ISSN 1682-296X (Print)
ISSN 1682-2978 (Online)

Bio Technology

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Biotechnology 13 (5): 226-232, 2014
ISSN 1682-296X / DOI: 10.3923/biotech.2014.226.232

Kernel-Based Informative Metric for Morphological Differences in Cell Images

Songhua Liu, Caiying Ding, Shaojun Zhong and Caiying Zhou
Institute of Medical Information Engineering, School of Science,
Jiangxi University of Science and Technology, Ganzhou, 341000, China

Abstract: The aim of this study is to propose a framework for Quantification and Calibration of Morphological Differences (QCMD) to extract nonlinear features. The framework based on our proposed Kernel informative energy which extends previous energy-based approaches in several important directions. First while previous research has defined Graph transition energy to quantify the similarity between collections of images, model both similar and dissimilar samples in a unified similarity measure framework. Further propose a Kernel informative energy to deal with potential similar samples and can be viewed as a Kernelized method (Kernel-QCMD). Finally, this study optimize the objective function and obtain global solutions. The results inferred that the proposed method can not only improve quality of quantification for morphological differences but also preserve an amount of useful information of cell images. Compared with the QCMD method, this fully extracted nonlinear feature is more suitable to quantification.

Key words: Cell images, kernel methods, distance metric, graph transition energy

INTRODUCTION

The knowledge of the morphological differences is not only useful to understand its specific function but it is also critical for drug discovery and disease diagnosis. For example (Nanni *et al*., 2010; Min *et al*., 2013) focused on the study of machine learning techniques for cell phenotype image classification and other applications. Some researchers have proposed a graph-based approach (Lin *et al*., 2010; Chen and Murphy, 2006) and quantitative method (Chen *et al*., 2013) to cell image analysis.

Despite its good performance, graph-based approaches are limited in its applicability to nonlinear problems. For example Loo *et al*. (2007) proposed a graph-based multivariate profiling method and support vector machine are applied to establish a hyperplane. However, this method fails when the images are not linearly separable. Also, the images may have a very high sampling variance.

In recent years, much effort has been devoted for morphological cell image analysis (Chen *et al*., 2012). Peng *et al*. (2011) detected and quantified target events to measure the composition of morphological subtypes of mitochondria (Peng *et al*., 2011). Sharma *et al*. (2012) describes a novel method for determining similarity between histological images through graph-theoretic description and matching for the purpose of content-based retrieval. A higher order graph-based representation of breast biopsy images has been attained and a tree-search based inexact graph matching technique has been employed that facilitates the automatic retrieval of images structurally similar to a given image from large databases (Sharma *et al*., 2012). Similar measurements used in this area is also extended by researchers (Tembe and Deodhar, 2010; Yu *et al*., 2008). Hochbaum *et al*. (2012) introduces a new framework for automatically ordering the performance of drugs. This general strategy takes advantage of graph-based formulations and solutions and avoids many shortfalls of traditionally used methods in practice (Hochbaum *et al*., 2012). Bagci *et al*. (2013) also introduces a novel computational framework to enable automated identification of texture and shape features of lesions on medical images through a graph-based image segmentation method (Bagci *et al*., 2013). Although, most of these methods are based on Euclidean geometry and consequently place limitations on the shape of the clusters, they have allowed morphological differences to be extensively used as a tool to understand biological activity within diseases. Smith *et al*. (2014) visualize morphological differences to test the relation between the cannabis and working memory impairments (Smith *et al*., 2014). Park *et al*. (2014) investigate the morphological properties of circulating tumor cell from patients in order to establish an important predictor of clinical outcome for a range of cancers (Park *et al*., 2014). Some special distance metric is also developed (Hilker *et al*., 2012).

Corresponding Author: Caiying Ding, Institute of Medical Information Engineering, School of Science,
Jiangxi University of Science and Technology, Ganzhou, 341000, China

Among them, graph-based approach, a class of methods which is based on graph transition energy computation often yields more superior experimental performance comparing to other algorithms (Huang and Murphy, 2004). While many graph algorithms are based on Euclidean geometry and consequently place limitations on the shape of the clusters. To avoid this limitation, morphological analysis method can adapt to a wider range of geometries and detect non-convex patterns and linearly non-separable clusters are appreciated.

Motivated by these observations, this study developed a framework to analyze the morphological differences based on Kernel informative energy. We employed this metric to examine both similar and dissimilar data points, following quantification and calibration of morphological differences. Experimental results show the effectiveness and efficiency of our proposed method.

MATERIALS AND METHODS

Notations: Let $\{x_i, c_i\}_{i=1}^{N}$ be a set of data points where, $x_i \in X \subset R^{N\times d}$ and $c_i \in C = \{1,2,\ldots,N_c\}$ a finite set of labels, N is the total number, d is dimensionality of the data points, N_c is the number of classes. For simplicity, this study make use of a dual notation for the data point x_{ci} in the input space, it is written with a single subscript x_i when its class is irrelevant, index $1 \le i \le N$. If the class is relevant, assume that we have J_c features for cth class, we write x_{ci}, where the class index $1 \le c \le N_c$ and the index within class $1 \le i \le j_c$.

For a given Kernel function k: $X \times X \to R$, the data points x_i, $1 \le i \le j_c$ are implicitly mapped to a feature space F with usually high dimensionality. Let $\phi(\cdot)$ denotes the mapping from X to F and $y_i = \phi(x_i) \in F$, then:

$$k_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (1)$$

We use Gaussian kernels which is defined as:

$$k(x, \sigma^2) = \exp\left(\frac{-x^T x}{2\sigma^2}\right) \quad (2)$$

where, σ is Kernel width to be chosen and the data points have been centered.

The graph transition energy in Lin *et al*. (2010) is defined as:

$$E(\sigma) = \frac{1}{N} \sum T_{ij} I(y_i \neq y_j)$$

where, $T_{ij} = w_{ij} / \Sigma_k w_{ik}$ and:

$$w_{ij} = \exp\left(-\sum_{k=1}^{d} \frac{(x_{ik} - x_{jk})^2}{\sigma^2}\right), I(.)$$

denotes the indicator function whose value is 1 if statement is true and 0 otherwise.

Similar/dissimilar energy functions: For the data point transformed in the Kernel space $y_{ci} = \phi(x_{ci})$, we consider each point in the Kernel space as a particle and pull or push other particles in this space. This means that the resultant effect of a particle is the sum of the separate effects between the same class and the different classes. Thsi study defined two informative energy functions as (Liu *et al*., 2010) similar and dissimilar one.

The similar energy is computed as follows:

$$E_c(y_{ci}) = \sum_{j=1}^{J_c} \exp\left(-\frac{(y_{cj} - y_{ci})^2}{2\sigma^2}\right) \quad (3)$$

Then the dissimilar one is computed as:

$$E_{p\neq c}(y_{ci}) = \sum_{p=1}^{N_c} \sum_{l=1}^{J_p} \exp\left(-\frac{(y_{pl} - y_{ci})^2}{2\sigma^2}\right) \quad (4)$$

A high E_c indicates that two data points in the same class are quite similar. But a low $E_{p\neq c}$ indicates that two points in the different class are far apart. We can use these two values to quantify the difference between any pairs.

Kernel-based informative energy: We continue our derivation of an objective function from the similar and dissimilar energy functions. In this study, we aim to learn some mappings to project the points in the same class to be close and those in different classes to be far apart, respectively. As mentioned above, this study give simple idea that $E_c(y_{ci})$ should be as large as possible and $E_{p\neq c}(y_{ci})$ should be as small as possible. This can ensure the separation between the different classes and the aggregation within the same class. Then, the total resultant effect of the candidate point y_{ci} can be computed as:

$$E(y_{ci}) = \alpha E_c(y_{ci}) - (1-\alpha) E_{p\neq c}(y_{ci}) \quad (5)$$

where, α is balance factor which will be determined later.

However, $y_{ci} = \phi(x_{ci})$ cannot be computed explicitly. So, we transform it to a coefficient matrix learning problem. In the Kernel space, we can project y_{ci} into a new feature space and define this space as F for simplicity. Owing to the Kernel trick, we can modify y_{ci} as following:

$$y_{ci} = \langle v, \phi(x_{ci}) \rangle = \sum_{s=1}^{C} \sum_{t=1}^{J_c} \beta_{st} K(x_{st}, x_{ci})$$

where, K is the Kernel matrix, $K(x_i, x_j) = k_{ij}$, β is the coefficient when project the original y_{ci} onto the direction v. In this form, we can compute the variables y_{pl}-y_{ci} as:

$$y_{pl} - y_{ci} = B\left(\sum_k K(x_k, x_{pl}) - \sum_l K(x_l, x_{ci})\right) \quad (6)$$

where, elements of B is constructed by β as well as the kernel matrix.

Given Eq. 6, we can rewrite $(y_i-y_j)^T (y_i-y_j)$ in Eq. 3 and 4 as:

$$\begin{aligned}&(y_i - y_j)^T (y_i - y_j)\\ &= \left(\sum_k K(x_k, x_i) - \sum_l K(x_l, x_j)\right)^T B^T B\left(\sum_k K(x_k, x_i) - \sum_l K(x_l, x_j)\right)\\ &= \left(\sum_k K(x_k, x_i) - \sum_l K(x_l, x_j)\right)^T M\left(\sum_k K(x_k, x_i) - \sum_l K(x_l, x_j)\right)\\ &= \text{Distance (M)}\end{aligned} \quad (7)$$

where $M = B^TB$, this term can be view as distance function with respect to M.

Substituting Eq. 7 into 5, we can transform the transformed data point y learning problem to the coefficient matrix M learning problem. For simplicity, we write $E_c(y_{ci})$ as E_c, $E_{p\neq c}(y_{ci})$ as $E_{p\neq c}$. Then, the similar energy can be rewritten as:

$$\xi_{pull}(M) = \sum_{c=1}^{N_c} \alpha E_c$$

and the dissimilar energy:

$$\xi_{push}(M) = \sum_{p=1}^{N_c} (\alpha - 1) E_{p\neq c}$$

which can be obtained using Eq. 3, 4.

As mentioned above, our object function can be rewritten as follows:

$$\xi(M) = \xi_{pull}(M) + \xi_{push}(M) \quad (8)$$

Maximizing the objective function (Eq. 8) using gradient ascent algorithm, we can get the final coefficient matrix M which can be used for classification and projection. We simply used a semidefinite programming in our implementation.

With Eq. 8, in order to optimize it using semidefinite programming (Mostafa, 2014; Suda and Tanaka, 2014) we reformulate it as follows:

$$\begin{aligned}&\max_M (1-\mu)\xi_{pull}(M) + \mu\xi_{push}(M) + E_0\\ &\text{s.t.}\begin{cases}\xi_{pull}(M) + E_0 \le \xi_{push}(M)\\ M \ge 0\end{cases}\end{aligned} \quad (9)$$

where, E_0 is margin which represents the change amount of energy when a neighboring particle moves from its potential location to a safe distance, μ is the parameter to adjust the potential which can be determined by cross validation.

Once the coefficient matrix M is determined, we can substitute it into Eq. 8 to quantify the difference.

Parameter determination: Here, it was discussed how to determine the parameter α in Eq. 5 for the total resultant effect of the candidate points y_{ci}. Most existing parameter selection algorithms empirically select it using cross validation. In the present case, the interaction between the candidate point and other points was considered. Hence, the effect of similar and dissimilar need to be consider:

- Consider there is only one class with 4 points where, k = 3 similar neighbors and $J_p = 0$, $J_c = 3$. Then, α is equal to 0 which means only similar energy is needed to be considered, it can be computed as:

$$\left(1 - \frac{J_c}{k+1}\right)^2$$

- If there are two class, one with 4 points, the other 5 points. As mentioned above, considering the candidate point in the first class, k = 8, $j_c = 4$ and $J_p = 0$. Then, the effect of those dissimilar points can be computed as:

$$\sum_{\substack{p=1\\p\neq c}}^{N_c}\left(\frac{J_p}{k+1}\right)^2$$

- Finally, this will be:

$$\alpha = \left[\left(1 - \frac{J_c}{k+1}\right)^2 + \sum_{\substack{p=1\\p\neq c}}^{N_c}\left(\frac{J_p}{k+1}\right)^2\right]$$

which presents the effects from the salme class and the different class k is the number of neighborhood of the candidate point y_{ci}. The first term of α means total effects of other points to y_{ci}, the second term means individual effects of other points to y_{ci}, except the features in class c

If α is determined, energy function can be tested from Eq. 5 for the difference quantification. Where, k can be determined using cross validation which may affect the difference quantification results.

Quantifying differences algorithm: This study summarize the Kernel-QCMD method as follows:

Algorithm 1: Kernel-QCMD

- Acquire data points in the Kernel space by $y_{ci} = \neq(x_{ci})$
- Define the similar energy function as in Eq. 3
- Define the dissimilar energy function as in Eq. 4
- Take y_{ci} as the particle in the Kernel space and express it by using Eq. 6 and 7
- Then Eq. 5 can be applied to quantify the difference and transform this form as a coefficient matrix learning problem as in Eq. 8
- The optimize problem can be obtained by Eq. 9 for classification or projecting applications

RESULTS AND DISCUSSION

The proposed Kernel QCMD method can be evaluated by conducting a number of morphological differences and experiments on a widely used cell image database. The following describes the details of the experiments and results.

Datasets: In this study, the 2D HeLa image data set was available from the Murphy lab (See http:// murphylab.web.cmu.edu/ data/2Dhela_ images.html). This data set contains 862 images. Each image contains a single cell with exactly one of the 10 distinct subcellular organelles tagged by a fluorescent protein. For comparison, the superset of SLF16 was used which consists of 47 features (See http://murphylab. web. cmu. edu/data/2Dhela_images_download.html).

Determining the parameter: For fair comparison, datasets of Lin *et al*. (2010) were used. Apart from their work, the parameter k was determined according to the output scores of the present proposed energy functions. All the algorithms have the same parameter k (The number of neighbors). In order to further examine the behaviors of these methods, we need to determine its value.

Figure 1 shows how the energy distribution changes as k varying from 2-16 on 392 samples. Our proposed energy values are computed with respect to the ground truth as shown in Fig. 1i where k = 391 means all other samples are considered as nearest neighbors. Other results are shown from Fig. 1a-i with different k's values. The results in Fig. 1 suggest that Kernel-QC MD depends on k (the number of nearest neighbors). The reasons lie in twofold: First, the sample from different classes was considered which may aggregate the similar samples and push away those dissimilar ones. Second, appropriate

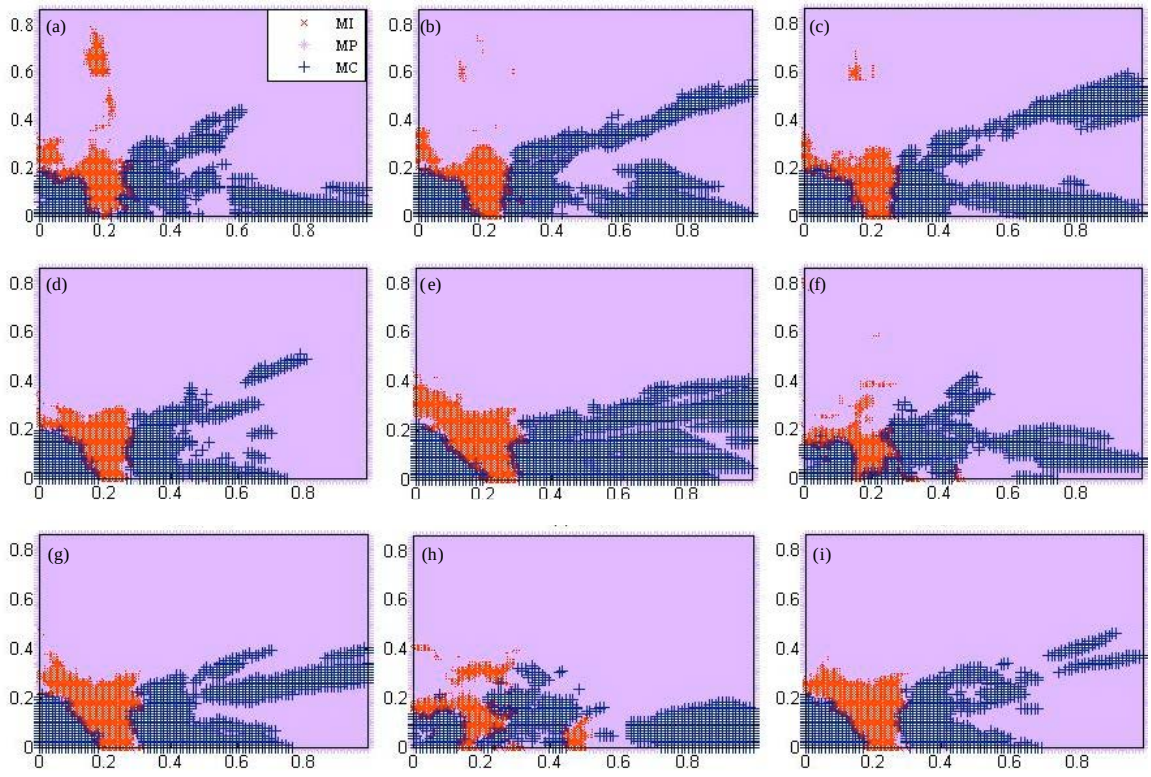

Fig. 1(a-i): Energy distribution of the Kernel-QCMD with different k (a) 2, (b) 4, (c) 6, (d) 8, (e) 10, (f) 12, (g) 14, (h) 16 and (i) ground truth

Table 1: Comparing the performance of estimating partial fragmentation of mitochondria

	Existing methods		Kernel-QCMD	
Parameters	QCMD (Lin *et al.*, 2010)	NCA (Goldberger *et al.*, 2005)	k = 6	k =14
Correlation	0.814	0.741	0.763	0.822
MSE	1.006	1.409	1.204	1.003

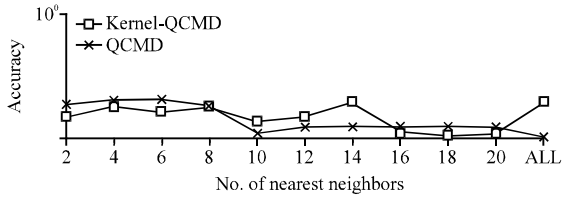

Fig. 2: Influence of the parameter k (Accuracy scale as log)

k can approximate the true distribution of the ground truth. Then, we choose k = 14 in all experimental settings.

Figure 2 shows the accuracy rate with respect to the nearest neighbors. Kernel based informative energy was used to quantify the difference in the transformed feature space by maximizing Eq. 5 with MI and MC as the training samples. Then, k is determined from 2,4,6,…, 20, ALL via leave-one-out validation based on the classification performance of the semi-supervised learning method of Zhu *et al.* (2003). For comparison, accuracy rate scale was ploted as log function, it is observed from Fig. 2 that the proposed method is sensitive to the number of nearest neighbors. Therefore, 14 nearest neighbors were used for this task.

Estimating partial fragmentation of mitochondria: Since, estimating fragmentation of a mitochondrion is important in cell image analysis, feature space transformation result were compared with NCA by Goldberger *et al.* (2005) and QCMD by Lin *et al.* (2010). In order to compare with QCMD, the same k were used for Kernel QCMD which is the number of the nearest neighbors. The result of NCA is cited from comparison results by Lin *et al.* (2010). Table 1 shows the performance comparison in terms of correlation coefficient with the human inspector's score and Mean Square Error (MSE).

Table 1 shows the performance of correlation coefficient and Mean Square Error (MSE). The result shows that our method can achieve the highest correlation coefficient and the lowest MSE when proper number of neighbors is selected. The result also shows that the transformation by learning k effectively improves the quality of the estimation. With the same k, QCMD is superior to the present proposed method which is mainly sensitive to the present method with the number of nearest neighbors. In this method, similar and dissimilar neighbors were considered, simultaneously.

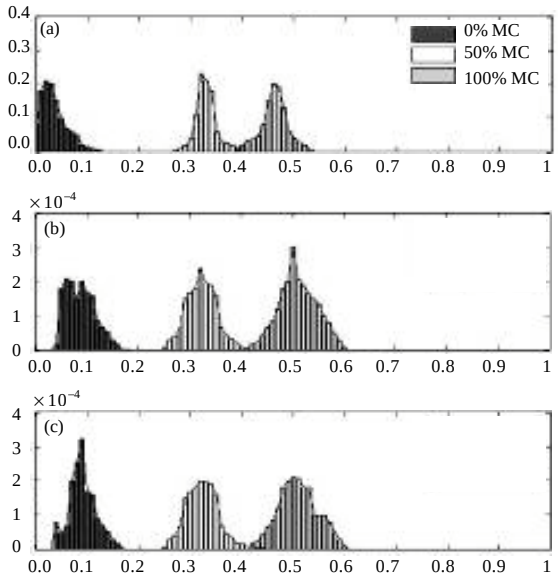

Fig. 3(a-c): Distributions of the similarity of energy as proposed by (Lin *et al.*, 2010) with (a) Optimal k (b) Distributions of energy estimated by Kernel-QCMD without optimal k and (c) with optimal k

The total effect of neighbors can be worked with proper number of nearest neighbors.

Quantifying collective morphological difference: The task to quantify morphological difference between collections of cell image was considered. We duplicate the method by the methods of Lin *et al.* (2010) and denoted as QCMD. Then the top 36 images denoted as MI and bottom 36 as MC to construct two extreme cases of mitochondria. The rest of the image are labeled as mitochondrial fragmented partially (MP) in three equal sized subsets, from the least fragmented to the most as described by Lin *et al.* (2010).

The present method was repeated 1000 times to approximate the distribution of the energy values to determine discrimination power. Figure 3 gives the distribution of the different treatment sets which 0% MC means pure MI, 50% MC means mixture of MI and MC with equal proportions and 100% MC means pure MC.

The result in Fig. 3 shows that the method is more sensitive to the energy difference which present actually

improves the quality of quantification. The reasons lie in twofold: Firstly, nonlinear feature can be extracted using Kernel-based methods effectively. Secondly, real distribution can be approximated when dissimilar samples are considered.

CONCLUSION

This study presents a new Kernel-based informative energy. It is used to solve the morphological difference analysis when the dimensionality is larger than the number of observed samples. There are several important observations which are given as follows:

- First, dissimilar samples are considered in this method. According to the property of K-nearest neighborhood, this method can effectively extract the high dimensional information from those cell images
- Second, Kernel-based informative energy is introduced which is sensitive to the small difference between cell images and identification of their morphological difference can be considered more precisely
- Finally, the performance of the experimental results confirms the proposed method is efficiently and effectively
- This study introduced and experimentally evaluated extensions of the original QCMD framework for morphological differences analysis. Kernel-based informative energy is integrated into this method which is suitable for high dimensional situation. Experimental results show that this method is sensitive to the small difference between those cell images. Moreover, energy distributions confirm the proposed method can reserve geometry information in high dimensional data

ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers and editors for their detailed and constructive comments which help them to improve the quality of this work significantly. This study was supported by the Natural Science Youth Foundation of Jiangxi Educational Committee (Grant No. GJJ13377) and the key foundation project of Jiangxi University of Science and Technology (Grant No. NSFJ2014-K19).

REFERENCES

Bagci, U., J. Yao, K. Miller-Jaster, X. Chen and D.J. Mollura, 2013. Predicting future morphological changes of lesions from radiotracer uptake in 18F-FDG-PET images. PloS One, Vol. 8. 10.1371/journal.pone.0057105

Chen, S.C. and R.F. Murphy, 2006. A graphical model approach to automated classification of protein subcellular location patterns in multi-cell images. BMC Bioinform., Vol. 7. 10.1186/1471-2105-7-90

Chen, S., M. Zhao, G. Wu, C. Yao and J. Zhang, 2012. Recent advances in morphological cell image analysis. Comput. Math. Methods Med., Vol. 2012. 10.1155/2012/101536

Chen, W.W., C.H. Chien, C.L. Wang, H.H. Wang, Y.L. Wang *et al*., 2013. Automated quantitative analysis of lipid accumulation and hydrolysis in living macrophages with label-free imaging. Anal. Bioanal. Chem., 405: 8549-8559.

Goldberger, J., G. Hinton, S. Roweis and R. Salakhutdinov, 2005. Neighbourhood Components Analysis. In: Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference, Saul, L.K., Y. Weiss and L. Bottou (Eds.). MIT Press, Cambridge, MA., pp: 513-520.

Hilker, R., C. Sickinger, C.N. Pedersen and J. Stoye, 2012. UniMoG-a unifying framework for genomic distance calculation and sorting based on DCJ. Bioinformatics, 28: 2509-2511.

Hochbaum, D.S., C.N. Hsu and Y.T. Yang, 2012. Ranking of multidimensional drug profiling data by fractional-adjusted bi-partitional scores. Bioinformatics, 28: i106-i114.

Huang, K. and R.F. Murphy, 2004. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. BMC Bioinform., Vol. 5. 10.1186/1471-2105-5-78

Lin, Y.S., C.C. Lin, Y.S. Tsai, T.C. Ku, Y.H. Huang and C.N. Hsu, 2010. A spectral graph theoretic approach to quantification and calibration of collective morphological differences in cell images. Bioinformatics, 26: i29-i37.

Liu, S., J. Zhang and K. Sun, 2010. Learning low-rank kernel matrices with column-based methods. Commun. Stat. Simul. Comput., 39: 1485-1498.

Loo, L.H., L.F. Wu and S.J. Altschuler, 2007. Image-based multivariate profiling of drug responses from single cells. Nat. Methods, 4: 445-453.

Min, W., G. Gao and J. Shi, 2013. Morphological and biochemical changes of *Microcystis aeruginosa* and anabaenaflos-aquae by low frequency ultrasonic. Biotechnology, 12: 87-92.

Mostafa, E.S.M., 2014. On the solution of a nonlinear semidefinite program arising in discrete-time feedback control design. J. Applied Math., Vol. 2014. 10.1155/2014/683797

Nanni, L., A. Lumini, Y.S. Lin, C.N. Hsu and C.C. Lin, 2010. Fusion of systems for automated cell phenotype image classification. Expert Syst. Applic., 37: 1556-1562.

Park, S., R.R. Ang, S.P. Duffy, J. Bazov, K.N. Chi, P.C. Black and H. Ma, 2014. Morphological differences between circulating tumor cells from prostate cancer patients and cultured prostate cancer cells. PloS One, Vol. 9. 10.1371/journal.pone.0085264

Peng, J.Y., C.C. Lin, Y.J. Chen, L.S. Kao and Y.C. Liu *et al.*, 2011. Automatic morphological subtyping reveals new roles of caspases in mitochondrial dynamics. PLoS Comput. Biol., Vol. 7. 10.1371/journal.pcbi.1002212

Sharma, H., A. Alekseychuk, P. Leskovsky, O. Hellwich, R.S. Anand, N. Zerbe and P. Hufnagl, 2012. Determining similarity in histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics. Diagn. Pathol., Vol. 7. 10.1186/1746-1596-134

Smith, M.J., D.J. Cobia, L. Wang, K.I. Alpert and W.J. Cronenwett *et al.*, 2014. Cannabis-related working memory deficits and associated subcortical morphological differences in healthy individuals and schizophrenia subjects. Schizophrenia Bull., 40: 287-299.

Suda, S. and H. Tanaka, 2014. A cross-intersection theorem for vector spaces based on semidefinite programming. Bull. London Math. Soc., 46: 342-348.

Tembe, R.P. and M.A. Deodhar, 2010. Clonal propagation of different cultivars of *Pelargonium graveolens* (L' Herit.) viz., reunion, bourbon and Egyptian. Biotechnology, 9: 492-498.

Yu, J., J. Amores, N. Sebe, P. Radeva and Q. Tian, 2008. Distance learning for similarity estimation. IEEE Trans. Pattern Anal. Mach. Intell., 30: 451-462.

Zhu, X., Z. Ghahramani and J. Lafferty, 2003. Semi-supervised learning using gaussian fields and harmonic functions. Proceedings of the 20th International Conference on Machine Learning, August 21-24, 2003, Washington, DC., USA., pp: 912-219.