Gene Prioritization Through Geometric-Inspired Kernel Data Fusion

Pooya Zakeri, Sarah Elshal, Yves Moreau

Department of Electrical Engineering - STADIUS
KU Leuven
3001 Leuven, Belgium

Medical IT
iMinds
Leuven, Belgium

Pooya.zakeri@esat.kuleuven.be

***Abstract*— In biology there is often the need to discover the most promising genes, among a large list of candidate genes, to further investigate. While a single data source might not be effective enough, integrating several complementary genomic data sources leads to more accurate prediction.**

We propose a kernel-based gene prioritization framework using geometric kernel fusion which we have recently developed as a powerful tool for protein fold classification [1]. It has been shown that taking more involved geometry means of their corresponding kernel matrices is less sensitive in dealing with complementary and noisy kernel matrices compared to standard multiple kernel learning methods. Since genomic kernels often encodes the complementary characteristics of biological data, this leads us to research the application of geometric kernel fusion in the gene prioritization task.

We utilize an unbiased and prospective benchmark based on the OMIM [2] associations. Experimental results on our prospective benchmark show that our model can improve the accuracy of the state-of-the-art gene prioritization model.

Keywords—gene prioritization; data fusion; geometric kernel data fusion; Log-Euclidean mean; OMIM;

I. INTRODUCTION

Gene prioritization addresses the need of discovering the most promising genes, among a large list of candidate genes, to further investigate them (in the wet lab). In particular, it takes care of genes-phenotype identification which is one of the most essential, prevalent, challenging and yet critical objective in modern biology and biomedical research. For example, hunting disease-associated genes is a demanding process and plays a crucial role in understanding the relationship between disease and genes. It has various applications ranges from functional genomics and drug design studies in both pharmacogenomics and personalized medicine. For instance, using the reliable automatic gene prioritization tool, the design of appropriate drugs could be sped up for many kinds of complex diseases linked to the unknown top-ranked genes.

The notion of gene prioritization was coined by Perez-Iratxeta et al. [3] in 2002. Since then, various approaches based on different genomic data sources and often machine learning strategies have been proposed for gene prioritization. Most of these strategies utilized the 'guilt-by-association' principle. They assumed that the most promising candidate genes for a disease are indeed the genes similar to the ones known to be linked to that disease. As a result, all of the strategies based on 'guilt-by-association' need a seed genes for training a model. Then, they rank a set of candidate disease genes for the biological process, phenotype or disease under study using the learned models.

Many successful gene prioritization methods use multiple genomic data sources to deliver more accurate ranking [4-7]. Finding an efficient technique for integrating heterogeneous biological data sources has received growing attention. Indeed, while a single data source might not be effective enough for these problems, fusing several complementary genomic data sources results in more accurate prediction. The concept of integrating several complementary genomic data source was carefully considered by Endeavour [4]. Up till now, many strategies have been proposed for integrating multiple genomic data sources applied in gene prioritization. For instance, the heterogeneous biological data sources can be integrated intelligently using kernel-based data fusion strategy. The use of kernel methods is a smart and flexible approach because it decouples the original data from the machine learning algorithms by using an identical representation of the data as a kernel matrix. As a result, any prioritization strategy can access the same data and, as such, the data heterogeneity disappears.

The first kernel-based genomic data fusion approach for gene prioritization was proposed in [7]. In their framework, all the genomic data sources are first transformed into kernels using a linear function or a Radial Basis. Then, they proposed a multiple kernel learning (MKL) formulation for one-class SVM to perform gene prioritization. Accordingly, they examine whether the optimal convex optimization of the genomic kernels performs better than the uniformly weighted average of base kernels. We already discussed in [1] about the limitation of convex linear combination of kernels in dealing with fusion of different biological data sources that carry complementary information. It has been showed that working with such approaches could cause a loss of potentially useful information. Moreover, we investigated the limitation of MKL approaches in terms of their sensitive behavior in dealing with kernel weights.

In this study, we propose a kernel-based gene prioritization framework using geometric kernel fusion (GKF) which we have recently developed as a powerful tool for protein fold classification, the protein remote homology detection and protein sub-nuclear localization [1]. It has been shown that

GKF is less sensitive in dealing with complementary and noisy kernel matrices compared to standard multiple kernel learning methods. Since genomic kernels often encodes the complementary characteristics of biological data, this leads us to research the application of GKF in the gene prioritization task. In particular, this study presents kernel-based method for gene prioritization through fusing heterogeneous biological data sources by taking Log-Euclidean mean [8] of their corresponding kernel matrices instead of convex linear combinations.

We consider several genomic data sources including annotation-based data sources, protein-protein interaction networks and literature, as well as sequence-based protein features such as information extracted directly from position-specific frequency matrices (PSFM) and pairwise local sequence alignment.

One of the most controversial issues in gene prioritization strategies is the assessment strategy. It is necessary to validate gene prioritization machines in a more realistic task setting such that their respective performance fairly reflect their ability to hunt down the unknown key genes among the thousands of genes for the phenotype under investigation . Most of gene prioritization tools have been evaluated through cross validation experiment. This often leads to an overestimate of the real performance since many biological data sources. In particular, the biological annotation-based and literature-based data sources are contaminated with the prediction of the disease-gene annotations [9]. In order to overcome this issue we develop an unbiased and prospective benchmark based on the OMIM associations [10]. It is at the heart of our assessment strategy to: (i) train the model based on disease-associated genes discovered before a certain time using genomic data sources released prior to that time, and (ii) evaluate the model on the prediction of disease-associated genes reported afterwards.

The main features of our work are listed below

- The integration of various heterogeneous biological data sources using Geometric-Inspired kernel data fusion. In this paper we specifically consider the Log-Euclidean mean as a tool for fusing kernel matrices.
- Investigating the performance of evolutionary-based information for gene prioritization. We use a sequence-based protein feature based on information extracted directly from PSFM.
- We develop a heuristic weighted approach for integrating kernel matrices using Log-Euclidean mean of kernel matrices.
- Evaluate our gene prioritization model based on an unbiased benchmark.

Experimental results on our prospective benchmark show that our model can improve the accuracy of the state-of-the-art gene prioritization model. It does achieve to offer true novel genes for 13 diseases of our benchmark (80 diseases under investigation) in the top 24 prioritized genes out of about 12,000 human genes (top 0.2% of the human genome).

II. KERNEL-BASED DATA FUSION FOR GENE PRIORITIZATION

Early and late integration are typical approaches to fuse various protein data sources [11] , the heterogeneous biological data sources can be integrated intelligently and efficiently using partial integration, such as kernel-based data fusion strategies [7, 12]. The main notion behind kernel methods is to construct the same representation for all heterogeneous data sources and integrate these representations systematically. In fact, data heterogeneity interestingly vanishes by using a representation of the original data as a Symmetric positive definite (SPD) matrix -called kernel matrix- which encodes the similarity between samples in their respective input space.

There are several methods for obtaining a valid and fitting kernel by tuning the kernel matrices weights [5]. Finding such weights from training data and replacing the single kernel by a convex linear combination of weighted base kernels is often referred to as MKL. These weights reflect the relative potential importance of the different data sources in the combined kernel. Several MKL approaches have been proposed in the literature [12, 13], during the last decade. Several complicated convex optimization-based approaches have been proposed [12, 13], to improve the efficiency of kernel data fusion. They try to optimize the kernel weights based on different optimization criteria. However, on the one hand, the optimized weights of base kernels convex combination of kernel matrices often leads to mixed results and causes an improvement in performance only when dealing with redundant or noisy kernel matrices [12]. Indeed, this type of averaging has usually sensitive behavior in coping with kernel matrices containing complementary and non-redundant information and fails to completely capture all the information for these kernels. Since genomic kernels often encode the complementary characteristics of biological data, applying convex linear combination of base kernels is not very appropriate for biological application.

On the other hand, it has been showed that even the results obtained by employing uniformly weighted kernel fusion are comparable to the results of the best existing MKL approaches in general applications. This is also supported by the equal weights theorem [14], which states when all optimized weights are uniformly distributed on the interval [0.25; 0.75], the performance is hardly changed using equal weights.

Using the Euclidean distance on a convex cone whose interior contains all SPD matrices $P(n)$, we can obtain the arithmetic mean (AM) which is uniformly weighted average of the base kernels. By contrast, since it has been illustrated that this type of averaging fails to completely capture all the information for kernels containing complementary, non-redundant information, Euclidean distance on SPD matrices might not be appropriate. Moreover, SPD matrices form a convex cone and not a vector space. This has an effect on the "natural" geometry of SPD matrices, which may not be Euclidean, but rather should rely on concepts from Riemannian geometry [1]. This inspires us to consider other means between SPD matrices that are not relative to the Euclidean distance on $P(n)$ and necessarily a linear

combination of SPD matrices [1]. For example, the mean corresponding to Riemannian distance on $P(n)$ is the geometric mean.

III. GEOMETRIC KERNEL DATA FUSION

In this study, we design and develop methods to integrate kernel matrices by taking more involved, geometry inspired means of these matrices instead of convex linear combinations [1]. Such averaging of the base kernels can be interpreted as a kind of integration that expresses the nonlinear relationship between the individual kernels. In particular, we focus on taking the matrix Log-Euclidean mean of base kernels. For a general number of matrices, the fused kernel is obtained by taking the geometric mean

$$K = F(K_1, K_2, \ldots, Kn) = \mathrm{E}_{LE}(K_1, K_2, \ldots, Kn) \tag{1}$$

A. Geometric Inspired Means

Similar to the arithmetic mean (AM), other types of means of SPD matrices such as the harmonic mean (HM), Log-Euclidean mean (LogE) [8] and geometric mean (GM) result in SPD kernels. However, computing the geometric mean of a general number of SPD matrices is a challenge. In fact, for a general number of SPD matrices, a proper definition of a geometric mean with some natural properties has only recently been developed [15]. We already presented two methods for computing the geometric mean [1]. The first approach is focused on computing the actual geometric mean using the definition of the Karcher mean (Karcher-KF) [14]. The second, however, only computes a rough approximation of the actual geometric mean using a proposed, heuristic method based on Arithmetic-Geometric-Harmonic (AGH) mean.

However, computing the geometric mean for a general number of SDP matrices is hard and computationally expensive, which is why we discuss here the Log-Euclidean mean (LogE) [8]. In fact, even the approximate geometric mean needs a computational cost of the order $n^2 \log(n)k$ per each iteration, where n and k denote the size and the number of kernel matrices respectively.

B. Log-Euclidean Mean

The Log-Euclidean mean can be considered as a consensus between the arithmetic and geometric mean. The Log-Euclidean mean of n kernel matrices can be obtained explicitly as [8]:

$$K_{LE}(K_1, \ldots, K_n) = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log(K_i)\right) \tag{2}$$

It can also be seen as a generalization of the notion of geometric mean. For n positive numbers $x_1, \ldots, x_n$, their geometric mean can be stated as below

$$G(x_1, \ldots, x_n) = (x_1, \ldots, x_n)^{1/n} = \exp\left(\frac{1}{n}\left(\sum_{i=1}^{n} \log(x_i)\right)\right) \tag{3}$$

C. The Links of Log-Euclidean with Arithmetic Mean and Geometric Mean

For SPD matrices, the Log-Euclidean mean can be seen an approximation to the geometric mean by considering both as an adaptation of the arithmetic mean towards positive definite matrices. Indeed, for a given set of kernel matrices $K_1, \ldots, K_n$, the arithmetic mean is given by

$$\mathrm{A}(K_1, \ldots, K_n) = \frac{1}{n}\sum_{i=1}^{n} K_i \tag{4}$$

It can also be stated as below:

$$\mathrm{A}(K_1, \ldots, K_n) = \min_{X \in \Re^{n \times n}} \sum_{i=1}^{n} \|X - K_i\|_F^2 \tag{5}$$

Here, each term in the cost function expresses the Euclidean distance between matrices X and Ai, and the minimizer will be exactly the arithmetic mean as we know it. Now, by switching to positive definite matrices, we restrict the search space of X to this set P_n and the Euclidean distance is replaced with the Riemannian distance between positive definite matrices. The new optimization problem has the geometric matrix mean (the Karcher mean) as its minimizer:

$$G(K_1, \ldots, K_n) = \min_{X \in P_n} \sum_{i=1}^{n} \left\| \log(K_i^{-1/2} X K_i^{1/2}) \right\|_F^2 \tag{6}$$

This shows the relation between the arithmetic and the geometric mean.

By contrast, there is an obvious one-to-one relation between the (Euclidean) vector space of symmetric matrices and the positive definite matrices using the matrix exponential and logarithm. This relation allows an interesting interpretation for the Log-Euclidean mean: we start with the positive definite matrices $(K_1, \ldots, K_n)$ which we transform to symmetric matrices using the matrix logarithm. Of these transformed matrices the arithmetic mean is taken, and finally, the result is transformed back to the positive definite matrices using the matrix exponential.

D. Weighted Log-Euclidean Mean

A weighted version of the Log-Euclidean (W-LogE) mean of n kernel matrices can be obtained simply by

$$K_{LE}(K_1, \ldots, K_n) = \exp\left(\frac{1}{n}\sum_{i=1}^{n} (\beta_i) \log(K_i)\right) \tag{7}$$

Where $\langle \beta_i \rangle_{i=1}^{n}$ are arbitrary positive weights such that $\sum_{i=1}^{n} \beta_i = 1$. We already offered a philosophical critique of the optimizing the kernel weights within biological applications, in particular where they encode the complementary characteristics of biological data. It has also been observed that geometric inspired means of kernel matrices are less sensitive in dealing with complementary and noisy kernels compared to typical multiple kernel learning approaches. Nevertheless, we are still interested in investigating the potential usefulness of weighting kernels for

our proposed kernel-based gene prioritization through geometric kernel data fusion.

The weights of kernels can be interpreted as their corresponding importance in the fused kernel. It is expected that the kernels that carry more information than other kernels receive higher weights in the fused kernel. We propose a heuristic and simple approach for combining kernels, which assigns the kernels weights based on the performance values obtained by each kernel separately. The basic idea behind of our approach is that a kernel matrix K_i that is more successful in hunting disease-associated genes among the top prioritized gene, in other words, a kernel with higher recall or true-positive rate (TPR) among the top rated genes, should contribute more to the fused kernel. Obtaining the weight of each kernel is formulated as

$$\beta_i = \frac{recall_{TopT}(K_i) - \delta}{\sum_{n}^{i=1} recall_{TopT}(K_i) - n\delta} \quad (8)$$

Where δ is a threshold less than the minimum of the TPR at the top T prioritized genes obtained from single-kernel performance. We set δ equal to 1/n of the minimum TPR. For each disease, its corresponding kernels performance in the top T prioritized genes is computed using leave-one-out cross validation on the training genes. If there is a kernel with 0 TPR at the top T prioritized genes, then we set δ equal to zero. If $recall_{TopT}$ for all kernels is zero, then each kernel will be assigned the same kernel weight. The parameter T is chosen using leave-one-out cross validation on the training genes. We search over a grid of top 10, 20, 40, 80 and 160 ranked gens. Here, we assign kernel weights based on their performance in hunting disease-genes in the top 40 ranked genes (top 0.33% of the human genome) because it gives the best average recall over all diseases on the validation part of training genes.

.

IV. MATERIAL AND METHODS

A. Benchmark and Assesment strategy

We developed a prospective benchmark based on the OMIM associations. For this we used an old version from OMIM (back in 2010), and a newer one (recent in 2013). Both versions have been automatically adjusted in order to have similar disease entries for which there is at least one delta gene (reported after 2010). Afterwards, we manually verified the delta genes which we used in our prospective analysis. OMIM released 6733 experimental-based disease-gene associations in 2013. Finally, we extracted 80 diseases that have at least 3 genes in the 2010 version and have at least one gene reported after 2010. This leads to 946 as the total number of disease-associated genes reported in OMIM 2013, which corresponds to a total of 727 genes listed in 2010, and a total of 219 genes reported after 2010. The 80 diseases that we tested in our study are listed in Supplementary Table I.

We first train our gene prioritization model on disease-genes annotation reported prior to 2010 using the genomic data sources constructed before 2010, and then we evaluate our model on the prediction of disease-associated genes discovered between 2010 and 2013.

B. Genomcis Data Kernels

We consider several genomic data sources including annotation-based data sources such as Gene Ontology (GO) and Swiss-Prot (SW) annotation, protein-protein interaction network based on STRING, and literature (TEXT), just as in [4], as well as sequence-based protein features such as information extracted directly PSFM and pairwise local sequence alignment-based feature spaces (based on Blast scores) (BLAST).

Sequence evolution information- It has been shown that the sequence evolution information extracted directly from PSFM or position specific scoring matrices (PSSM) profiles can be very helpful for challenging tasks in bioinformatics such as protein fold recognition and protein localization [1].

A protein sample P with L amino acid residues can be represented by its evolutionary information through PSSM or

PSFM, which both have L columns and 20 rows. Each row of PSSM $(M_{i\rightarrow})$ represents the log-likelihood of the residue substitution at the corresponding position in the protein sequence. In particular, the *(i, j)*th entry of the PSSM matrix $(M_{i\rightarrow})$ represents the possibility of the amino acid type *j* appearing in the *i*-th position of the protein domain during the evolution process. The PSSM entries are obtained using the PSI-BLAST program to search versus the non-redundant protein database, like the Swiss-prot database, through three iterations with the E-value cut-off set to 0.001. We use a 20 dimensional feature vector created by summing up each column in the PSFM profile and dividing by the length of the gene as a common profile-based representative models of genes.

In our framework, all the genomic data sources are first transformed into kernels using a linear function followed by normalization, except for STRING and PSFM. For PSFM, we use a Radial Basis Function (RBF) with kernel width equal to twice the average distance of a data point to its nearest neighbor in the union of the training and the test set. For STRING, we use two different kernel functions: (i) Tanimoto kernel and (ii) diffusion kernel with parameter β tuned using leave-one-out cross validation on training data.

Next, to see the advantage of fusing heterogeneous genomic data sources for gene prioritization through kernel data fusion, we focus on combining all kernel matrices derived from each view on human genes. The kernel matrices are fused through LogE, AM and weighted version of LogE. Accordingly, the fused kernel is used to determine the performance.

Same as in [7], we exploit the one-class support vector machine (SVM) approach to perform gene prioritization. As a result, a one-class SVM algorithm is applied to perform novelty detection using the fused kernel. For each disease, a one-class SVM algorithm uses the fused kernel to seek the

hyperplane that best separates the training genes from the origin.

V. RESULTS

We report the average TPR (recall) results, when considering the top 5%, 10%, and 30% of the ranked genes. The performance of the individual genomic kernels on our prospective benchmark is listed in Table I.

According to Table I, among all genomic data sources, the TEXT performs best when measuring the recall among the top ranked human genes. This implies that TEXT contains the most important and consistent information relevant to disease gene hunting.

Next, to see the advantage of fusing heterogeneous data sources for gene prioritization through our proposed data integration setup, we focus on combining kernel matrices derived from each view on human genome. The kernel matrices are combined through AM, LogE and the weighted version of LogE (weighting kernels based on their TPR among the top 40 prioritized genes). Afterwards, the combined kernel is used to determine the performance. Table II lists the recall of our proposed kernel fusion approaches based on averaging of the kernel matrices.

TABLE I. A COMPARISON OF THE AVERAGE TPR RESULTS FOR THE INDIVIDUAL GENOMIC KERNELS AT THE TOP 5%, 10%, AND 30% PRIORITIZED GENES (AVERAGE TPR OVER ALL DISEASES).

Genomic Data sources	TPR results for at top of single data sources		
	TPR in top 5%	*TPR in top 10%*	*TPR in top 30%*
TEXT	**35%**	**47%**	**72%**
SW	32%	39%	59%
GO	34%	38%	60%
BLAST	14%	25%	51%
STRING (Tanimoto)	27%	33%	48%
STRING (Diffusion)	30%	35%	45%
PSFM	22%	27%	47%

In order to compare the results of our proposed methods on our prospective benchmark with the gene prioritization tool, already proven to be successful, we run Endeavour with our OMIM disease-genes annotation reported until 2010 using genomic data sources released prior to 2010. Endeavour is trained using genomic data sources listed in Table I except the PSFM. Table II also provides the performance of Endeavour on our prospective benchmark.

According to Table II, TPR of all our proposed kernel-based data fusion for gene prioritization in top 5%, 10% and 30% is consistently outperforming Endeavour. We observe that using LogE results in an average TPR of 40%, 52%, and 78 % at the top 5%, 10%, and 30% ranked genes, respectively. This compares to 46%, 52%, and 68% using AM. This indicate that, however, AM achieves better TPR at the top 5% prioritized genes, LogE achieves higher TPR at the top of 10% and 30% ranked genes. Our results show that the weighted version of LogE propose higher recall at the top 5%, 10%, and 30% prioritized genes.

TABLE II. A COMPARISON OF THE AVERAGE TPR RESULTS FOR OUR PROPOSED KERNEL-BASED DATA FUSION APPROACHES FOR GENE PRIORITIZATION AND ENDEAVOUR AT THE TOP 5%, 10%, AND 30% PRIORITIZED GENES

Methods	TPR results for at top of prioritized genes		
	TPR in top 5%	*TPR in top 10%*	*TPR in top 30%*
LogE	40%	52%	78%
W-LogE	**47%**	**54%**	**79%**
AM	46%	52%	68%
Endeavour	40%	49%	71%

Table 3 shows the results of incorporating the sequence evolution information extracted from PSFM to our proposed kernel-based data fusion for gene prioritization. According to Table III, while the ability of uniformly weighted kernel integration approach in hunting disease genes at top 5%, 10% and 30% decreases continuously, our geometric-inspired kernel data fusion model can improve their average TPR at top 5%,10% and 30% prioritized genes, giving promising average recall of 49% in top 5% prioritized genes. According to Table I, the PSFM kernel carries less information in comparison with other biological kernels. Table 3 suggests that the geometric-based averaging of kernel matrices is less sensitive for dealing with a noisy kernel and that the performance of both LogE and W-LogE increase as a kernel which carries much less information is integrated. Contrary to these approaches, the linear-based averaging of kernel matrices leads to mixed results and its performance on top 5 and 10 decreases if we add a less informative kernel. We can also discuss that although PSFM kernel is not very successful in hunting disease-accosted genes, it still carries complementary information when it is integrated with other biological data in an intelligent way, which could improve the performance of the gene prioritization task. Table 3 shows that W-LogE can succeed in performing that intelligent combination and can exploit such a complementary biological kernel. However, the convex linear combination of kernel fails to capture this information.

TABLE III. THE RESULTS OF INCORPORATING THE SEQUENCE EVOLUTION INFORMATION TO OUR PROPOSED KERNEL-BASED DATA FUSION FOR GENE PRIORITIZATION AT THE TOP 5%, 10%, AND 30% PRIORITIZED GENES

Methods	TPR results for at top of prioritized genes			*p-value of gene ranking*
	TPR in top 5%	*TPR in top 10%*	*TPR in top 30%*	
LogE	42%	53%	**80%**	2.1813e-18
W-LogE	**49%**	**55%**	**80%**	5.9909e-20
AM	44%	49%	69%	9.0616e-15

Table 3 also provides the p-value for our results obtained with statistical analysis using Kolmogorov-Smirnov test.

Clearly, our ranking results are highly significant for all our proposed kernel-based approaches.

VI. Conclusion

In this study, we develop a novel kernel data fusion framework using Log-Euclidean kernel data fusion for gene prioritization. We present a heuristic approach for weighted Log-Euclidean kernel data fusion.

We also develop an unbiased and prospective benchmark based on the OMIM associations. Experimental results on our prospective benchmark demonstrate that our geometric-inspired model can improve the accuracy of the state-of-the-art gene prioritization model. Our geometric-inspired kernel data fusion model for gene prioritization is developed further by incorporating the sequence evolution information into our kernel data fusion framework, giving average recall of 49% in top of 5% prioritized genes. It succeeds to propose true novel genes for 13 diseases of our benchmark in the top 24 prioritized genes out of about 12,000 human genes (top 0.2% of the human genome) considered in this study.

VII. Acknowledgment

This research is supported by KUL GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, KUL IOF 3M120274 Immunosuppressive drugs. IWT TBM Haplotyping. Hercules Stichting: Hercules III PacBio RS. iMinds: SBO 2013. EU COST Action BM1006: NGS Data analysis network. FCT Neuroclinomics. The scientific responsibility is assumed by its authors.

References

[1] P. Zakeri *et al*., "Protein fold recognition using geometric kernel data fusion, Bioinformatics," vol. 30(13), 2014, pp. 1850-1857.

[2] J. Amberger *et al*., "A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®)," Hum Mutat, vol. 32, 2011, pp. 564-567.

[3] C. Perez-Iratxeta *et al*., "Association of genes to genetically inherited diseases using data mining," Nat Genet vol. 31, 2002, pp.316–9.

[4] S. Aerts *et al*., "Gene prioritization through genomic data fusion. Nat Biotech," vol. 24(5), 2006, pp. 537–544.

[5] J. Chen *et al*., "Improved human disease candidate gene prioritization using mouse phenotype," BMC Bioinformatics, vol. 8(1), 2007, p.392.

[6] J.E. Hutz *et al*., "CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genetic Epidemiology," vol. 32(8), 2008, pp. 779–790.

[7] T. .De Bie *et al*., "Kernel-based data fusion for gene prioritization," Bioinformatics , vol 23(13), 2007, pp. i125-i132.

[8] V. Arsigny *et al*., "Geometric means in a novel vector space structure on symmetric positive-definite matrices," SIAM J. Matrix Anal. Appl., vol. 29, 2007, pp. 328–347.

[9] D. Börnigen1 *et al*., "An unbiased evaluation of gene prioritization tools," vol. 28(23), 2012, pp. 3081-3088.

[10] S. ElShal *et al.,* "Beegle: from literature mining to disease-gene discovery," Nucleic Acids Res. 2015, gkv905v1-gkv905.

[11] P. Zakeri *et al*., "Prediction of protein submitochondria locations based on data fusion of various features of sequences," J. Theor. Biol., vol. 269, 2011, pp. 208–216.

[12] G.R.G. Lanckriet *et al*., "A statistical framework for genomic data fusion,". Bioinformatics, vol. 20, 2004, pp. 2626–2635.

[13] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," J. Mach. Learn. Res., vol. 12, 2011, pp. 2211–2268.

[14] H. Wainer, "Estimating coefficients in linear models: it don't make no nevermind,".Psychol. Bull., vol. 83, 1976, pp. 213–217.

[15] R. Bhatia, Positive Definite Matrices. Princeton University Press, Princeton Series in Applied Mathematics., 2007.

Supplementary Table I: The 80 diseases that we tested in our study

#	diseases	#training genes	#test genes
1	Amyloidosis	6	1
2	Aortic aneurysm	3	1
3	Autoimmune lymphoproliferative syndrome	3	1
4	Bardet-Biedl syndrome	16	1
5	Brachydactyly	7	1
6	Cataract	24	4
7	Ceroid lipofuscinosis	8	2
8	Ciliary dyskinesia	9	10
9	Coenzyme Q10 deficiency	5	1
10	Combined oxidative phosphorylation deficiency	5	10
11	Congenital disorder of glycosylation	24	9
12	Congenital heart defects	1	1
13	Conotruncal heart malformations	5	1
14	Cornelia de Lange syndrome	3	2
15	Cowden syndrome	3	2
16	Cutis laxa	6	1
17	Deafness	57	16
18	Diamond-Blackfan anemia	7	3
19	Dyskeratosis congenita	5	2
20	Dystonia	11	2
21	Epidermal nevus	3	1
22	Epidermolysis bullosa	8	1
23	Hypochromic microcytic anemia	6	1
23	Exudative vitreoretinopathy	3	1
25	Fanconi anemia	13	3
26	Glycogen storage disease	16	1
27	Hermansky-Pudlak syndrome	8	1
28	Heterotaxy	4	1
29	Holoprosencephaly	6	1
30	Hypertriglyceridemia	3	1
31	Hypogonadotropic hypogonadism with or without anosmia	11	9
32	Hypomagnesemia	5	1
33	Hypophosphatemic rickets	4	1
34	Hypothyroidism	5	2
35	Ichthyosis	10	4
36	Left ventricular noncompaction	42	4
37	Leigh syndrome due to mitochondrial complex deficiency	14	5
38	Cardiac arrhythmia	11	1
39	Lymphoproliferative syndrome	3	1
40	Maple syrup urine disease	3	1
41	Meningioma	4	2
42	Microcephaly	8	7
43	Microphthalmia with coloboma	3	1
44	Migraine	6	1
45	Mitochondrial complex IV deficiency	21	5
46	Myelogenous leukemia	5	2
47	Centronuclear myopathy	26	4
48	Joubert syndrome	14	13
49	Nephrotic syndrome	5	2
50	Neurodegeneration with brain iron accumulation	3	1
51	Noonan syndrome	5	1
52	Obesity	24	1
53	Albinism	5	1
54	Paragangliomas	4	1
55	Peroxisome biogenesis disorder	9	1
56	Pheochromocytoma	6	2
57	Premature ovarian failure	9	1
58	Progressive external ophthalmoplegia with mitochondrial DNA deletions	3	1
59	Senior-Loken syndrome	4	1
60	Short stature	3	2
61	Spastic paraplegia	18	13
62	Spherocytosis	16	2
63	Amyotrophic lateral sclerosis	28	14
64	Metaphyseal anadysplasia	12	2
65	Stickler syndrome	4	1
66	Surfactant metabolism dysfunction	4	1
67	Tetralogy of Fallot	4	2
68	Thrombocytopenia	9	2
69	Usher syndrome	9	2
70	Ventricular tachycardia	3	1
71	Weill-Marchesani syndrome	2	1
72	Malaria	15	1
73	Bone marrow failure	4	2
74	Breast-ovarian cancer	2	2
75	HIV type	14	1
76	Multiple sclerosis	3	2
77	Mycobacterium tuberculosis	9	1
78	Psoriasis susceptibility	1	2
79	Schizophrenia	9	2
80	Stroke	8	1