Data Type Integration for Protein Identification using Kernel Based Classification Methods

Ito Wasito[1*], Hadaiq R Sanabila[2] and Aulia N Istiqlal[3]

[1] Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia; *Corresponding author

[2] Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia

[3] Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia

Abstract

The integrated biological data is expected to obtain a higher exactness, better performance and greater robustness compared to single dataset. In this work, we present data integration using kernel-based approach to identify protein class in yeast, ribosomal proteins and membrane proteins. By using intermediate stage of integration, we change the single data source into kernel matrix format. Kernel weighting was used in the establishment of integrated data. We propose three weighting methods approach i.e. KTA (Kernel Target Alignment), FSM (Feature Space-based kernel matrix evaluation Measure), and AI (Alignment Index). We also perform the combination of these three methods. These integrated kernels will be analyzed using Support Vector Machine (SVM). Our proposed data integration methods achieve a higher performance compared to single data source. KTA is the best kernel weighting measurement method and always obtain a better performance to recognize membrane and ribosomal proteins classes than others.

Keywords: *Data Integration, Kernel Matrix, Kernel Target Alignment, Feature Space-based kernel matrix evaluation Measure, Alignment Index, Support Vector Machine.*

1. Introduction

Protein is one of the crucial compounds in living organism. Proteins take an important role in cell signalling, immune responses, cell adhesion, metabolic pathway of cells and the cell cycle. It consists of one or more amino acid monomer which folded together by polypeptides. Several proteins functions are as energy source, cell and tissue constructions, as main hormone and enzyme establishment, and as acid-base cell regulator. The structure and functions of proteins are reviewed in the field of proteomics. Proteomics can be defined as the qualitative and quantitative comparison of proteomes under various conditions to further untangle biological processes.

Nowadays, with the rapid technological advances, biological data with assorted structures, measurements, formats, and sizes have become openly available. Biological data are stored in various formats and structures files such as sequence, vector, and graph. Each biological data type such as in protein, has different and independent perspective of the whole genome/protein. The difference perspectives of the protein are the consequence of the various protein measurements. In order to obtain whole view of the biological data, data integrations are employed. The integrated data is expected to obtain a higher exactness, better performance and greater robustness compared to single dataset. Furthermore, it can be used to compare and evaluate experimental results from various datasets and measurements. In the near future, bioinformatics research will concern on data fusion methods in various approaches [1].

In this research, we present data integration using kernel-based approach. Many types of data can be represented by kernel matrix. Kernel matrix transforms the similarity /relations measures among the data point within input space into numerical data. By using intermediate stage of integration, we change the single data source into kernel matrix format before we combine them as integrated data [2]. Multiple Kernels Learning (MKL) is one of the Machine Learning terms which discusses the fusion of multiple kernel matrices. Multiple kernels which originated from heterogeneous single data are fused using linear combination. The information quality contained in each data format has various levels. Kernel weighting was used so that the kernel which has better information

quality has a larger portion in the establishment of integrated data. In this research, we measure the performance of these methods for recognizing ribosomal and membrane protein in yeast. The ribosome is responsible for mRNA translation into the certain amino acid sequence through the general genetic code [3][4]. Meanwhile, membrane protein is a protein which associated with the cells membrane. Its function including to assure the cell stability, get involved in immune response, produce significance material for cell function, maintain the ion concentration, and manage the connections between internal and external cell environment.

On the previous research, Lankcriet et al [1] using semi-definite programming framework to obtain a set of weights μ_i which reflects the quality of different information sources from the various kernel matrices. Malossini et al [5] consider using von Nuemann entropy to measures the quality of kernel matrix. The entropy value is purely associated to the notion of data sparseness. The higher von Nuemann value means the kernel matrix has better quality. Ying et al [6] consider using information-theoretic technique based on a Kullback-Leibler (KL) divergence. It measures the difference between output and input kernel matrix. In this research, we consider using three weighting methods approach i.e. Kernel Target Alignment (KTA) [7], Feature Space-based kernel matrix evaluation Measure (FSM) [8], and Alignment Index (AI) [9]. We also perform kernel weighting using the combination of these methods to obtain precise kernel weight. Furthermore, these integrated weighted kernels will be analyzed using SVM.

2. Methods

2.1 Kernel Methods

Kernel methods perform a mapping from the input space into higher dimensional space [10][11][12][13]. This method provides the way to merge and integrate different type of data. Kernel represents the similarity or relations measures among the data point. For pair of data x_1 and x_2, denote their embedding as $\Phi(x_1)$ and $\Phi(x_2)$, respectively. We define the embedded data inner product, $\langle \Phi(x_1), \Phi(x_2) \rangle$, through a kernel function/operator $K(x_1, x_2)$ [1].

In this research, we used 7 types of data with various similarity measures. For sequence data, there are 3 kernel matrices to be further analyzed. The first two kernels (K_{SW} and K_B) are constructed using Smith-Waterman [14] and BLAST [15]. The last sequenced based kernel is K_{HMM}. It is contain the pairwise comparison score which derived from HMM (Hidden Markov Models) in protein families (Pfam) database [16]. The fourth kernel (K_{FFT}) contains the information of the hidrophobicity pattern. This pattern is extracted using FFT kernel. Furthermore, for the protein interaction data, there are two kernels i.e. K_L and K_D. Meanwhile, we employed radial basis kernel for gene expression data. Gene expression is required to distinguish ribosomal proteins. The kernels details are depicted in table 1.

Table 1 : The kernel list

Kernel Data	Data Type	Similarity Measure
K_{SW}	sequences	Smith-Waterman
K_B	sequences	BLAST
K_{HMM}	sequences	Pfam HMM
K_{FFT}	hydropathy profile	FFT
K_L	Protein-protein interactions	linear kernel
K_D	Protein-protein interactions	diffusion kernel
K_E	Microarray gene expression	radial basis kernel

2.2 Kernel Weighting

The goodness of kernel matrix reflects the information quality of the data. There are many methods to evaluate the kernel matrix quality, especially for classification. Several methods which generally used as kernel matrix evaluation are negative log-posterior [17], regularized risk [11], and hyperkernels [18]. These evaluations only maintain certain standard in form of regularities in particular spaces and do not give a particular score. However, for evaluation, these kernel matrix measurements need an optimization routine. Therefore, these measurements are high-priced to be integrated with other costly process such as feature and model selections. That is the reason why kernel matrix measurement must be efficiently and effectively calculated before used in feature and model selection.

2.3 Kernel Target Alignment (KTA)

Kernel Target Alignment (KTA) is one of efficient kernel measurements which generally employed. This method was proposed by N.Cristianini [7] in 2002. Because of its simplicity and effectiveness, KTA has been used in several tasks for two fundamental problems in kernel methods i.e. learning kernels from data and designing

kernels. KTA is used to evaluate the degree of kernel matrix aligns to its target [7]. The degree of kernel matrix aligns to its target is defined as the normalized Frobenius inner product among the kernel matrix (K) and the target vector covariance matrix ($t.t^T$). This alignment interpreted as cosine distance between these two bi-dimensional vectors.

Denote the sample set $\{x_i\}_{i=1...n} \in X$ with the corresponding target vector $t = \{t_1, t_2, ... ,t_n\} \in \{-1,1\}^n$. Frobenius inner product of K and K* is calculated as formula 1.

$$\langle K, K^* \rangle_F = \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij} k^*_{ij} \quad (1)$$

Given kernel K and target t, the KTA value defined as

$$KTA(K,t) = \frac{\langle K, t.t^T \rangle_F}{\sqrt{\langle K, K \rangle_F \langle t.t^T, t.t^T \rangle_F}} \quad (2)$$

The value of KTA resides between 0 and 1 ($0 \leq$ KTA (K, t) ≤ 1). The two bi-dimensional vectors K and $t \cdot t^T$ are linear when KTA (K, t) = 1. The higher KTA value of common kernel matrix, it contains a higher information quality.

2.4 Feature Space-based kernel matrix evaluation Measure (FSM)

Nguyen et.al .in 2008 [8] proposed Feature Space-based kernel matrix evaluation Measure (FSM). The idea of this kernel goodness measurement is using data distribution in a feature space. There are two factors which consider in this measurement, the within-class variance in the direction of among class centers and the gap among the class centers. The illustration of these factors depicted in figure 1. FSM defines the proportion of the summed within-class standard deviation in the direction between the class centers to the gap distance among the class centers.

$$FSM(K,t) \underline{\underline{def}} \frac{std}{\|\phi_- - \phi_+\|} \quad (3)$$

Where the summed within-class standard deviation of class$^+$ and class$^-$ in the direction is based on formula 4.

$$std = \sqrt{\frac{\sum_{i=1}^{n_+} \langle \phi(x_i) - \phi_+, e \rangle^2}{n_+ - 1}} + \sqrt{\frac{\sum_{i=n_+ +1}^{n_-} \langle \phi(x_i) - \phi_-, e \rangle^2}{n_- - 1}} \quad (4)$$

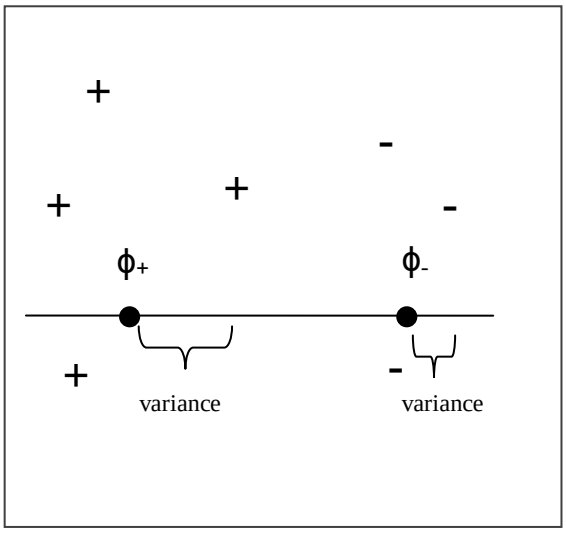

Fig. 1 Data distribution illustration in the feature space

The center of a class denote as the mean of the sample data in particular class in the feature space. Hereafter, the center of class$^+$ is $\phi_+ = \sum_{i=1}^{n_+} \frac{\phi(x_i)}{n_+}$ and the center of class$^-$ is $\phi_- = \sum_{i=n_+ +1}^{n_-} \frac{\phi(x_i)}{n_-}$. Furthermore, $e = \frac{\phi_- - \phi_+}{\|\phi_- - \phi_+\|}$ denote as the unit vector in the way among two class centers. The FSM (K, t) ≥ 0 and smaller FSM value of particular kernel matrix K, the better quality of kernel matrix K is.

2.5 Alignment Index (AI)

Another kernel matrix measurement is alignment index. This method was proposed by Tang et al in 2010 [9]. The degree of matching among a kernel matrix and its target vector consider as the idea of this measurement. For kernel K and target t with the n number of data, the alignment index defined as formula 5.

$$AI(K, tt') = \frac{t' K t}{n\|K\|} \quad (5)$$

2.6 Multiple Kernel Integration

Each data which comes from heterogeneous formats transforms into kernel matrix. The entry of kernel matrix encodes the certain impression of one protein resemblance to another. Primary algebraic operations as well as addition, multiplication and exponentiation preserve the fundamental positive semi-definiteness property. Hence, it is enable simple and powerful kernels algebra [1]. In

order to integrate several kernel matrices, we can use this property with the kernel weight value greater equals to zero, to assure the kernel positive definiteness.

For a set of kernel K $\{K_1, K_2,...,K_n\}$, we can construct the linear kernel combination with the kernel weight μ_i, i=1,...,n.

$$K = \sum_{i=1}^{n} \mu_i K_i \quad (6)$$

We used several kernel weighting methods and the combinations to measure the goodness which represent the information quality of kernel matrix. The brief explanations of linear kernel combination for each kernel weight are described in the formula below.

$$K_{IntegKTA} = \sum_{i=1}^{n} \mu_{KTAi} K_i \quad (7)$$

$$K_{IntegFSM} = \sum_{i=1}^{n} \frac{K_i}{\mu_{FSMi}} \quad (8)$$

$$K_{IntegAI} = \sum_{i=1}^{n} \mu_{AIi} K_i \quad (9)$$

$$K_{Integ(KTA\&AI)} = \sum_{i=1}^{n} (\mu_{KTAi} + \mu_{AIi}) K_i \quad (10)$$

$$K_{Integ(KTA\&FSM)} = \sum_{i=1}^{n} (\mu_{KTAi} + \frac{1}{\mu_{FSMi}}) K_i \quad (11)$$

$$K_{Integ(FSM\&AI)} = \sum_{i=1}^{n} (\frac{1}{\mu_{FSMi}} + \mu_{AIi}) K_i \quad (12)$$

$$K_{Integ(KTA\&FSM\&AI)} = \sum_{i=1}^{n} (\mu_{KTAi} + \frac{1}{\mu_{FSMi}} + \mu_{AIi}) K_i \quad (13)$$

The integrated kernels are attempts to be further analyzed using SVM. The illustration of kernel matrix integration is depicted in figure 2.

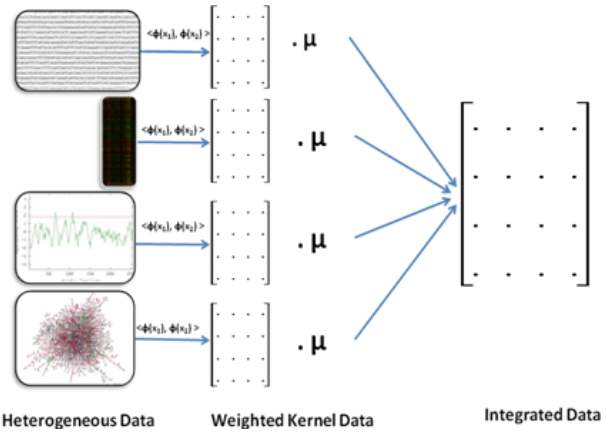

Fig. 2 The illustration of Kernel matrix integration

3. Results

Yeast protein data that we analyzed is proposed in [1] where heterogeneous data sources are combined together to improve the recognition of ribosomal and membrane protein. We divided the data into training and testing set in certain ratio randomly. Furthermore, we repeated the entire procedure 10 times and computed the average result. Particular ratio of training data is used in the training phase and the remaining ratio used as testing set. In this research, we used 60%, 70%, 80%, and 90 % of dataset as a training set. As a performance measurement, we used area under ROC curve (AUC). AUC value represents the chance that a classifier will place an arbitrarily takes positive case higher than an arbitrarily takes the negative.

In order to integrate various kernel matrices using linear combination, it should be made to be comparable. Therefore the centering and normalization should be conducted. Centering is a process to translate each sample $\phi(x)$ on the center of mass $\phi_C = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$ of the dataset. Meanwhile, normalization consists in normalizing the norm of mapped samples $\phi_N(x) = \frac{\phi(x)}{\|\phi(x)\|}$.

In the table 2 are illustrated the kernel weight of each kernel matrix. The best kernel matrix in KTA is linear kernel for ribosomal and diffusion kernel for membrane. Meanwhile for the FSM, the best kernel matrix is gene expression kernel for ribosomal and Pfam HMM kernel for membrane. Furthermore, the AI best kernel matrix for

Table 2 : The Weight of each kernel

Kernel	Membrane			Ribosomal		
	KTA	FSM	AI	KTA	FSM	AI
KerB	0.0392	14.4144	0.04051	0.005	1.4119	0.005195
KerD	0.1764	4.8498	0.18251	0.1059	1.9648	0.1103
KerE	0.0178	4.6662	0.026231	0.2184	0.4955	0.28802
KerFFT	0.0188	2.7279	0.020579	0.1205	0.9784	0.13028
KerHMM	0.0628	2.719	0.071671	0.0684	2.3946	0.079161
KerL	0.0726	4.3381	0.072823	0.3571	2.7652	0.36016
KerSW	0.0021	4.8833	0.014811	0.0035	2.4296	0.022252

Table 3 : The average AUC score in particular training-testing configuration

Kernel Weight	Average AUC Score							
	60-40		70-30		80-20		90-10	
	Ribosomal	Membrane	Ribosomal	Membrane	Ribosomal	Membrane	Ribosomal	Membrane
Integ(KTA)	***0.961197***	***0.90989***	***0.963154***	***0.912288***	***0.967815***	***0.91056***	***0.979645***	***0.9171608***
Integ(FSM)	0.959972	0.772312	0.962036	0.775736	0.96644	0.774345	0.978297	0.7725877
Integ(AI)	0.9606	0.772303	0.962653	0.775731	0.967003	0.774349	0.97894	0.7726249
Integ(KTA&FSM)	0.959963	0.772308	0.962037	0.775738	0.96644	0.774345	0.97823	0.7725989
Integ(KTA&AI)	0.959972	0.772309	0.962036	0.775732	0.96644	0.774337	0.97823	0.7725919
Integ(FSM&AI)	0.959963	0.772306	0.962036	0.775727	0.966459	0.774351	0.97823	0.7726128
Integ(KTA&FSM&AI)	0.959977	0.772309	0.962028	0.775746	0.96644	0.774358	0.97823	0.7725665

ribosomal is linear kernel and the best kernel matrix for membrane is diffusion kernel.

For the 60%-40% training-testing configuration, the higher average AUC value is obtained by KTA in ribosomal and membrane. AUC score of KTA for ribosomal is 0.961197 and for membrane is 0.90989. Furthermore, for the 70%-30% training-testing configuration, KTA obtains the higher AUC score in ribosomal and membrane specifically 0.963154 and 0.912288. KTA also obtains the higher AUC score for the 80%-20% training-testing configuration in ribosomal and membrane. The AUC score for ribosomal is 0.967815 and for membrane is 0.91056. Meanwhile, for the 90%-10% training-testing configuration, KTA obtains a higher AUC score as well. The AUC score for ribosomal is 0.979645 and for membrane is 0. 9171608. The complete average AUC score of each kernel weight are illustrated in table 3.

The KTA AUC score for the ribosomal is higher than the other kernel weighting methods even it slightly. However, for the membrane, KTA AUC score is superior compared to FSM, AI, and its combinations. When the KTA combines with the other kernel weighting methods, it decreases the AUC scores. It is verify that KTA is suitable method for measure the goodness of kernel matrix. Besides that KTA is not proper to combine with other kernel weighting methods because it will decrease the information quality in the integrated kernel matrix.

4. Discussion

We have conducting a kernel based method for integrating heterogeneous genome data, especially for recognizing ribosomal and membrane protein. The information quality and goodness contained in each data format has various levels. Kernel weighting was used to overcome it, so that the kernel which has better information quality has a larger portion in the establishment of integrated data. Kernel goodness measurement such as KTA (Kernel Target Alignment), FSM (Feature Space-based kernel matrix evaluation Measure), and AI (Alignment Index) are employed as a kernel weighting methods. Besides that, the combinations

of these kernels weighting methods are conducted to obtain the best integrated kernel matrix.

When using KTA as kernel weight, the classifier gives a better performance compared to the other kernel matrix measurements. When the combination of kernel weight was employed, the classifier performance is decreased. FSM and AI are improper kernel weighting methods which used to linear combination of kernel matrix. It is lead to the unfit integrated kernel of heterogeneous data. Even though KTA represents the goodness and the quality of a kernel matrix, when it combines with the other kernel weighting methods, it will decrease the classifier performance. KTA method is properly used as kernel weight in linear kernel combination solely.

5. Conclusions

Many types of data can be represented by kernel matrix. Kernel matrix is used to integrate from heterogeneous data. In this research, we proposed a simple and efficient kernel matrix evaluation which used as a kernel weighting in linear kernel matrix combination. Kernel Target Alignment (KTA) obtains the higher average AUC score compared to the other kernel weighting methods and its combinations. The combination of KTA with the other kernel measurements will decrease the AUC score. The combinations of KTA with the other kernel measurements decrease the information quality and the goodness of the integrated kernel matrix. In the next research, we attempt to examine the kernel weight in the data sample level.

References

[1] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and S. Noble, "A statistical framework for genomic data fusion", Bioinformatics, vol. 20, no. 16, 2004, pp. 2626–2635.

[2] Hamid JS, Hu P, Roslin NM, et al, "Data integration in genetics and genomics. Methods and challenges. Human Genomics and Proteomics", 2009; doi:10.4061/2009/869093.

[3] Chluenzen, F., Tocilj,A., Zarivach,R., Harms,J., Gluehmann,M., Janell,D. Bashan,A., Bartels,H., Agmon,I., Franceschi,F. And Yonath,A, "Structure of functionally activated small ribosomal subunit at 3.3 Å resolution", Cell, 102, 615–623, 2000.

[4] Harms,J., Schluenzen,F., Zarivach,R., Bashan,A., Gat,S., Agmon,I., Bartels,H., Franceschi,F. and Yonath,A, "High resolution structure of the large ribosomal subunit from a meshophilic eubacterium", Cell, 107, 679–688, 2001.

[5] Malossini.A. "Kernel methods for quality control and data integration in microarray data analysis", PhD Dissertation, 2007.

[6] Yiming Ying, Kaizhu Huang, Colin Campbell, "Enhanced Protein Fold Recognition Through a Novel Data Integration Approach", BMC Bioinformatics, Vol. 10, No. 1, 2009, 267 doi:10.1186/1471-2105-10-267.

[7] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. Kandola, On kernel-target alignment, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), "Advances in Neural Information Processing Systems", MIT Press, Cambridge, MA, 2001.

[8] Canh Hao Nguyen, Tu Bao Ho: "An efficient kernel matrix evaluation measure", Pattern Recognition 41(11)(2008)): 3366-3372.

[9] Tang KL, Yao WJ, Li TH, Li YX, Cao ZW, "Cancer classification from the gene expression profiles by Discriminant Kernel", PLS, Journal of Bioinformatics and Computational Biology, 2010 Dec;8 Suppl 1:147-60.

[10] Cristianini,N. and Shawe-Taylor,J. "An Introduction to Support Vector Machines", Cambridge University Press, Cambridge,UK, 2000.

[11] Schölkopf,B. and Smola,A. " Learning with Kernels". MIT Press, Cambridge, MA, 2002.

[12] Wahba,G. "Spline Models for Observational Data". SIAM, Philadelphia, 1990.

[13] Vapnik,V.N. "Statistical Learning Theory", Wiley-Interscience, 1998.

[14] Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. "Basic local alignment search tool", J. Mol. Biology. 215, 403–410, 1990.

[15] Smith,T.F. and Waterman,M.S. "Identification of common molecular subsequences", J. Mol. Biol., 147, 195–197, 1981.

[16] Sonnhammer,E., Eddy,S. And Durbin,R. "Pfam: a comprehensive database of protein domain families based on seed alignments". Proteins, 28, 405–420, 1997.

[17] S. Fine, K. Scheinberg, "Efficient SVM training using low-rank kernel representations", J. Mach. Learn. Res. 243--264.2, 2002.

[18] C.S. Ong, A.J. Smola, R.C. Williamson, "Learning the kernel with hyperkernels", J.Mach. Learn. Res.1043--1071.6, 2005.

Ito Wasito is a senior lecturer and researcher at Faculty of Computer Science, Universitas Indonesia. He did PhD in Computer Science at University of London. His research interests are Data Mining and Bioinformatics.

Hadaiq R. Sanabila is a graduate student at Faculty of Computer Science, Universitas Indonesia.

Aulia N. Istiqlal is a graduate student at Faculty of Computer Science, Universitas Indonesia.