ing paradigm could be viewed as an integration of supervised learning and unsupervised learning.

The problem of high dimensionality can be alleviated by discriminant analysis. It is to find a mapping such that the data are clustered in the reduced feature space, in which the probabilistic structure could be simplified and captured by simpler model assumption, e.g., Gaussian mixtures.

Discriminant-EM (DEM) [1] effectively combines self-supervised learning and discriminant analysis for content-based image retrieval. It applies self-supervised learning techniques in a lower dimensional space projected by discriminant analysis. The basic idea is to learn discriminating features and a classifier simultaneously by inserting a multi-class linear discriminating step in the standard expectation-maximization (EM) [2] iteration loop. However, since its discriminating step is linear, DEM has an obvious drawback in handling data that are not linearly separable.

In this paper, we generalize the DEM algorithm from a linear setting to a nonlinear one, and propose a self-supervised learning framework for microarray analysis using Kernel Discriminant-EM (KDEM). KDEM transforms the original data space **X** to a higher dimensional kernel *"feature space" F* and then projects the transformed data to a lower dimensional discriminating subspace, such that nonlinear discriminating features could be identified and training data could be better classified in a nonlinear feature subspace. Extensive experiments are performed on the microarray dataset of malaria parasite *Plasmodium falciparum* for classification of specific functional classes.

The rest of the paper is organized as follows. In Section 2, we present kernel DEM algorithm. In Section 3, we apply KDEM and other algorithms to gene classification and use KDEM to identify putative genes of specific functional classes. Finally conclusions and future work are given in Section 4.

2 Kernel Discriminant-EM Algorithm

Kernel DEM (KDEM) is a generalization of DEM [1] in which instead of a simple linear transformation to project the data into discriminant subspaces, the data is first projected nonlinearly into a high dimensional feature space F where the data are better linearly separated. After that, the original Multiple Discriminant Analysis (MDA) [2] algorithm is applied in a kernel feature space F, which is related to the original space by a non-linear mapping $\phi : x \rightarrow \phi(x)$. To avoid working with the mapped data explicitly, the kernel function computes a dot product in a feature space F: $k(x, z) = (\phi(x)^T \cdot \phi(z))$. Formulating MDA using only dot products of the form $\phi_i^T \cdot \phi_j$, the reproducing kernel matrix can be substituted into the formulation and eliminate the need for direct nonlinear transformation.

Using superscript $^\phi$ to denote quantities in the new space, we have the objective function of MDA in the following form:

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B^\phi W|}{|W^T S_W^\phi W|} \tag{1}$$

$$S_B^\phi = \sum_{j=1}^{C} N_j \cdot (m_j^\phi - m^\phi)(m_j^\phi - m^\phi)^T \tag{2}$$

$$S_W^\phi = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (\phi(\mathbf{x}_i^{(j)}) - m_j^\phi)(\phi(\mathbf{x}_i^{(j)}) - m_j^\phi)^T \tag{3}$$

with S_B and S_W are between-class and within-class scatter matrices, $m^\phi = \frac{1}{N}\sum_{k=1}^{N}\phi(\mathbf{x}_k)$, $m_j^\phi = \frac{1}{N_j}\sum_{k=1}^{N_j}\phi(\mathbf{x}_k)$, where $j = 1,\ldots,C$, and N is the total number of samples.

In general, there is no other way to express the solution $W_{opt} \in F$, either because F is too high or infinite dimension, or because we do not even know the actual *feature space* connected to a certain kernel. Hence the goal of kernel multiple discriminant analysis (KMDA) [3] is to find

$$\mathbf{A}_{opt} = \arg\max_{A} \frac{|\mathbf{A}^T K_B \mathbf{A}|}{|\mathbf{A}^T K_W \mathbf{A}|} \tag{4}$$

Where $A = [\vec{\alpha}_1, \cdots, \vec{\alpha}_{C-1}]$, K_B and K_W are $N \times N$ matrices which require only kernel computations on the training samples [3].

Kernel DEM can be initialized by selecting all labeled data as kernel vectors, and training a weak classifier based on only labeled samples. Then, the three steps of Kernel DEM are iterated until some convergence criterion is satisfied:

- E-step: set $\hat{Z}^{(k+1)} = E[Z \mid D; \hat{\Theta}^{(k)}]$

- D-step: set $\mathbf{A}_{opt}^{k+1} = \arg\max_{A} \frac{|\mathbf{A}^T K_B \mathbf{A}|}{|\mathbf{A}^T K_W \mathbf{A}|}$, and project a data point $\mathbf{x}$ to a linear subspace of feature space F.

- M-Step: set $\hat{\Theta}^{(k+1)} = \arg\max_{\Theta} p(\Theta \mid D; \hat{Z}^{(k+1)})$

The same notation is used as in [1]. The E-step gives probabilistic labels to unlabeled data, which are then used by the D-step to separate the data.

3 Experiments and Analysis

3.1 Dataset

The microarray dataset used in this study is a time-course expression profile of malaria parasite *Plasmodium falciparum*, during the 48-hour red blood cell cycle [4]. The original data are downloadable from http://malaria.ucsf.edu/SupplementalData.php, which include the profiles of 46 consecutive time points excluding 23-hour and 29-hour during which synchronized samples were not available. After standard quality control filtering and normalization, a

complete dataset consisted of signals for 7091 oligonucleotides corresponding to over 4000 Open Reading Frames (ORFs) [5]. Note that the spots with array features were recorded as empty and thus were not included for further study if the sum of median intensities is smaller than the local background plus two times its standard deviation.

In the original paper [4], 14 functional classes of proteins were shown to exhibit distinct developmental profiles by Fourier Transform, including components involved in genetic information flow, metabolic pathways, cellular regulatory networks, organ-ellar activities, and parasite-specific activities. Because the number of genes in class 3 and 7 were too small to train, we combined class 3 with class 7 to form a large group given that they are naturally consequential in metabolic pathways, and combine class 4 with class 5 given that they both represent nucleotide synthetic pathways. Finally, the entire ground truth dataset included the expression of 472 annotated genes in a total of 12 functional classes (Table 1).

3.2 Experiments

In a well-cited microarray classification study [6], SVM, decision tree and multi-layer perceptrons (MLP) etc. have been investigated, and SVM, especially SVM with ra-dial basis (RBF) kernel significantly outperformed the other algorithms in the func-tional classification. Therefore we were focused on the comparison of KDEM with SVM using the same radial basis kernel (RBF) functions. RBF functions used are $K(\mathbf{X}, \mathbf{Y}) = \exp(- \| \mathbf{X} - \mathbf{Y} \|^2 / 2\alpha^2)$. In our testing, α was set to be a widely used value, the median of the Euclidean distances from each positive example to the nearest negative example [6].

We performed a two-class classification with positive genes from one functional class and the negative genes from the remaining classes. Hence, each gene can be classified in one of the four ways: true positive (TP), true negative (TN), false posi-tive (FP) and false negative (FN), according to the ground truth and classifier results. The malaria dataset is an imbalanced dataset, in which the number of negative genes is much larger than the number of positive genes. For example, for class *Mitochon-drial*, the number of positive instances was only 16 whereas the number of negative instances reached 456. In this case, we chose to use *f*_measure=2·(recall×precision)/(recall+precision) to measure the overall performance of each classifier. Besides, for each class, we randomly selected 2/3 positive genes and 2/3 negative genes as training set and the remaining gene data as testing set for classification. This procedure was repeated for 30 times. Finally, we obtained the average values of *f*_measure of 30 rounds for each class.

Table 1 shows the *f*_measure for three different classifiers on the twelve functional classes. From this table, we can clearly see that: (i) KDEM outperformed SVM for total of eleven classes. SVM failed for most classes with small size and yielded zero *f*_measure. It is because given small sample size, SVM could hardly find sufficient labeled data to train classifiers well. By contrast, KDEM releases the pain of small sample size problem by incorporating a large number of unlabeled data. (ii) When the sample size was large, for example, for class 2, KDEM also performed at least as good as SVM. (iii) Compared to DEM, KDEM also achieved superior performance

on most classes except for class 5. This shows that KDEM, provided good kernel functions, has a better capacity than DEM to separate linearly non-separable data. For example, for the functional class *Proteasome*, the *f*_measure of KDEM was 87.4, whereas DEM could only achieved 28.8. For class 5, KDEM doesn't show significant advantages over DEM, which is probably due to that the data are more likely linearly separable. We will further investigate this to prove our assumption. (iv) Figure 1 shows the performance of KDEM, DEM and SVM as the size of training samples drops from 2/3 to 1/5 of the total samples. It is clear that the performance of KDEM is good and stable while the performance of DEM and SVM declines with smaller training samples.

Overall, the superior classification results of KDEM over other methods demonstrate its promise for classifying microarray gene expression data.

Table 1. Comparison of *f*_measure for KDEM, DEM, and SVM on twelve classes

Class_ID	Functional Class	Number	SVM-rbf	DEM	KDEM-rbf
1	Transcription	23	0.0	16.0	**30.0**
2	Cytoplasmic Translation	149	**87.5**	79.7	87.2
3	Glycolysis pathway and TCA cycle	23	1.33	17.6	**35.6**
4	nucleotide synthesis	21	0.0	22.0	**23.6**
5	DNA replication	36	17.6	**59.9**	58.4
6	Proteasome	35	70.9	28.8	**87.4**
7	Plastid genome	18	57.9	67.3	**81.3**
8	Merozoite invasion	80	84.1	80.7	**86.5**
9	Actin myosin motors	13	0.0	32.7	**35.3**
10	Early ring transcripts	31	91.3	90.6	**91.4**
11	Mitochondrial	16	0.0	27.3	**35.5**
12	Organellar Translation	27	0.0	26.2	**42.4**
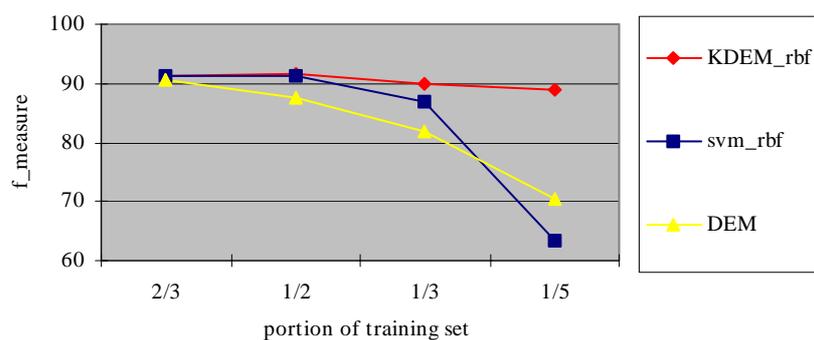
Fig. 1. Comparison of *f*_measure for KDEM, DEM, and SVM on *Mitochondrial* class with different sizes of training set.

3.3 Putative genes of specific functional classes identified by KDEM

After validation of our algorithms on a small set of genes (472) with ground truth from *P. falciparum* microarry dataset, we applied KDEM to the rest of unknown 3776 putative malaria genes to classify six specific functional categories.

Table 2 shows several representative genes that were predicted to belong to six functional classes. Their potential functionality is confirmed by independent predictions based on Gene Ontology [8], demonstrating that self-supervised learning is a powerful expression classification method. Such classification could shed light on novel network components and interactions:

(1) Transcription, translation, and DNA replication machineries are complex networks that involve fine regulations of DNA (RNA)-protein and protein-protein interactions. For instance, besides essential enzymes (DNA-directed RNA polymerase complex), transcriptional factors such as Gas41 and Sir2 homolog and transcriptional activators may participate in the regulation of transcription (Table 2). The promoter regions of these regulators are yet to be discovered.

(2) Glycolysis/TCA cycle and Nucleotide (DNA or RNA) synthesis exemplify metabolic networks which involve protein-metabolite interactions. For example, the presence of a cascade of co-expressed enzymes, including glucose-6-phosphate isomerase, glycerol-3-phosphate dehydrogenase, pyruvate kinase, lactate dehydrogenase (Table 2), not only suggests that malaria parasite possesses conserved key components in carbohydrate metabolism, but also portrays the various co-factors and metabolites that are involved in the activity of each enzyme.

(3) Proteasome is a tightly-wrapped complex of threonine proteases and regulatory proteins that mediate protein-protein interactions in cell cycle control and stress response. In previous work [9], we predicted a number threonine proteases and ubiquitin hydrolases, sketching the core elements of malarial proteasome. A concerted regulation pattern revealed by this study is consistent with the postulation of an essential ATP-dependent ubiquitin-proteasome pathway, which was inferred from the results of inhibition assays [10].

4 Discussions

Kernel Discriminant-EM (DEM) proposed a framework to address the small sample problem and the high dimensionality problem by applying self-supervised learning in an optimal nonlinear discriminant subspace. The proposed algorithm is applied on gene classification on the *Plasmodium falciparum* dataset, and KDEM outperforms linear DEM and SVM in the extensive tests.

The insights provided by self-supervised learning on transcriptomic data into the dynamics of gene networks could shed light on as yet unrecognized network interactions [11]. A significant roadblock on the use of genomic data to better understand infectious diseases is our inability to assign gene functionality. Malaria parasite *Plasmodium falciparum* appears among the most problematic: 60% of the open reading frames are annotated as "hypothetical" [5]. Our study may provide an effective

means to circumvent this problem. By identifying co-expressed genes in developmental cycle, it also helps us to identify what could conceivably be network modules. Any network module could contain a range of proteins and regulatory elements [11]. The key components of these modules may have stringent functional constraint and hence are conserved across species [12]. Subtracting these known from the modules, the remaining "hypothetical" in transcriptomic maps represent lineage-specific gaps in gene networks. The ability to assign a "hypothetical" gene to a specific network module opens an opportunity toward a tempo-specific functional characterization, because for a parasite with multiple hosts (human and mosquito) and a dynamic life cycle, "when and where" to initial wet-lab experiments is of critical importance. This network view should allow us to locate choke points in the parasite - potential vulnerabilities that could result in new malarial control strategies.

Acknowledgement

References

1. Wu, Y., Tian, Q., and Huang, T. S.: Discriminant EM algorithm with application to image retrieval, *Proc. of IEEE Conf. Computer Vision and Pattern Recognition* (2000)
2. Duda, R. O., Hart, P. E., and Stork, D. G.: 2nd *Pattern Classification*, John Wiley & Sons, Inc. (2001)
3. Schölkopf, B. and Smola, A. J.: *Learning with Kernels*. Mass: MIT Press. (2002)
4. Bozdech, Z. Llinas M., Pulliam, B. L., Wong, E. D., Zhu, J. DeRisi, J. L.: The transcriptome of the intraerythrocytic development cycle of plasmodium falciparum. *Plos Biology*, (2003) 1(1): 1-16
5. Gardner, M.J., N. Hall, E. Fung, O. White, M. Berriman, R.W. Hyman, J.M. et al.: Genome sequence of the human malaria parasite *Plasmodium falciparum. Nature,* (2002) 419: 498-511
6. Brown, M.P., Grundy, W.N., Lin, D., et al.: Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl, Acad. Sci.* USA, (2000), 97(1): 262-267
7. The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology," *Nature Genet.* (2000) Vol. 25, 25-29
8. Wu, Y., Wang, X., Liu, X., and Wang Y.: Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite, *Genome Res.* (2003) 13:601-616
9. Gantt S.M., Myung J.M., Briones M.R., Li W.D., Corey E.J., Omura S., Nussenzweig V., Sinnis, P.: Proteasome inhibitors block development of Plasmodium spp., *Antimicrob Agents Chemother.* (1998) 42:2731-2738
10. Kitano, H.: Systems biology: A brief overview. *Science,* (2002) 295: 1662-1664
11. Bowers, P.M., Cokus, S.J., Eisenberg, D., Yeates, T.O.: Use of logic relationships to decipher protein network organization, *Science.* (2004) 306:2246-2249

Table 2. Representative co-expressed genes of specific functional classes. The classification is based on their expression profiles during erythrocytic developmental cycle in malaria parasite.

Class	Oligo_ID	Gene_ID	Annotation	Prob(%)
Transcrip-tion	f22770_1	PFC0805w	DNA-directed RNA pol II	96.4
	opfc0750			77.9
	m44300_14	PF13_0152	sir2 homologue	99.6
	f21506_2	MAL8P1.131	Gas41 homologue	51.3
	M33088_1	MAL13P1.213	transcription activator	98.2
Translation	c430	PFC0635c	TIF E4	89.8
	f26262_1	PF07_0117	eukaryotic TIF2 α	55.4
	j346_4	PF10_0103	eukaryotic TIF2, β	96.3
	j353_17	PF10_0136	Initiation factor 2 subunit	76.9
	ks142_1			92.1
	popfj52810			88.9
	opfI0097	PFL2430c	eukaryotic TIF2b	63.5
	a3310_7	PFA0495c	elongation factor	69.5
	c578	PFC0870w	elongation factor 1	97.9
	f64345_2	PFL1590c	elongation factor g	85.0
	f41218_2	MAL7P1.20	peptide chain release factor	50.9
	opff72453	MAL6P1.210	nascent polypeptide associated complex alpha c	92.8
DNA replication	D17715_47	PFD0475c	replication factor a protein	99.9
	D12635_36	PFD0950w	ran binding protein 1	94.5
	F64125_2	PFE0520c	topoisomerase I	51.2
	F16271_1	PF07_0105	exonuclease I	99.9
	F16210_1	MAL7P1.145	DNA mismatch repair protein pms1 homologue	79.6
	oPFG0045			95.1
	F57777_1	MAL6P1.125	DNA polymerase epsilon	99.9
Nucleotide synthesis	m38941_10	PF13_0349	diphosphate kinase b	67.2
Glycolysis TCA	j21_14	PF10_0363	pyruvate kinase	89.7
	ks152_12	PF11_0157	glycerol-3-phosphate dehydrogenase (GPDH)	89.7
	L2_270	PFL0780w	GPDH	81.7
	Z_5_70	PF13_0141	L-lactate dehydrogenase	99.9
	Z_5_80			99.2
	Z_5_90			99.9
	m16243_2	PF13_0269	glycerol kinase	99.9
	N132_136	PF14_0341	glucose-6-phosphate isomerase	54.1
	E714_14	PFE0225w	3-methyl-2-oxobutanoate dehydrogenase	97.5
	J158_3	PF10_0218	citrate synthase	97.3
	PFBLOB0009	PF10_0334	succinate dehydrogenase	91.8
Proteasome	D6287_29	PFD0165w	ubiquitin-specific protease	74.6
	D23156_23	PFD0680c	Ubiquitin terminal hydrolase a	99.9
	Z_7_90	MAL8P1.142	proteasome β-subunit	93.3
	oPFL0014	PFL2345c	tat-binding protein homolog	99.9