Note to Aalt-Jan: If you read the .pdf is fine, but I suggest you open later the .odt to see a few comments in the margin that are not printed in the .pdf. This is for all 3 documents (not for the appendix)

Computational methods for PFP and network-based methods

Protein function prediction (PFP) is one of the most important aims of modern biology. In crop and livestock species, PFP is conventionally based on annotation transfer from the few well-studied species, such as Arabidopisis and humans. While successful, these methods rely on the assumption that homologs proteins share function, which has been proved wrong in many cases [1]. It is thus desirable to complement the orthology-based methods with other approaches.

Network-based methods, for instance, infer the function of proteins exploiting the principle of guilt-by-association. Based on this, proteins that interact are likely to have similar function [2]. These methods have a lot of potential because they can utilize the information generated by high-throughput biological experiments, such as co-expression [3] or protein-protein-interactions [4] to construct networks from which to infer function. Furthermore, network-based may allow to identify regulatory elements that are relevant for a set of functions or functional cascades. Other applications of networks data include the discovery of translational modifications or aligned modules (signs of conservation) via network-comparison. Explanations about these and other network analyses are given somewhere else [19].

Networks in poorly annotated species

Because a wide range of data can be combined in these networks, the network approaches seem particularly relevant for poorly annotated species, such as agricultural species, where the validated data of a particular kind (i.e. coexpression, protein-protein-interactions…) is scarce [12]. The identification of putative transcription factors or hub genes via networks, for instance, may be more relevant for poorly annotated species in which knowledge about key regulatory elements is more limited. Network approaches, however, are also more challenging for these species because the data may be insufficient to carry a network analysis. A previous study, has shown that it is possible to develop network-based methods that can utilize the limited network resources of some crop species like rice, poplar, soybean and tomato, and achieve accurate PFP [5] by combining different data sources. In their approach, they combined data co-expression and protein-protein-interaction data, as well as information from other well annotated species, such as *Arabidopsis Thaliana*.

In livestock species, network data is even more limited than for the aforementioned crop species. Nevertheless, it is expected that the data available will increase in the coming years. Efforts such as The Functional Annotation of Animal Genomes consortium (FAANG) [6], are currently generating functional annotations for some relevant species such as pig and chicken. This information could be utilized by the network-based methods. A previous study has used coexpression networks in chickens to infer function via defining GO-enrichment-modules [12]. For this, they defined a Conditional dependent network, in which a large set of unrelated experiments were combined. Thus, the network became robust and reproducibility of the analysis greatly improved. This approach based on GO-enrichment, however, is indirect in that it first identify identify functional modules in the network. Sharan et. al. [16] judged the direct methods (directly predict the function of a protein) as slightly superior to the indirect ones. Furthermore, the method does not enjoy the advantage of the direct statistical-learning-based methods. Methods based on statistical learnig have more potential than other methods because they can identify combinations of features that correlate with certain functions [20]. An interesting question therefore is whether it is possible to use some of the direct statistical-learning-based methods that are efficient for PFP, in livestock species.

BMRF

In order to develop a statistical-learning-based network method that is efficient for livestock species, a logical approach is to utilize one of the state-of-art methods used for crop species. To our knowledge, the Bayesian Markov Random Field (BMRF) [1] is the highest-performance method of this kind. BMRF was one of the highest performabnce methods in the Critical Assessment of protein Function Annotation (CAFA) experiment (Illustration 1), but it is particularly efficient for poorly annotated species for two reasons mainly: First, it can synthesizes heterogeneous data into one network. For instance, it can integrate co-expression from the same species, as well as from related species; and second, though Gibbs sampling, BMRF can take into account unlabeled proteins to estimate the parameters of the model. Because BMRF directly exploits the information from the neighbors, we would expect that PFP via BMRF will be more accurate for the genes with a large number of neighbors (co-expressed genes) and whose neighbors information is accurate. Thus, we would expect better prediction performance for the most general GO-terms and, in particular, for the hub genes that are located in high-density regions in the network. However, on the other hand, from a biological perspective we expect that the genes are more co-expressed in the most specific GO-terms and therefore the principle o guilt-by-association holds better in this cases. BMRF, therefore, may not only be a good starting point for inferring PFP in livestock, but it also may be useful to identify regulatory elements.

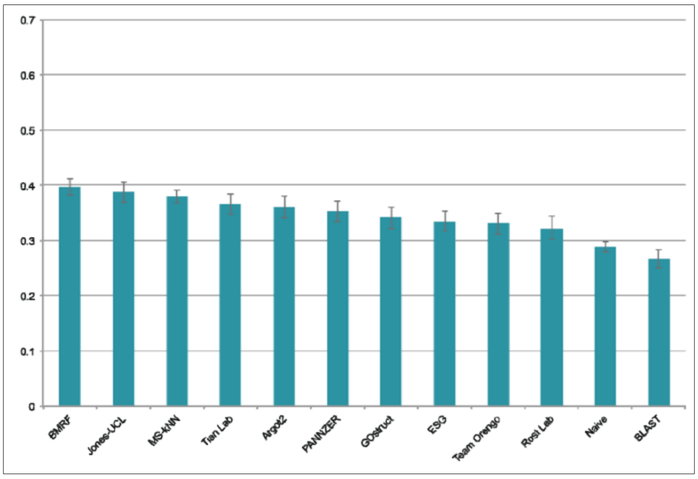

Illustration 1: Comparison of BMRF with other PFP methods. Evaluation for the Biological process category in H. Sapiens. Source:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3584181/f

Positive-unlabbeled learning

We discovered an important potential problem with BMRF. The learning process may be biased because BMRF attempts to solve a one class classification problem where the annotated data consists solely on positive cases. This is because it is very difficult to be certain that a protein does not perform a function. In fact, this problem applies to many situations in bioinformatics because from a biological perspective, the lack of evidence for a connection does not imply that such a connection does not exist. As an example, for rice 415 proteins have experimental evidence for a biological process, but not a single protein has a validated proof of no-connection with a function. Since only positive associations are reported, the negative set is composed of all unlabeled data. This leads to some bias in the prediction because the unlabeled data may contain some positive cases. This problem increases with increasing numbers of unannotated proteins, such as in the case of network data from livestock species. To overcome this problem, a new type of machine learning has emerged called Positive Unlabeled learning (PU). With PU it is possible to identify the proteins that are more unlikely to have a given function. Hence, the number of unlabeled cases can be minimized by extracting some “reliable negatives” proteins from the set of unlabeled. It has been proved theoretically that, by identifying sets of reliable negatives, PU improves the performance of machine learning algorithms in situations where only positive labels are known [7]. PU has been successfully applied to a variety of problems related to PFP [7-10,13,14]. In [7], for instance, the authors extracted a set of reliable-negatives proteins from the unlabeled dataset by computing the euclidean distance between a set of positive proteins and the set of unlabeled, and defining a threshold of similarity. Yang et al. [8] developed a multi-label version of PU learning to identify genes associated with diseases; [9] developed two novel approaches to identify reliable negatives that can be applied in different algorithms. Jiang et al. [10] applied PU on a support vector machine and outperformed all pre-existing methods for pupylation sites prediction, and then Nan et al. [13] improved their method by adding as a first step to the algorithmn in [10], the method described in [7]. Lastly, in another recent study, Nusrath et al. [14] used a Self organizing map to extract reliable negatives from unlabeled data-set of drug-drug interactions. None of these studies, however, have applied PU learning on a BMRF. BMRF is one of the few methods that can handle networks with a large portion of unknown cases, and PU can extract reliable information from these unknowns. Our hypothesis therefore is that a PU implementation of BMRF will be particularly effective for PFP in species with limited network resources, such as pig and chicken. Furthermore, since with PU it is possible to create an alternative database with negative examples, the task of identification of regulatory elements is, to some extend, simplified.

The aim of this study is to develop a PU implementation of an existing Bayesian Markov Random Field algorithm that can efficiently assign proteins to GO term categories and identify putative regulatory elements using network data from chickens.