heterogeneous datasets, such as independence and identical distribution of the datasets. We benchmarked this system with large-scale annotation transfer (based on the MIPS FunCat ontology) to proteins of the prokaryotes *Bacillus subtilis*, *Helicobacter pylori*, *Listeria monocytogenes*, and *Listeria innocua*. hRMN consistently outperformed two competitors for annotation of four bacterial genomes. The developed code is available for download at <http://mips.gsf.de/proj/bfab/hRMN.html>.

Keywords: automatic functional annotation; belief propagation; bacterial genomes; heterogeneous data sources; probabilistic graphical models

INTRODUCTION

The assignment of function to protein sequences remains an essential goal at the very heart of the bioinformatics enterprise (Frishman, 2007, Gattiker, et al., 2003). Given the volume of genomic sequence data, the only feasible, reliable annotation relies on the combination of computational tools based on features of proteins, most prominently homology by sequence similarity, with the laborious endeavor of manually curated annotation. Numerous efforts created compilations of annotated data, like PDB, SwissProt, KEGG, InterPro, GO and FunCat, based to varying degrees on expert curation and automated annotation by computational methods such as sequence similarity, 3D structure, protein-protein interaction, gene expression and recently, 3D motifs among homologues. All these widely used resources cover different or complementary aspects of gene functions, such that an integration of such features naturally emerges as a challenge. Integration is expected to improve on predictive power, coverage and to increase the reliability of the assignments.

Probabilistic graphical models (Jordan, 1998, Pearl, 1988) are promising candidates to this challenge. First, they can cope with incompleteness, variability, conditionality, noise and uncertainty inherent in annotation transfer. Second, they provide a probabilistic framework for coherent integration (Gerstein, et al., 2002) of heterogeneous data "enabling conclusions that are not supported by either data source independently" (Friedman, 2004). Third, such systems are scalable, and easily extended to new data sources. Finally, their output maybe filtered by confidence thresholds or probability cut-offs for a desired level of sensitivity or specificity.

Here, we propose a compound probabilistic graphical model, termed heterogeneous Relational Markov Network (hRMN), for large-scale automated annotation transfer. Notably, this system (1) avoids common problems to such computational methods, such as inappropriate independence assumptions among features and/or categories, and a general lack of consistently propagated noise and missing data (Pachter and Sturmfels, 2004)), and (2) integrates heterogeneous datasets. hRMN also allows us to combine a multitude of relational features, describing fundamentally different species relationships between proteins, e.g. domain composition,

gene expression profiles, and sequence similarity scores, as well as only slightly differing instantiations of the same relational feature, e.g. sequence similarity scores obtained with different BLOSUM matrices, or annotations from different ontologies. We stipulate only that the features be expressible as pairwise relations between proteins. As an application, we benchmarked our system on the assignment of functional labels to proteins of *B. subtilis*, *H. pylori*, *L. monocytogenes*, and *L. innocua* according to the MIPS Functional Catalog (Ruepp, et al., 2004).

METHOD

The Bayesian paradigm is a powerful probabilistic framework for making inference, and is here deployed to annotate M proteins m of unknown function X_m from proteins of known function via heterogeneous feature sets relating the proteins of unknown function to those of unknown function. From this data D , focus is on the joint posterior

$$p(X = x | D) \tag{1}$$

for the entire proteome, which is proportional to the likelihood $p(D | X = x)$ and the prior $p(X = x)$ via Bayes' Theorem. Here, datasets are each viewed as "objects" o that are composed of a Markov Net, representing pairwise relationships among all proteins, and a number of attributes that characterize the dataset. To increase the power in annotation transfer by multiple lines of evidence, sets of pairwise protein relations are combined into a Markov network. Each protein, either labeled (known function) or unlabeled, corresponds to a node in an undirected graph. An edge is created between two proteins for which a relationship score is available. The inclusion of additional attributes allows us to evaluate the contribution of each Markov net to overall annotation transfer. A Bayesian network on the attributes of all objects is employed to compute "object weights" η , accounting for the dependencies across all Markov nets. This way, we introduce a flexible and extensible framework to account for different interpretations of the features in varying (attributed) contexts. For instance, the transfer of an annotation referring to a certain subcellular localization may be justified for paralogous sequences, yet not between sequences from different species. Figure 1 exemplifies the full system.

To solve equation (1), we propose a weighted counting model:

$$p(X = x | D) \propto \prod_m \phi_m(X_m = x_m) \prod_{\langle m, m' \rangle} \psi_{m, m'}(X_m, X_{m'}) = \exp - \left(\sum_{m, x_m} \phi_m(X_m = x_m) + \sum_{o \in O} \sum_{\langle m, m' \rangle} \eta_{m, m'}^o N_{m, m'} \right), \quad (2)$$

following the above notation; $N_{m, m'}$ counts the pairwise annotations $x_m, x_{m'}$ within the first order neighborhood $\langle m, m' \rangle$ of m . Each node is assigned a potential ϕ_m , which can be interpreted as the prior probability in the realized label assignment x_m for protein m . The compatibility functions $\psi_{m, m'}$ can be regarded as edge weights between nodes whose values are derived from the feature score, reflecting the strength of relationship between the respective two proteins (Lu, et al., 2005). Object weights capture interrelations between heterogeneous sources. Intuitively, integration maybe understood by coloring all edges, one color for each object, and then superimposing all Markov nets. Notably, in the superposition, each edge is drawn with a specific width corresponding to its weight $\eta_{m, m'}^o$. We close this section with two final observations: equation (2) explicitly correlates annotations to similar proteins, and many cycles are expected in each Markov net. Generalized Belief Propagation (GBP) is a suitable inference algorithm (Yedidia, et al., 2001, Yedidia, et al., 2004) in this situation.

RESULTS

To benchmark hRMN for automatic functional annotation, we used a dataset of four bacterial genomes, *B. subtilis*, *H. pylori*, *L. monocytogenes*, and *L. innocua* as proposed previously (Tetko, et al., 2005a). To illustrate the power of hRMN on integrating similar relational features, we computed sequence similarity scores based on FASTA, BLAST, PSIBLAST, and INTERPRO domains.

We implemented two competing, neighborhood-based algorithms, subsequently termed "Neighbor's Best List"(NBL) and "Closest Neighbor" (CN). In brief, both approaches included the following steps. NBL: (1) create a list of proteins with sparse mutual distances; (2) for any given protein, collect labels from all neighboring proteins and (3) count the occurrences of each label; (4) assign the labels most often agreed upon, if more than half of the labels agree and

otherwise do not assign any label. CN: (1) assign the labels of the closest neighboring protein and (2), in case of integration, assign only the labels on which the closest neighboring proteins among all feature sets agree. CN, traditionally used in annotation transfer based on sequence similarity, was chosen as a baseline comparison, embodying a pure form of 'guilt by association'. NBL resembles the underlying counting model in hRMN, thus serving to test the performance of hRMN's belief propagation (Yedidia, et al., 2001, Yedidia, et al., 2004).

TO ASSESS PERFORMANCE, we used a two-fold approach. First, we performed a five-fold cross-validation. Second, we predicted the labels of all sequences of each of the four genomes from the three other genomes.

We first evaluated the performance of the hRMN and two other algorithms integrating over the 4 feature sets, for all FunCat hierarchy levels (but report only the performance for the most specific level), and the performance based on each of the 4 feature sets in terms of cross-validation as shown in Figure 2, and the supplementary Figures S1-S3 (available on-line at <http://mips.gsf.de/proj/bfab/hRMN.html>). For any feature, hRMN outperforms its two competitors in both sensitivity and overlap (recognition rate). This holds also true for the integration of similarity scores obtained by InterPro and BLAST, the feature performing best as measured by cross-validation. Moreover, in this last case, hRMN does not exhibit a performance drop common to its competitors.

Turning to genome-wide prediction, hRMN, when compared to CN, achieved improved prediction based on each of the individual features, and for the integration of Blast and InterPro domains. Notably, integrating BLAST similarity scores with InterPro domains performed at least as well as any of the features individually, even without any weighting or other tuning. The hard dip for FunCat category '04.*' in the Figures S2-S3 is mainly due to the very rare occurrences of these categories, so it is hard to transfer. We believe that an implementation of the Bayesian net on dataset attributes to weight relational evidence will further boost the predictive power of this first implementation. As a next step, it is possible to integrate more complementary heterogeneous features such as protein interactions, protein structure, motif scores and gene neighborhood and orientation (Wu, et al., 2008). Our finding that features differ significantly in their predictive accuracy for some FunCat categories indicates that both the implementation of ob-

ject weights and the integration of further datasets are likely to improve on the predictive power of hRMN in annotation transfer of whole genomes. It is also possible to include category-specific scores, and to supplement hRMN with edge and label permutation tests (Balasubramanian, et al., 2004), and/or mutual information scores for labels and features (Lu, et al., 2005). Recently, we also applied hRMN and Super Paramagnetic Clustering algorithm (Blatt, et al., 1996, Tetko, et al., 2005b) to the large-scale automated annotation of mouse protein functions (Ruepp, et al., 2006). The developed system integrated manual annotations data from public (SwissProt) and commercially available data sources (Biomax Informatics AG). The marginal beliefs scores calculated with hRMNN provided the confidence of the automated annotation, which was further used during the manual curation of the results. However, the previous article did not describe the details of the developed algorithm neither performed any benchmarking, that has been done in this study.

In summary, this article described a new graph-based algorithm for annotation of protein sequences and benchmarked it using four bacterial genomes. The calculated results demonstrated high performance of the proposed method compared to traditional approaches used in the field. The source code of the algorithm is freely available and can be applied/used by other users to annotate their data or to develop new approaches.

ACKNOWLEDGEMENTS

This study was partially supported by the DFG grant TE 380/1-1.

REFERENCES

- [1] Balasubramanian R., LaFramboise T., Scholtens D., Gentleman R., A graph-theoretic approach to testing associations between disparate sources of functional genomics data, *Bioinformatics* 20 (2004) 3353-3362.
- [2] Blatt M., Wiseman S., Domany E., Superparamagnetic clustering of data, *Physical Review Letters* 76 (1996) 3251-3254.
- [3] Friedman N., Inferring Cellular Networks Using Probabilistic Graphical Models, *Science* 303 (2004) 799-805.
- [4] Frishman D., Protein annotation at genomic scale: the current status, *Chem Rev* 107 (2007) 3448-3466.

- [5] Gattiker A., Michoud K., Rivoire C., Auchincloss A.H., Coudert E., Lima T., Kersey P., Pagni M., Sigrist C.J., Lachaize C., Veuthey A.L., Gasteiger E., Bairoch A., Automated annotation of microbial proteomes in SWISS-PROT, *Comput Biol Chem* 27 (2003) 49-58.
- [6] Gerstein M., Lan N., Ronald J., PERSPECTIVES : PROTEOMICS: Integrating Interactomes, *Science* 295 (2002) 284-287.
- [7] Jordan M.I., *Learning in graphical models*, Kluwer Academic Publishers, Boston, MA, USA, 1998.
- [8] Lu L.J., Xia Y., Paccanaro A., Yu H., Gerstein M., Assessing the limits of genomic data integration for predicting protein networks, *Genome Research* 15 (2005) 945-953.
- [9] Pachter L., Sturmfels B., Parametric inference for biological sequence analysis, *PNAS* 101 (2004) 16138-16143.
- [10] Pearl J., *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufman Publishers, Inc, San Mateo, CA, USA, 1988.
- [11] Ruepp A., Doudieu O.N., van den Oever J., Brauner B., Dunger-Kaltenbach I., Fobo G., Frishman G., Montrone C., Skornia C., Wanka S., Rattei T., Pagel P., Riley L., Frishman D., Surmeli D., Tetko I.V., Oesterheld M., Stumpflen V., Mewes H.W., The Mouse Functional Genome Database (MfunGD): functional annotation of proteins in the light of their cellular context, *Nucleic Acids Res* 34 (2006) D568-571.
- [12] Ruepp A., Mewes H.W., Prediction and Classification of Protein Functions, *Drug Discov. Today: Tech* 3 (2006) 145-151.
- [13] Ruepp A., Zollner A., Maier D., Albermann K., Hani J., Mekrejs M., Tetko I., Guldener U., Mannhaupt G., Munsterkotter M., Mewes H.W., The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Res* 32 (2004) 5539-5545.
- [14] Tetko I.V., Brauner B., Dunger-Kaltenbach I., Frishman G., Montrone C., Fobo G., Ruepp A., Antonov A.V., Surmeli D., Mewes H.W., MIPS bacterial genomes functional annotation benchmark dataset, *Bioinformatics* 21 (2005a) 2520-2521.
- [15] Tetko I.V., Facius A., Ruepp A., Mewes H.W., Super paramagnetic clustering of protein sequences, *BMC Bioinformatics* 6 (2005b) 82.
- [16] Wu H., Mao F., Olman V., Xu Y., On application of directons to functional classification of genes in prokaryotes, *Computational Biology and Chemistry* 32 (2008) 176-184.
- [17] Yedidia J.S., Freeman W.T., Weiss Y., *Understanding Belief Propagation and Its Generalizations*, Mitsubishi Electric Research Laboratories, 2001.
- [18] Yedidia J.S., Freeman W.T., Weiss Y., *Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms*, Mitsubishi Electric Research Laboratories, 2004.

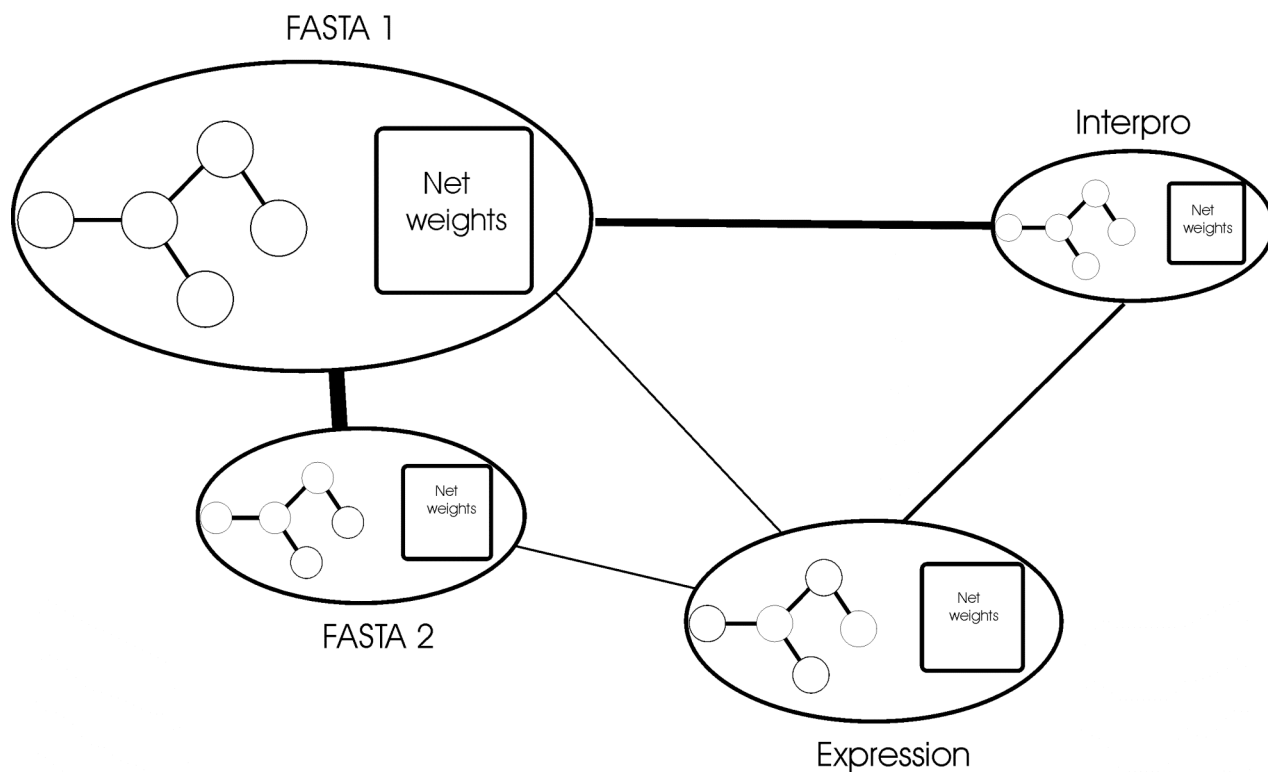


Figure 1. The full hRMN composed of several source objects, each containing a graph from the pairwise scores and attributes determining the source class weights.

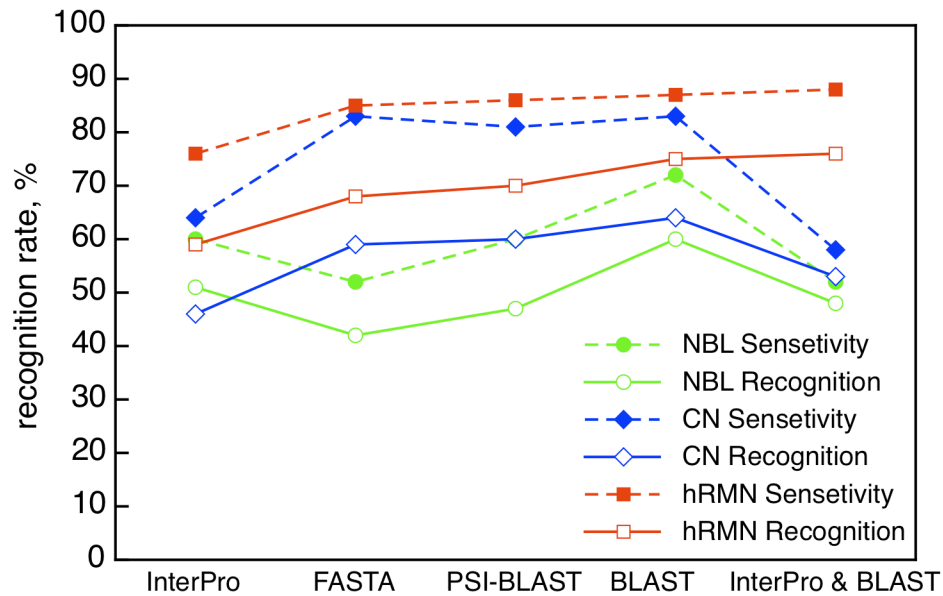


Figure 2. Results for cross-validation at the fifth FunCat level across all genomes, for all features as well as the integration of BLAST and InterPro. Dashed lines represent sensitivity, solid lines recognition rates; coded in red are results for hRMN, blue-Neighbor's Best List, green-Closest Neighbor.