MicroRNA Prioritization based on Target Profile Similarities

Péter Marx[1,3], Bence Bolgár[1], András Gézsi[2], Attila Gulyás-Kovács[1] and Péter Antal[1]
[1]*Department of Measurement and Information Systems,*
Budapest University of Technology and Economics, Budapest, Hungary
[2]*Department of Genetics, Cell- and Immunobiology, Semmelweis University, Budapest, Hungary*
[3]*MTA-SE Neuropsychopharmacology and Neurochemistry Research Group, Budapest, Hungary*

Keywords: microRNA, microRNA Target, Kernel Methods, Multiple Kernel Learning, Gene Prioritization.

Abstract: microRNAs form a complex regulatory network with thousands of target genes. This network is known to suffer specific, but largely elusive, genetic perturbations in various types of disease. Accurate prioritization of microRNAs for each disease type would elucidate those perturbations and so facilitate therapeutic and diagnostic design. The multiple target profiles of microRNAs stemming from various experimental and *in silico* methods allow the definition of wide range of similarities over microRNAs, but the combined use of these of heterogeneous similarities was not utilized in the gene prioritization approach. Using microRNAs as bases, prioritization with a disease-specific query set of microRNAs is straightforward once a microRNA-microRNA similarity matrices have been derived. Here we demonstrate the application of a one-class version of the multiple kernel learning framework in order to fuse heterogeneous characteristics of microRNAs. We evaluate the method with breast cancer-specific queries, illustrate its technological aspects, and validate our results not only by standard leave-one-out cross validation, but also with a prospective evaluation.

1 INTRODUCTION

The growing availability of omic measurements and multiple characterizations of entities like genes or proteins led to the emergence of data and knowledge fusion as a central challenge in various fields. From the 90's, the similarity based virtual screening became a popular method in chemoinformatics, relying mostly on rank fusion and heuristic fusion of data sources (Johnson and Maggiori, 1990; Ginn et al., 1997; Ginn et al., 2000; Eckert and Bajorath, 2007). The aim of these methods were to make predictions based on different heterogeneous information sources in various research fields. From the 2000's, gene prioritization emerged as a separate task with the fusion of wide range of information about genes and gene products, such as sequence similarities, transcriptional regulation similarities or expression level similarities (Freudenberg and Propping, 2002; Aerts et al., 2006; Kohler et al., 2008; Moreau and Tranchevent, 2012). A theoretically sound foundation for omic data fusion using similarities was proposed in 2004 by Lanckriet et al., which utilized the kernel methods for large-scale similarity integration (Lanckriet et al., 2004). Currently, wide range of methods were reported for data and knowledge fusion, such as the aforementioned kernel-based methods and the more traditional text mining and network based fusion methods (De Bie et al., 2007; Liekens et al., 2011; Lee et al., 2011). Although there are many open questions, such as the management of systematically incomplete similarities, uncertainty in similarities, biological relevance of ranks and proper evaluation, the kernel-based fusion methods allows a universal framework for any set of entities and it achieved an impressive performance in multiple settings and domains, compared to for example rank fusion (Bornigen et al., 2012; Arany et al., 2013). Despite the universality of the kernel-based fusion and prioritization framework, the basis for the applications so far were nearly exclusively gene centric, with some exceptions such as its application for protein-ligand interaction prediction (Iacucci et al., 2012) and for drug repositioning (Arany et al., 2013; Bolgár et al., 2013). In this paper we report the application of the kernel-based fusion and prioritization framework over miRNAs, as a step towards the extension of this methodology to integrate information over heterogeneous sets of entities.

MiRNAs are short approximately 22 base pairs

Marx P., Bolgár B., Gézsi A., Gulyás-Kovács A. and Antal P..
MicroRNA Prioritization based on Target Profile Similarities.
DOI: 10.5220/0004925502780285
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2014), pages 278-285
ISBN: 978-989-758-012-3

long non-coding RNAs and have an important role in *cis*-regulation. Generally miRNAs recognize their target by the 2-8 bps long seed region at the 5' end of the miRNA. They bind to the 3' end of the target mRNA by Watson-Crick base pairing and can repress the mRNA translation in multiple ways such as mRNA degradation, mRNA de-deadenylation, inhibition of initiation and premature ribosome drop off. MiRNAs can also have binding sites in the coding region or at the 5' end of the target moreover one miRNA can have multiple targets and one target can bind to multiple miRNAs. Although there is an exponentially increasing interest towards miRNAs (more than 2000 miRNAs are known in human cells), plenty of their target genes are only connected to them based on *in silico* prediction. There is a wide variety of *in silico* methods for target exploration, prediction and validation, but these methods are somewhat complementary and suffer from different biases. To avoid these errors we used only experimentally supported data in this paper. For additional information on miRNAs and gene regulation see (Nilsen, 2007) and (Chen and Rajewsky, 2007).

Many studies suggested an important role for miRNAs in the development of cancer based on the observation that their expression levels changed in tumor tissues which led to altered mRNA quantity inside the cell, and similarly both inherited and somatic genetic variations related to miRNAs were reported as associated with diseases. In the paper we illustrate the new miRNA-based prioritization method with general, disease aspecific kernels and with breast cancer queries. Breast cancer is one of the leading cancer types among the female population, and both miRNA related genetic variations and expression levels were reported to be associated with various phenotypic features of breast cancer classes.

Our goal is to examine the efficiency of miRNA prioritization based on the already available data which could extend the current gene centric prioritization. Predicting new miRNA-disease connections can lead to a more detailed description of the regulation of the disease in focus. Since the miRNA data is coming from different experiments and including *in silico* target predictions would increase heterogeneity we use a well known information fusion method as well and compare it to the other one kernel prioritization algorithm. For evaluation of the performance of the prioritization we apply the leave-one-out cross-validation and because of its potential positive bias through knowledge contamination, see e.g. (Kohler et al., 2008; Bornigen et al., 2012), we performed a prospective evaluation using a recent summary from Jacobsen et al. (Jacobsen et al., 2013).

2 EARLIER WORKS

The data fusion based prioritization methods through integration of similarities and rankings from multiple information source has a long and somewhat redundant history, with its roots in virtual screening in chemoinformatics (Johnson and Maggiori, 1990; Ginn et al., 1997; Ginn et al., 2000; Eckert and Bajorath, 2007), kernel methods in chemoinformatics (Burbidge et al., 2001; Warmuth et al., 2003), in the one-class classification and prioritization (Moya and Hush, 1996; Schlkopf et al., 2001), in multiple kernel learning (Lanckriet et al., 2004; Rakotomamonjy et al., 2008) and in gene prioritization methods (Freudenberg and Propping, 2002; Aerts et al., 2006; Kohler et al., 2008; De Bie et al., 2007; Liekens et al., 2011; Lee et al., 2011; Moreau and Tranchevent, 2012).

Compared to the growing admittance of the central role of miRNAs, the use of this information source is surprisingly rare. For example, transcription regulation based gene-gene similarities (kernels) are dominantly constructed base on the commonality of the transcription factor binding sites of the gene pairs, and not by the commonality of miRNAs (see e.g. (Aerts et al., 2006)). Two notable exceptions reported the prioritization of miRNAs using a network approach (Xu et al., 2011) and using a support vector machine classifier (Li et al., 2012). The work reported in this paper continues this line as it uses miRNAs for the bases of prioritization, instead of the prevailing gene centric approach. However, this work adopts a more principled approach for fusion by adopting the multiple kernel learning framework in the one-class classification (/prioritization) settings.

3 METHODS

The number of databases containing miRNA-target pairs is growing just like the methods which utilize machine learning algorithms to find further potential targets for a miRNA (Ficarra et al., 2012). However applying such methods largely increase the number of miRNA-target pairs and extend the miRNA set using predicted connections, it leads to higher degree of uncertainty at the results. Therefore we used only experimentally supported targets for every miRNA. Using the predictions based results requires an extensive evaluation of the methods which will be aimed in a future study.

3.1 Computing miRNA Kernel

We used miRTarBase (Hsu et al., 2011) (downloaded 11/14/2013 Release: 4.5) to collect experimentally validated miRNA-target pairs. We filtered out the non-human target genes from the data resulting in a dataset of 596 unique miRNAs and their experimentally supported targets. Similarity between two miRNAs is defined by the number of common target genes in a given context, e.g. using a given miRNA target validation method. The primary diagonal of the matrix X contains the degree of a miRNA which equals with the number of targets in the present study. To extend our definition of miRNA similarity and to ensure positive definiteness we squared the matrix and normalized it with a diagonal matrix D containing the reciprocal of the square root of the diagonal elements of X.

$$K = DX^2D$$

This way two miRNAs i, j will be more similar (x_{ij} will be higher) if they share more targets or they have common similar miRNAs. miRTarBase contains information on the experimental methods used to validate the miRNA-target pairs. These methods can be classified into two groups based on the strength of the evidence they provide. E.g. microarrays give indirect proof as the prediction is based on the differential expression miRNA-gene pairs therefore it is a 'weak' validation. On the contrary attaching reporter genes to the target gene proves directly the miRNA effects on the gene (Ficarra et al., 2012). We used the classification of miRTarBase (functional, functional-weak, nonfunctional, nonfunctional-weak) and built different kernels for each of them with the above described method.

3.2 Multiple Kernel Learning

The most important benefit of Multiple Kernel Learning is that it provides a way to perform both data fusion and prioritization in a joint manner, by solving one single optimization problem. This is achieved by incorporating kernel "weights" into the objective function, which essentially amounts to determining an optimal weighting of (possibly very heterogeneous) information sources by exploiting the information content of the query itself. Further favorable properties include computational efficiency, flexibility (depending solely on pairwise similarities), and often very good empirical performance. Multiple kernel methods have already been successfully applied in the context of genomic data fusion (Lanckriet et al., 2004), however, first formulations did not support gene prioritization. The first tool which performed fusion and prioritization in a coupled way involved a modification of the multiple kernel one-class SVM algorithm (Schölkopf et al., 2001) to perform the ranking of the entities based on their projections to the normal of the hyperplane (De Bie et al., 2007). In order to attain better computational performance, we modified the formulation of Vishwanathan et al. (Sun et al., 2010):

$$\begin{aligned} \min_{\boldsymbol{f},\rho,\boldsymbol{\xi},\boldsymbol{d}} \quad & \frac{1}{2}\sum_k \frac{\|f_k\|^2_{\mathcal{H}_k}}{d_k} - \rho + \frac{1}{\nu l}\sum_i \xi_i + \frac{\lambda}{2}\left(\sum_k d_k^p\right)^{\frac{2}{p}} \\ s.t. \quad & \sum_k f_k(\boldsymbol{x_i}) \geq \rho - \xi_i, \\ & \boldsymbol{\xi} \geq 0, \quad \boldsymbol{d} \geq 0, \quad i = 1,2,\ldots,l, \end{aligned}$$

where $\boldsymbol{f}$ parameterizes the hyperplane, d_k is the weight of the kth kernel, ρ denotes the margin, ν controls the model complexity, $\boldsymbol{\xi}$ are the slack variables and the last term stands for the Lp-norm regularization of the kernel weights. Note that this is just the primal of the one-class SVM augmented by the kernel weights.

In view of earlier findings (Yu et al., 2010), we utilized the squared L2-norm as weight regularizer. In this setting, the dual reads

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{8\lambda}\sum_k \left(\boldsymbol{\alpha}^T K_k \boldsymbol{\alpha}\right)^2 \\ s.t. \quad & 0 \leq \alpha_i \leq 1, \quad \mathbf{1}^T\boldsymbol{\alpha} = \nu l, \end{aligned}$$

where $\boldsymbol{\alpha}$ are the dual variables and the optimal kernel weights can be recovered from the solution. Note that this dual is differentiable and convex, which is very easy to optimize. The score of a sample can be computed as

$$score(\boldsymbol{x}) = \frac{\sum_i \alpha_i \sum_k d_k K_k(\boldsymbol{x_i}, \boldsymbol{x})}{\sqrt{\sum_k d_k \boldsymbol{\alpha}^T K_k \boldsymbol{\alpha}}},$$

i.e. here we determine the distance to the hyperplane instead of the side on which the sample $\boldsymbol{x}$ lies; samples are then ranked on the basis of their respective scores.

4 RESULTS

We used multiple kernels and training sets (queries) to cover the different subtypes of breast cancer classified by Samantarrai et al. (Samantarrai et al., 2013) as follows: Ductal carcinoma in situ, Lobular carcinoma in situ, Invasive ductal- or invasive lobular carcinoma. Our first approach was to collect a subset of miRNAs from one subtype of breast cancer and

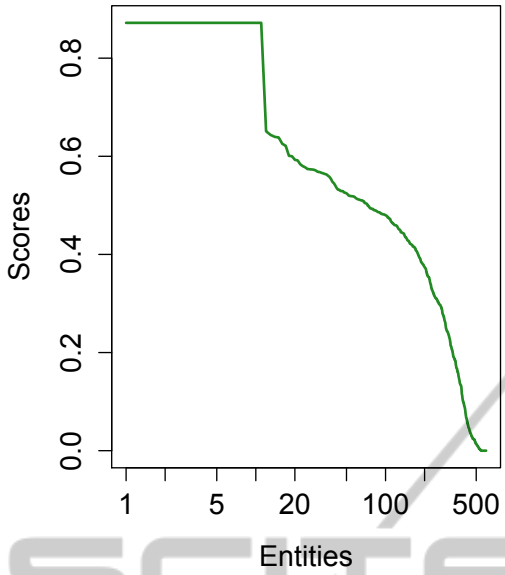

Figure 1: The scores of the miRNAs for the MKL kernel and the third query (axis X is log scaled).

one type of differential expression (up- or downregulated). We will refer to these subset of miRNAs as a class. This way the prioritization should recover the remaining set of miRNAs of the same class, and it should give a higher score to other miRNAs connected with breast cancer. The other approach was to choose miRNAs which are present in more subclasses to catch the different properties of each subtype and prioritize with this breast cancer "hyperclass". In both cases we prioritize with the MKL algorithm and also the kernel built from the full human miRTarBase. The first query (Query1) contained the hsa-miR-182, hsa-miR-183, hsa-miR-200c, hsa-miR-21. The second query (Query2) contained miRNAs of the subtype ductal carcinoma in situ downregulated miRNAs (hsa-miR-125b, hsa-miR-127, hsa-miR-210, hsa-miR-7). We included those miRNAs in the last query (Query3) which were present in more classes to span the set of miRNAs which are connected to the disease. These are hsa-let-7d, hsa-miR-182, hsa-miR-183, hsa-miR-21, hsa-miR-210, hsa-miR-221. When it was not specified and for one pre-miRNA more mature miRNA were available we used both the 3p and 5p miRNA. We analyzed other queries also which gave similar results and for this reason those are not included in the manuscript.

Query1 with the full kernel ranked the other miRNAs in the same class hsa-miR-361-5p, hsa-miR-374a, hsa-miR-93 113th, 52th, 12th respectively and resulted an average 126.46 rank for the all miRNAs whereas the MKL kernel ranked them 29th, 84th, 13th with an average of 121.96. Both methods ranked miRNAs from other classes in the first 10% of the data. The rank of the miRNA is in brackets. Full: hsa-miR-221 (8), hsa-miR-96 (44); MKL: hsa-miR-221 (15), hsa-miR-96 (42), hsa-miR-18 (44). Query2 ranked the miRNAs from the same class (first number belongs to the full kernel and the second to the MKL kernel set) hsa-let-7d (222,123), hsa-miR-221 (88,60), hsa-miR-320 (23,66) and others full: hsa-miR-182 (51), hsa-miR-361-5p (43), hsa-miR-9 (50), hsa-miR-93 (20) and MKL: hsa-miR-182 (25), hsa-miR-361-5p (33), hsa-miR-9 (41), hsa-miR-93 (38), hsa-miR-10b (40) and hsa-miR-21 (57). The average ranks of miRNAs from all subtypes are 120.83 for the full kernel and 112.92 for the MKL set. The results for the last query is similar to the first two, generally using MKL kernel set gives a lower average rank. Figure 1 shows the scores of the last query for the MKL kernel set on a log scale. The cut-off point shows a break between the query elements and the other miRNAs. The MKL method computes a weight for every kernel. In our case these weights were almost uniform and not varied between the different queries. Generally, the weak functional set had 0.4 weight and the other kernels 0.2.

Table 1: Leave-one-out cross validation results for the ductal carcinoma in situ breast cancer type upregulated miRNAs. The numbers indicate the rank of the left out miRNA.

microRNA	Full	MKL
hsa-miR-21	169	113
hsa-miR-200c	257	57
hsa-miR-182	41	31
hsa-miR-183	60	54
hsa-miR-361-5p	101	28
hsa-miR-374a	63	119
hsa-miR-93	16	14
Average	101	59.43

For validation we used not only the same set of miRNA, but also the results from Jacobsen et al. (Jacobsen et al., 2013) for a prospective evaluation. They computed the correlation R^2 between miRNAs expression levels and DNA copy number variation. Based on their results the variation of the expression levels in case of cancer related miRNAs can be explained by DNA copy number variation. We selected the miRNAs which has a higher than 0.1 correlation and which are known to play a role in breast cancer. This routine makes it possible to validate our results against a newer breast cancer dataset (see Table 2). Furthermore, leave-one-out cross validation (LOOCV) was performed on the Ductal carcinoma in situ upregulated miRNA set. The average rank of the left out miRNA was 101 with the full kernel and 59.43 with the MKL algorithm (see Table 1).

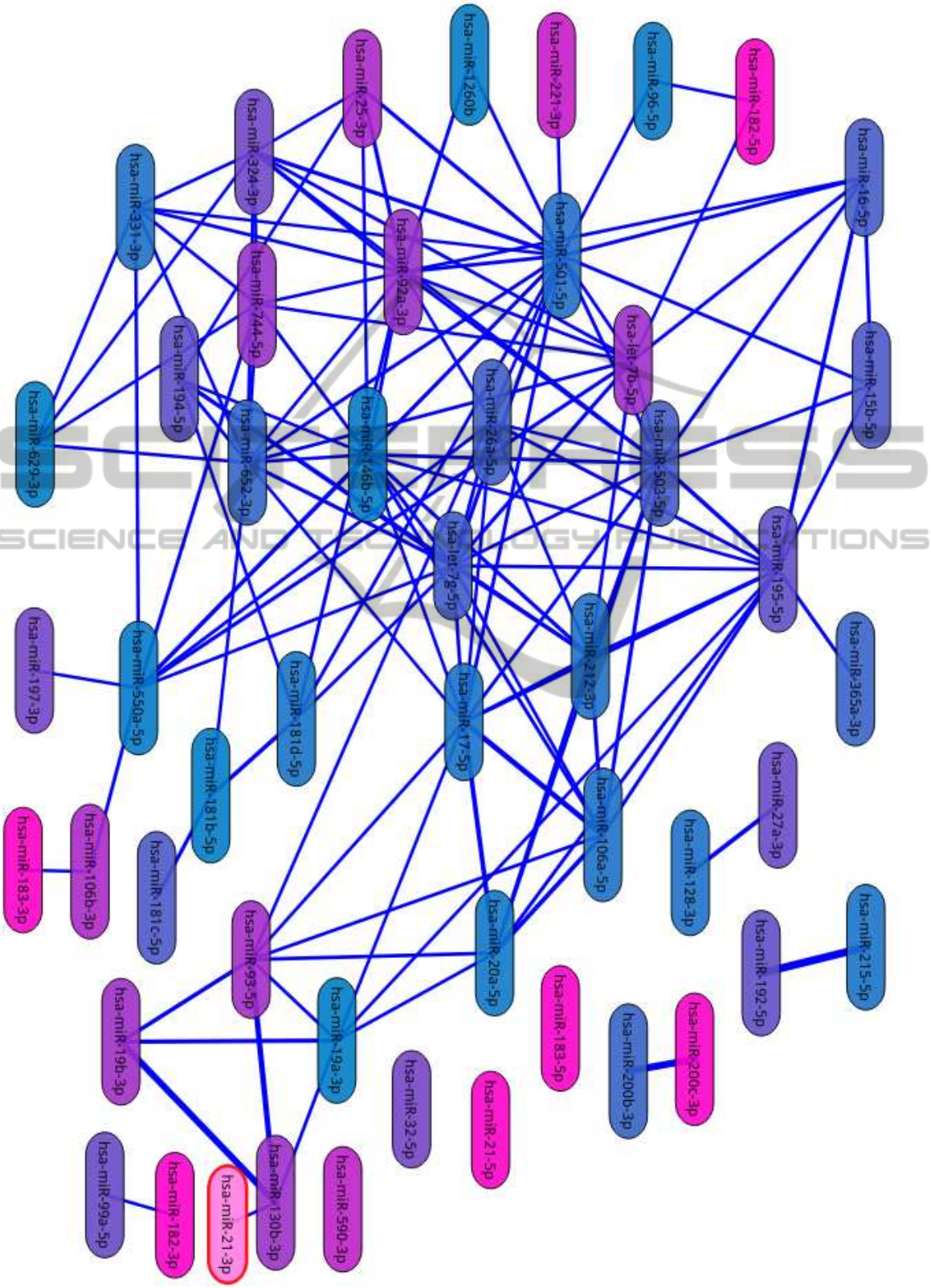

Figure 2: The graph of the first 50 ranked miRNAs for the full kernel and the first query. Only edges with higher than 0.4 similarity scores are shown. The color of a node is defined by its score, where blue indicates lower score.

Table 2: The miRNAs ranked below 50 from the Jacobsen dataset. *The miRNA was part of the query.

microRNA	Query1		Query3	
	Full	MKL	Full	MKL
hsa-miR-106b	11	51	12	79
hsa-miR-141	51	27	375	386
hsa-miR-15b	26	11	47	30
hsa-miR-17	41	20	85	41
hsa-miR-19b	13	24	34	36
hsa-miR-200c	6*	5*	280	95
hsa-miR-20a	43	28	92	88
hsa-miR-21	2*	2*	14*	3*
hsa-miR-92a	16	31	3	18
hsa-miR-140	61	22	121	80
hsa-miR-182	3*	6*	16*	2*
hsa-miR-183	1*	1*	15*	1*
hsa-miR-186	76	18	68	34
hsa-miR-25	10	25	9	37
hsa-miR-320a	107	87	32	43

5 DISCUSSION

Despite the popularity of gene prioritization, there are only a few studies in which researchers used miRNA prioritization for finding novel disease related miRNAs. To our knowledge, those are only validated by LOOCV. The quantitative evaluation of the performance of our method for the queries (Query 1,2,3) shows that it ranks most of the miRNAs related to breast cancer according to the review from Samantarrai et al. (Samantarrai et al., 2013) in the first 20% of the miRNA set used in our study. Bornigen et al. (Bornigen et al., 2012) used the top 5 %, 10 %, 30 % to validate the performance of the prioritization methods for real examples. Taking the number of miRNAs into account this rate (top 20% includes almost all miRNAs related to breast cancer) can be enough for experimental validation with a real situation with limited budget also.

On Figure 2 miRNAs are connected based on the similarity of miRNA-miRNA pairs. Nodes even with lower rank has plenty of neighbors, on the other hand nodes with higher scores, typically query members, share less properties with the other miRNAs in the graph (i.e. connected with edge above the threshold 0.4). The manual evaluation of the graph also suggested that the prioritization gives higher score not only for those miRNAs which are in the same miRNA cluster as the query miRNAs. Generally the nodes with more connections belongs to a miRNA family represented by more members in the first 50 ranked entities, which verifies our similarity definition.

The prospective validation on the results of Jacobsen et al. ranked the miRNAs in Table 2 on the first 10% of the whole set. This means that the applied method was able to identify new candidates which were experimentally validated later without retrospective bias. The role of these miRNAs in cancer development is supported by experimental evidence. For example miR-17 and miR-200c are oncosuppressors and miR-92 is an oncomiR (Samantarrai et al., 2013). Furthermore (Kim et al., 2011) found inverse correlation between hsa-miR-17, hsa-miR-92, hsa-miR-106b, hsa-miR-20a, hsa-miR-19 and ZBTB4 gene which has a significant correlation to relapse-free survival. The almost uniform kernel weights most likely caused by the similar input data as for plenty of miRNA-target pairs are supported with more than one experimental evidence. The higher weight of weak functional kernel is possibly the consequence of the higher number of connections available. Our results show that the MKL method which includes both prioritization and fusion outperformed the single kernel prioritization in all of the above cases. With even more heterogeneous information source the application of the MKL method can be even more advantageous.

6 CONCLUSIONS

The high heterogeneity of experimental and *in silico* methods for miRNA target exploration, prediction and validation result in highly heterogeneous miRNA profile and respectively, miRNA similarities. We applied the multiple kernel learning framework for the fusion of these similarities in breast cancer. Results indicate that this methodology can be a promising approach to find new miRNAs related to cancer.

Furthermore, this approach could be used to refine existing miRNA families and to define more detailed gene-gene similarities . Our results also indicate that miRNA prioritization can support study design and interpretation of miRNA discovery, and it also indicates the universal applicability of kernel-based fusion methods over different sets of entities, such as miRNAs, beside the current gene centric approach. With the growing number of discovered miRNAs and the deeper understanding of their role in the post-transcriptional regulation miRNA prioritization can lead to a more detailed description of miRNA-gene regulatory networks.

However, the approach described in this paper

seems to be successfully identify even novel miRNAs for breast cancer still there are many question to be answered. MiRNA networks or miRNA similarity can be defined in other ways such as comparing the sequence or based on single nucleotide polymorphisms (SNP) or by using conservation scores. After a thorough investigation of the target prediction algorithms we can include predicted miRNA-target pairs in the analysis to increase the number of miRNAs in the study. Last with the growing number of available data it will be possible soon to build more tissue specific miRNA networks. This way we will have a heterogeneous data source and we can make a good use of the MKL method as it can be used to compare the importance of the different similarity scores or data sources based on the learned weights of the kernels. Future work will concentrate on answering these questions.

REFERENCES

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544.

Arany, A., Bolgár, B., Balogh, B., Antal, P., and Mátyus, P. (2013). Multi-aspect candidates for repositioning: Data fusion methods using heterogeneous information sources. *Current Medicinal Chemistry*, 20(1):95–107.

Bolgár, B., Arany, A., Temesi, G., Balogh, B., Antal, P., and Mátyus, P. (2013). Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies. *CURRENT TOPICS IN MEDICINAL CHEMISTRY*, 13(18):2337–63.

Bornigen, D., Tranchevent, L., Bonachela-Capdevila, F., Devriendt, K., De Moor, B., De Causmaecker, P., and Moreau, Y. (2012). An unbiased evaluation of gene prioritization tools. *Bioinformatics*, 28(23):3081–3088.

Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem. (Oxford, U. K.)*, 26(1):5–14.

Chen, K. and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and micrornas. *Nat Rev Genet*, 8(2):93–103.

De Bie, T., Tranchevent, L.-C., Van Oeffelen, L. M., and Moreau, Y. (2007). Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–i132.

Eckert, H. and Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today*, 12(5-6):225–233.

Ficarra, E. et al. (2012). One decade of development and evolution of microrna target prediction algorithms. *Genomics, proteomics & bioinformatics*.

Freudenberg, J. and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18:S110–S115.

Ginn, C., Turner, D., Willett, P., Ferguson, A., and Heritage, T. (1997). Similarity searching in files of three-dimensional chemical structures: Evaluation of the eva descriptor and combination of rankings using data fusion. *J. Chem. Inf. Model.*, 37(1):23–37.

Ginn, C., Willett, P., and Bradshaw, J. (2000). Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.*, 20(1):1–16.

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., et al. (2011). mirtarbase: a database curates experimentally validated microrna–target interactions. *Nucleic acids research*, 39(suppl 1):D163–D169.

Iacucci, E., Tranchevent, L., Popovic, D., Pavlopoulos, G., De Moor, B., Schneider, R., and Moreau, Y. (2012). Reliance: a machine learning and literature-based prioritization of receptor-ligand pairings. *Bioinformatics*, 28(18):I569–I574.

Jacobsen, A., Silber, J., Harinath, G., Huse, J. T., Schultz, N., and Sander, C. (2013). Analysis of microrna-target interactions across diverse cancer types. *Nature structural & molecular biology*.

Johnson, M. and Maggiori, G. (1990). *Concepts and applications of molecular similarity*. Wiley.

Kim, K., Chadalapaka, G., Lee, S., Yamada, D., Sastre-Garau, X., Defossez, P.-A., Park, Y.-Y., Lee, J.-S., and Safe, S. (2011). Identification of oncogenic microrna-17-92/zbtb4/specificity protein axis in breast cancer. *Oncogene*, 31(8):1034–1044.

Kohler, S., Bauer, S., Horn, D., and Robinson, P. (2008). Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4):949–958.

Lanckriet, G., De Bie, T., Cristianini, N., Jordan, M., and Noble, W. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

Lee, I., Blom, U., Wang, P., Shim, J., and Marcotte, E. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121.

Li, X., Wang, Q., Zheng, Y., Lv, S., Ning, S., Sun, J., Huang, T., Zheng, Q., Ren, H., Xu, J., Wang, X., and Li, Y. (2012). Prioritizing human cancer microrna based on genes? functional consistency between microrna and cancer. *Nucleic Acids Research*, 40(16):7653–7665.

Liekens, A., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., and Del-Favero, J. (2011). Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology*, 12(6).

Moreau, Y. and Tranchevent, L. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536.

Moya, M. and Hush, D. (1996). Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474.

Nilsen, T. W. (2007). Mechanisms of microrna-mediated gene regulation in animal cells. *TRENDS in Genetics*, 23(5):243–249.

Rakotomamonjy, A., Bach, F., Canu, S., and Grandvalet, Y. (2008). Simplemkl. *J. Mach. Learn. Res.*, 9:2491–2521.

Samantarrai, D., Dash, S., Chhetri, B., and Mallick, B. (2013). Genomic and epigenomic cross-talks in the regulatory landscape of mirnas in breast cancer. *Molecular Cancer Research*, 11(4):315–328.

Schlkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Sun, Z., Ampornpunt, N., Varma, M., and Viswanathan, S. (2010). Multiple kernel learning and the smo algorithm. In *Advances in neural information processing systems*, pages 2361–2369.

Warmuth, M., Liao, J., Ratsch, G., Mathieson, M., Putta, S., and Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Model.*, 43(2):667–673.

Xu, J., Li, C.-X., Lv, J.-Y., Li, Y.-S., Xiao, Y., Shao, T.-T., Huo, X., Li, X., Zou, Y., Han, Q.-L., Li, X., Wang, L.-H., and Ren, H. (2011). Prioritizing candidate disease mirnas by topological features in the mirna target?dysregulated network: Case study of prostate cancer. *Molecular Cancer Therapeutics*, 10:1857–1866.

Yu, S., Falck, T., Daemen, A., Tranchevent, L.-C., Suykens, J. A., De Moor, B., and Moreau, Y. (2010). L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC bioinformatics*, 11(1):309.