BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 21 2005, pages 3970–3975
doi:10.1093/bioinformatics/bti653

Gene expression

Improving molecular cancer class discovery through sparse non-negative matrix factorization

Yuan Gao and George Church*
Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Received on April 7, 2005; revised on July 27, 2005; accepted on August 30, 2005
Advance Access publication September 8, 2005

ABSTRACT

Motivation: Identifying different cancer classes or subclasses with similar morphological appearances presents a challenging problem and has important implication in cancer diagnosis and treatment. Clustering based on gene-expression data has been shown to be a powerful method in cancer class discovery. Non-negative matrix factorization is one such method and was shown to be advantageous over other clustering techniques, such as hierarchical clustering or self-organizing maps. In this paper, we investigate the benefit of explicitly enforcing sparseness in the factorization process.

Results: We report an improved unsupervised method for cancer classification by the use of gene-expression profile via sparse non-negative matrix factorization. We demonstrate the improvement by direct comparison with classic non-negative matrix factorization on the three well-studied datasets. In addition, we illustrate how to identify a small subset of co-expressed genes that may be directly involved in cancer.

Contact: g1m1c1@receptor.med.harvard.edu, ygao@receptor.med.harvard.edu

Supplementary information: http://arep.med.harvard.edu/snmf/supplement.htm

1 INTRODUCTION

Accurate classification of cancer types or subtypes is of great importance for better treatment and prognosis. Traditionally, such classification is based on clinical and histopathological evidences and thus subject to a pathologist's interpretation. With the advent of microarray technology, which can simultaneously monitor the expression of all genes in the genome, it is natural to ask if molecular markers, such as gene expression patterns, can be used to diagnose and classify cancer types in a systematic and objective fashion.

Many classification methods from statistical and machine-learning area have been proposed for molecular cancer classification using gene expression data (Alon *et al*., 1999; Golub *et al*., 1999; Alizadeh *et al*., 2000; Ben-Dor *et al*., 2000; Bittner *et al*., 2000; Ross *et al*., 2000; Slonim *et al*., 2000; Pomeroy *et al*., 2002; Nguyen and Rocke, 2002; Brunet *et al*., 2004). We are interested in the clustering-based class discovery methods that do not need or have the luxury of known types as training set, which are required by supervised learning methods. Several well-known unsupervised methods, such as hierarchical clustering (HC) and self-organizing maps (SOM) are powerful approaches that have been used successfully (Eisen *et al*., 1998; Tamayo *et al*., 1999; Golub *et al*., 1999; Alizadeh *et al*., 2000; Perou *et al*., 2000).

However, such methods tend to be unstable, producing different clusters with slightly different input or different choice of initial conditions.

There are at least three characteristics of gene-expression data that present challenging problem for traditional statistical and machine-learning methods. First, gene-expression data have very high dimensionality, with tens of thousands of measured variables (genes). Second, there are only a few observations or experiments available, usually <100. Therefore, the number of variables easily exceeds the number of observations, making traditional statistical methods powerless. Third, although there is a large number of measured genes, only a handful of gene components account for most of the data variation.

It is obvious that dimension reduction to a much lower dimension (smaller than the number of observations) is appropriate. Principle component analysis (PCA) or singular value decomposition and partial least squares (PLS) are two such methods that have been applied to cancer classification with satisfactory results (Nguyen and Rocke, 2002; Bicciato *et al*., 2003; Tan *et al*., 2004). However, due to the holistic nature of PCA, the resulting components are global interpretations and lack intuitive meaning. To solve this problem, Lee and Seung (1999) demonstrated that non-negative matrix factorization (NMF) is able to learn localized features with obvious interpretation. Their work was applied elegantly to image and text analysis. Inspired by the work, Brunet *et al*. (2004) applied non-negative matrix factorization to describe all the genes in a genome in terms of a small number of metagenes and summarized the sample gene-expression patterns by that of the metagenes. The metagene expression patterns were then used to cluster the samples into distinct tumor types and subtypes. Brunet *et al*. (2004) showed that NMF appeared superior to HC and SOM in the gene expression datasets.

One problem with the basic NMF formulation is that it gives no control over the sparseness of the decomposition. Li *et al*. (2001) demonstrated that depending on the data, the basic NMF may give parts-based but holistic representation. For such datasets, NMF does not give an intuitive decomposition into parts that would correspond to the idea of 'build blocks' of the data (Hoyer, 2002, 2004). However, in many applications, such control may give better representation of or reveal localized features or latent structures in the data. Thus Li *et al*. (2001) proposed local NMF (LNMF) with additional constraints to enforce the sparseness of the decomposition. Hoyer (2002, 2004) also showed that explicitly incorporating the notion

*To whom correspondence should be addressed.

of sparseness allowed the discovery of parts-based representations that were qualitatively better than those given by classic NMF. Wang *et al*. (2004) introduced yet another variant of NMF, Fisher non-negative matrix factorization (FNMF) for learning local features by imposing Fisher constraints to the original NMF formulation. This approach seeks to maximize the between-class scatter and minimize the within-class scatter of the encoding matrix H.

Naturally, we would like to apply sparse NMF to cancer classification problem. Since the work of Brunet *et al*. (2004) demonstrated that basic NMF appeared to perform better over HC and SOM, we would like to compare the sparse NMF directly with basic NMF using the same datasets. The results, though very limited in scope, seem to indicate that explicitly enforcing sparseness of H improves class discovery over direct application of the basic NMF. In addition, we have investigated the biological significance of sparseness factors by using the well-studied leukemia dataset as a show case.

2 METHODS

In this study, we attempt to classify cancer types purely from their gene-expression data. Without prior knowledge of the features that distinguish one cancer type from the other, or cancer from normal tissue, we adopt the unsupervised approach: clustering through sparse non-negative matrix factorization (SNMF). Ideally, samples with distinct disease state will form distinct clusters.

The gene-expression data obtained from a typical microarray experiment can be represented as a $N \times M$ matrix A. Each row represents the expression level of a gene across all samples, and each column represents the expression level of all genes in one sample. All the entries in the gene-expression matrix are non-negative.

Previously, Brunet *et al*. (2004) took advantage of the non-negative property of gene-expression matrix and applied the basic NMF with divergence update rule for clustering and achieved respectable classification results.

In the following, we summarize their method briefly. Detailed description can be found in their paper (Brunet *et al*., 2004). Mathematically, their clustering method resorts to factor the gene-expression matrix A into the product of two matrices of non-negative entries, $A \sim WH$. Matrix W has size $N \times k$ and Matrix H has size $k \times M$. k is much smaller than M. The column of W defines a 'metagene', with entry w_{ij} the coefficient of gene i in metagene j. The columns of Matrix H represent the metagene expression pattern of the corresponding sample, with each entry h_{ij} represent the expression level of metagene i in sample j.

Given such a factorization, the matrix H can be used to determine the cluster membership: sample j is placed in cluster i if the h_{ij} is the largest entry in column j (Brunet *et al*., 2004).

As discussed previously, the basic NMF method has no control over the sparseness of the decomposition and therefore does not always yield a parts-based representation. A few groups have proposed ways to add sparseness constraints to NMF (Li *et al*., 2001; Hoyer, 2002, 2004; Shahnaz *et al*., 2004). In this study, we attempt to enforce sparseness by combining the goal of minimizing reconstruction error with that of sparseness. Specifically, we adopt the point-count regularization approach that enforces sparseness of H by penalizing the number of non-zero entries rather than the sum of entries ΣH_{ij} in H (Hoyer, 2002, 2004; Shahnaz *et al*., 2004).

The SNMF algorithm is described below (Fig. 1). The sparseness of H is controlled by the parameter ($\lambda > 0$). Larger λ value results in sparser matrix H at the expense of accurate reconstruction of the original matrix A.

Note that the minimization problem is convex in W and H separately but not convex in both simultaneously. The idea of fixing W and solving the optimization with respect to H then reversing the roles of the variables and iterating until convergence was originally proposed by Paatero and Tapper (1994) and subsequently described by others (Hoyer, 2002, 2004; Pauca *et al*., 2004; Shahnaz *et al*., 2004).

```
Sparse NMF Algorithm (SNMF)
1. Initialize $W$, $H$ to random positive matrices of dimension
   $N \times k$ and $k \times M$ respectively, rescale the column of W to unit norm
2. Iterate until convergence or reach maximum number
   of allowed iterations.
   a. Solve. $W_{(ia+1)} := W_{ia} \frac{(AH^T)_{ia}}{(WHH^T)_{ia}}$
   b. Rescale the column of $W$ to unit norm
   c. Solve for each $j$
      $\min_{H_j} \left\{ \frac{1}{2} \| A_j - WH_j \|^2 + \frac{1}{2} \lambda \| H_j \|^2 \right\}$
   d. if ( $H_{ij} < 0$) then $H_{ij} := 0$
```

Fig. 1. Sparse non-negative matrix factorization.

Under such optimization scheme, the resulting H matrix should contain as many zero entries as possible. Larger λ will force the H matrix becomes more and more sparse, resulting in more localized basis vectors.

2.1 Biological motivation for sparseness

Originally, Lee and Seung (1999) applied NMF to decompose facial images and derived parts-based representation of whole images. Parts correspond to localized features that are building blocks for the whole. However, Li *et al*. (2001) demonstrated that basic NMF may give parts-based but holistic representation. Therefore, explicitly enforcing sparseness is desired. Analogously, in the gene-expression study, parts correspond to 'metagenes' that represent genes tend to be coexpressed in samples (Brunet *et al*., 2004). These metagenes can overlap, indicating that a single gene can participate in many pathways or processes. The more sparse the matrix of H, the more sparse is the feature matrix W. Therefore, enforcing the sparseness of H will give rise to metagenes that comprised few dominantly co-expressed genes. In the context of cancer classification, such small subset of co-expressed genes may indicate genes that are involved in cancer and thus good local features for specific cancer types. Specifically, we first decompose the gene-expression matrix A into W and H. Entry w_{ij} is the coefficient of gene i in metagene j. Entry h_{ij} is the expression level of matagene i in sample j. Therefore, for each cluster of samples, i.e. cluster i, we investigate the gene components that have relatively large coefficient in the corresponding i-th column of W. In Section 3, we investigate the biological meanings of the sparse factors by analyzing the functions of component genes in the corresponding metagenes.

Model selection. As discussed in the Brunet *et al*. (2004) paper, for any rank k, the basic NMF algorithms group the M samples into k clusters. The choice of k automatically presents a difficult problem as it is not known a priori which k decomposes the samples into meaningful clusters. Another problem is that the NMF algorithm may not always converge to the same solution from different starting point, thanks to the stochastic nature of the method. Brunet *et al*. (2004) developed a nice model selection method based on consensus clustering (Monti *et al*., 2003; Brunet *et al*., 2004). The basic idea is that if a clustering into k classes is strong, sample assignment to clusters should not vary much from random starting points. After running with many different random initial points, a consensus matrix for class assignment can be calculated. Its entries reflect the probability that each pair of samples is clustered together. Thus the dispersion between 0 and 1

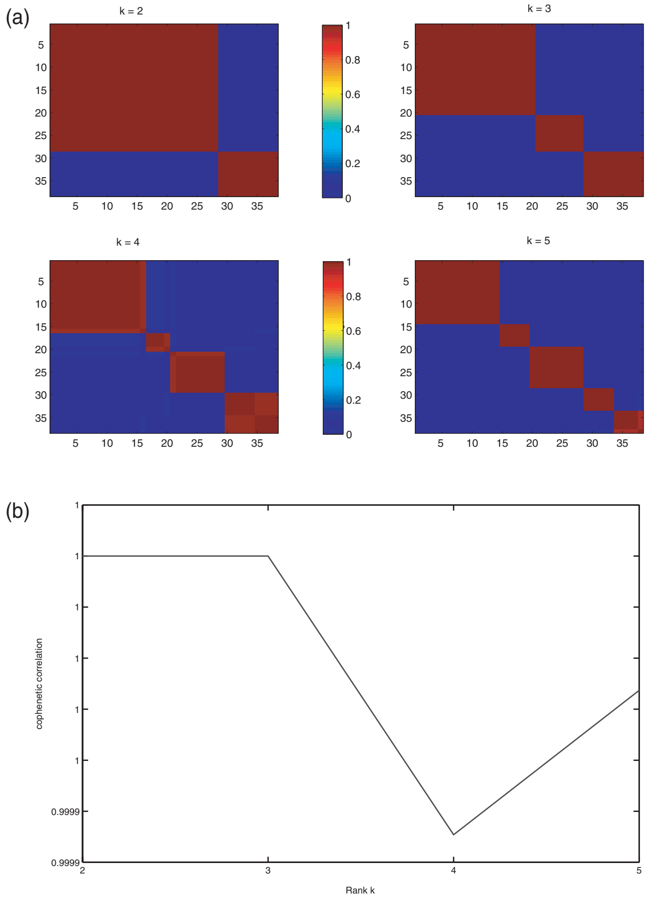

Fig. 2. The reordered consensus matrices and corresponding cophenetic correlation coefficients ρ for hierarchially clustered matrices. (**a**) The reordered consensus matrices and ρ are derived from 50 connectivity matrices calculated at $k = 2$–5, as done by Brunet *et al*. (2004). (**b**) ρ drops when k increases from 3 to 4, indicating a three-cluster split of the data is more stable than four-cluster split. Note that ρ goes back up when k increases from 4 to 5, contrasting the continuous drop observed by Brunet *et al*. Therefore, it seems that further division of the cancer subtypes may be possible.

indicates the reproducibility of the class assignments with respect to random starting points. The off-diagonal entries of the resultant consensus matrix can serve as similarity measure among samples. Brunet *et al*. (2004) use average linkage HC to reorder the samples and thus the rows and columns of it. The degree of dispersion of the reordered consensus matrix can be visually inspected. In Figures 2, 3 and 4 in Section 3, deep blue color corresponds to a numerical value of 0 and means that the samples are never assigned to the same cluster. Dark red color corresponds to 1 and means that the samples always appear in the same cluster. Quantitatively, the stability of clustering associated with a given rank k can be measured through cophenetic correlation coefficient ρ_k that ranges from 0 to 1 (Brunet *et al*., 2004). ρ_k can be easily calculated as the Pearson correlation of the distance matrix between samples induced by the consensus matrix and the distance matrix induced by the linkage used in the reordering of the consensus matrix (Brunet *et al*., 2004). Simply speaking, the bigger is the coefficient, the more stable is the cluster assignment. Therefore, by observing how ρ_k changes as k increases, one can select the values of k where the magnitude of the coefficient starts to fall. Interested readers should consult the original paper for detailed description of how this coefficient is calculated. To make the comparison valid, we use the same model selection criteria.

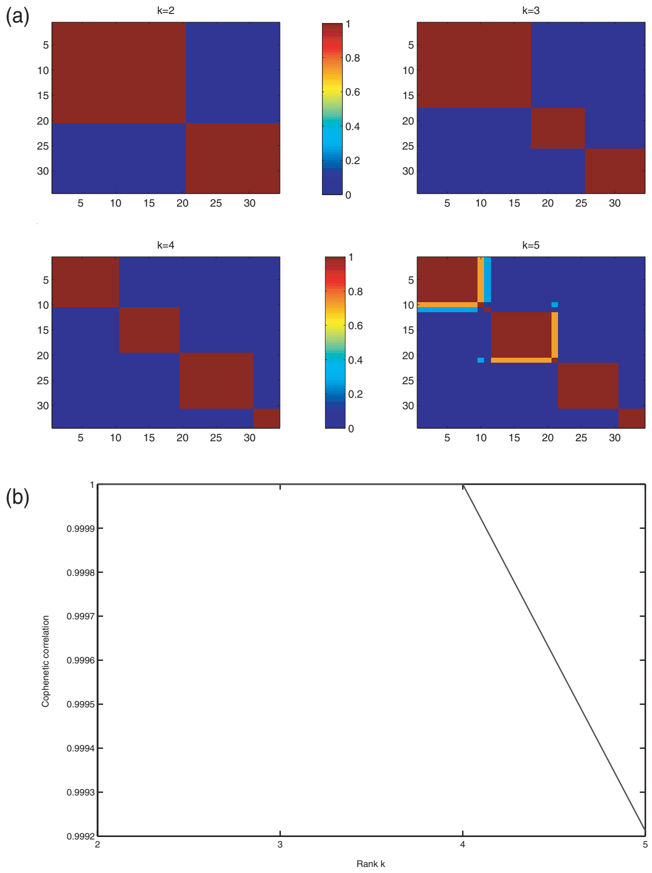

Fig. 3. (**a**) The reordered consensus matrices and corresponding cophenetic correlation coefficients for hierarchially clustered matrices. (**b**) ρ drops when k increases from 4 to 5, indicating a four-cluster split of the data is more stable than a five-cluster split.

3 RESULTS

First, we compare the clustering results achieved by enforcing the sparseness of the decomposition with the basic NMF on the three datasets reported by Brunet *et al*. (2004). The accuracy of the clustering is measured by the following formula (Xu *et al*., 2003):

$$\mathrm{AC} = \frac{\sum_{i=1}^{n} I(j_i)}{n}$$

where $I(j_i)$ is 1 if the cluster assignment is correct for sample j_i, and 0 if the cluster assignment is incorrect. The results are found in Table 1, the λ used is 0.01. Detailed analysis can be found in text and in supplement tables. It appears that SNMF outperform NMF in leukemia and central nervous system tumors dataset. However, for the Medulloblastoma dataset, the result is not clear cut and may indicate the histological subclasses are not as well understood as the other two cases. Second, we attempt to investigate the biological meaning of sparseness by analyzing genes that tend to co-occur in each cancer types. The results indicate that NMF or SNMF can identify sets of genes that seem to be involved in the underlying cancer.

3.1 Leukemia dataset

Acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL) can be easily distinguished. In addition, ALL

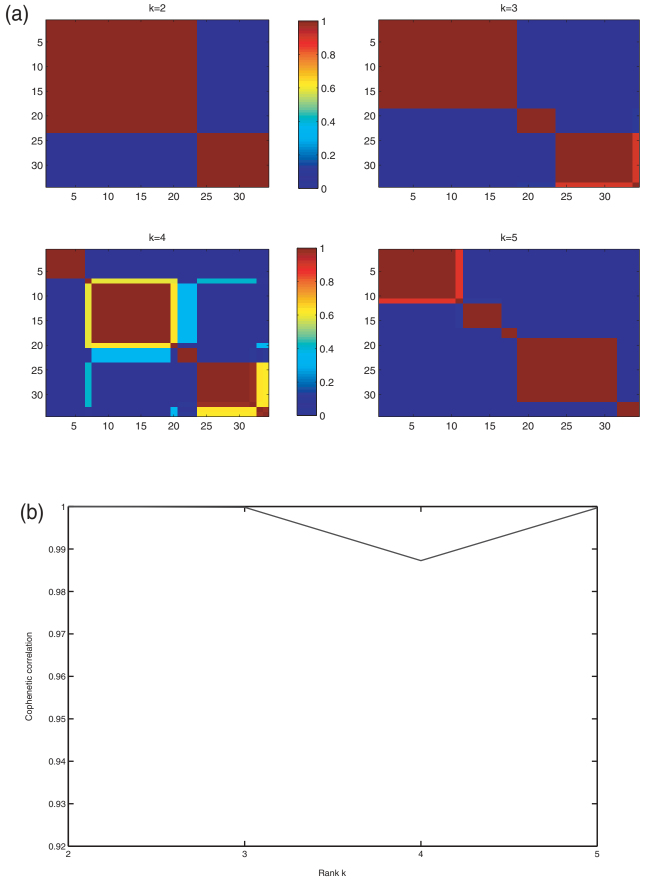

Fig. 4. (**a**) The reordered consensus matrices and corresponding cophenetic correlation coefficients for hierarchially clustered matrices. (**b**) ρ drops as k increase from 3 to 4, but goes back up at $k = 5$. When k further increases, we also observed the continuous drop of ρ, in agreement with that of Brunet *et al*. (2004) (data not shown).

Table 1. Performance comparisons of NMF versus SNMF

Datasets	Number of Types	Number of patient samples	Accuracy NMF	SNMF
Leukemia	3	38	0.947	0.974
CNS	4	34	0.941	0.971
Medulloblastoma	2	34	N/A	N/A

Because the pathogenesis of medulloblastoma is not well established, we did not calculate the accuracy for this dataset. See text for detail.

can be further divided into T and B subtypes. This dataset contains 38 bone marrow samples that can be assigned to the aforementioned three subtypes with high confidence based on clinical and histopathological evidences. This dataset is well established and has served as a benchmark dataset for comparing the performance of different clustering algorithms. In general, most clustering algorithms work well. For example, SOM could rediscover these distinctions on this dataset (Slonim *et al*., 2000). HC can also perform well depending on the choice of linage metric and the number of input genes (Brunet *et al*., 2004). However, HC was proved unstable because its performance was subject to the number of input genes. However, there are two ALL samples that are consistently misclassified or classified with low confidence by most methods (Brunet *et al*., 2004). One possible explanation mentioned by Brunet *et al*. (2004) is the incorrect diagnosis of the samples. Brunet *et al*. included them in the analysis but expected them to be outliers. When NMF was applied to the dataset, with rank $k = 2$, it consistently rediscovered the distinction between AML and ALL. However, it did misclassify two ALL B subtypes to AML (ALL_14749_B-cell and ALL_7092_B-cell to AML). At $k = 3$, it further partitioned the ALL subtypes into ALL-B and ALL-T. Again, there were two misclassification made (Supplementary Table S1). Although the same ALL_14749_B-cell was once again incorrectly assigned to AML, ALL_7092_B-cell was assigned correctly. However, a new ALL-B sample (ALL_21302_B-cell) was now incorrectly assigned to AML, indicating some kind of instability. Increasing k showed increased dispersion that was quantitatively measured nicely by decreased value of the cophenetic correlation (Brunet *et al*., 2004).

In contrast, at $k = 2$, the SNMF correctly classified the two difficult ALL cases that were missed by NMF (Supplementary Table S1). However, it did make one mistake. One AML sample (AML_12) was incorrectly assigned to ALL. At $k = 3$, SNMF nicely split the ALL samples into two subtypes with no mistake made. However, it still misclassified the same AML sample to ALL. One possible explanation may be incorrect diagnosis of this sample. The SNMF is consistent when k changes (Fig. 2), unlike NMF. This is also an improvement. Another improvement is that increasing k did not show significant dispersion, compared with NMF (Fig. 2). Therefore, we suspect that there may exist more than just three subclasses in the leukemia dataset.

3.2 Central nerve system tumors

This dataset is composed of four types of central nervous system embryonal tumors (Pomeroy *et al*., 2002). There are 34 samples representing four distinct morphologies: 10 classic medulloblastomas, 10 malignant gliomas, 10 rhabdoids and 4 normals. Both HC and SOM failed to reveal the correct clusters. The normal and malignant glioma samples was consistently clustered together by both methods (Brunet *et al*., 2004). NMF method suggested a four-cluster split with high cophenetic coefficient (Brunet *et al*., 2004). NMF method made only two mistakes. One mistake is to assign a glioma (Brain_MGlio_8) to rhabdoid. However, the other mistake is more serious. It incorrectly assigned a rhabdoid sample (Brain_Rhab_10) to normal. Such an assignment will definitely delay the treatment, if at all, of the patient and thus highly undesired.

However, the SNMF method correctly split the samples into four clusters stably with high cophenetic coefficient (Fig. 3). The drop of cophenetic coefficient at $k = 5$ indicates that the samples can be best split into four clusters. Impressively, it made only one mistake: the same glioma (Brain_MGlio_8) was assigned to rhabdoid. However, significantly, it correctly clusters the four normal samples into a distinct group.

3.3 Medulloblastoma dataset

Medulloblastoma are childhood brain tumors and the pathogenesis of these tumors is not well understood. The major histological subclass of medulloblastoma is desmoplastic medulloblastoma whose diagnosis is highly subjective (Pomeroy *et al*., 2002). Nevertheless, two known histological subclasses, classic and desmoplastic, are generally accepted as their differences can be clearly seen under microscope (Brunet *et al*., 2004). Genes whose expression was statistically correlated with those two histological classes have been reported (Pomeroy *et al*., 2002). The samples can be divided into 25 classic and 9 desmoplastic medulloblasstomas. HC and SOM failed to expose a clear desmoplastic cluster at any level of the hierarchical tree. However, basic NMF predicted robust clustering for $k = 2$, 3 and 5 and revealed a clusters made up almost entirely of desmoplastic samples. This cluster included seven out of nine desmoplastic samples. However, this cluster also contained a classic medulloblastoma sample. In contrast, the SNMF also predicted the existence of robust classes for $k = 2$, 3 and 5 (Fig. 4). For $k = 5$, we were able to expose a subclass made up entirely of desmoplastic sample. However, it contains only three out of nine desmoplastic samples, and the other desmoplastic samples were scattered among other clusters. This result is more or less consistent with the results by HC and SOM. It could mean that the straightforward use of NMF is better in this case. Or it may also raise doubt about the sample assignments, given the fact that the pathogenesis of these tumors is not well understood and desmoplastic medulloblastoma diagnosis is highly subjective.

3.3.1 Biological investigation: a case study For the three datasets, we can investigate genes that tend to co-occur in each cancer type. We illustrate our method by using the well-established leukemia dataset. To identify genes that may be dominantly involved in each subtype, top 20 genes with the largest coefficient in the W matrix are extracted for each corresponding clusters, namely ALL-B, ALL-T and AML. In general, a PubMed search for the functions of the genes indicate that the three sets of 20 genes each are enriched in chemokines, oncogenes, tumor suppressor genes and DNA repair genes. For example, in the AML cluster, the 20 co-occurred genes include GRO3 oncogene, which is a chemokine and belongs to the small inducible cytokine subfamily b. Chemokines play fundamental roles in the development, homeostasis and function of the immune system. They also affect cells of the central nervous system as well as on endothelial cells involved in angiogenesis or angiostasis. We also identified cellular oncogene c-fos, which has an important role in signal transduction, cell proliferation and differentiation. For the ALL cluster, 1 gene that does not appear in the top 20 of the AML cluster is PI5 Protease inhibitor 5 (maspin), which is a tumor suppressor and angiogenesis inhibitor. For the ALL-T subtype, the well-known tumor suppressor gene, retinoblastoma susceptibility protein (RB1) gene, with a 3 bp deletion in exon 22, is found to co-occur with SKI V-ski avian sarcoma viral oncogene homolog, Proto-Oncogene Trk and PI5 Protease inhibitor 5. For the ALL-B subtype, the tumor suppressor and angiogenesis inhibit PI5 Protease inhibitor 5 is found to co-occur with XP-C repair complementing protein (p58/HHR23B), a gene that is involved in DNA repair.

Interestingly, a single gene, AP-3 complex beta3A subunit mRNA, appeared to be the only gene that appeared simultaneously in all three subtypes, indicating shared pathways or processes. (For a compete list of the 20 genes for each subtype, refer to Supplementary Table (4). Similarly, we can also apply the same analysis to the other two datasets (data not shown).

CONCLUSIONS

NMF has been used successfully in image analysis, text clustering and cancer class discovery and classification. In this paper, we observed improved clustering results by enforcing an additional sparseness constraint on H to the basic NMF in cancer class discovery. We have shown that SNMF improves cancer class discovery on the same three datasets that were used by Brunet *et al*. (2004). Systematic studies on larger datasets are required to yield more convincing arguments for imposing the sparseness constraints on cancer class discovery. We also investigated the biological significance of enforcing the sparseness factor by using leukemia dataset as a show case.

This study is apparently very limited in scope, more detailed investigation of the theoretical and biological basis is desired in the longer term.

ACKNOWLEDGEMENTS

We would like to thank John Aach, Patrik D'haeseleer, Dat Nguyen and Xiaoxia Lin for their discussion and comments. This study was conducted while Y.G. and G.C. were supported by Department of Energy – Genomes to Life Program, Defense Advanced Research Project's Agency, and Pharmaceuticals Research & Manufacturers of America Foundation.

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U. *et al*. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci., USA*, **96**, 6745–6750.

Ben-Dor,A. *et al*. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol*., **7**, 559–583.

Bicciato,S. *et al*. (2003) PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics*, **19**, 571–578.

Bittner,M. *et al*. (2000) Molecular Classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Brunet,J-P., Tamayo,P., Golun,T.R., and Mesirov,J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA*, **101**(12): 4164–4169.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, **286**, 531–537.

Hoyer,P.O. (2002) Non-negative sparsecoding Neural Networks for Signal Processing XII. In: *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland, pp. 557–565.

Hoyer,P.O. (2004) Non-negative Matrix Factorization with sparseness constraints. *J. Mach. Learn. Res*., **5**, 1457–1469.

Lee,D.D. and Seung,H.S. (1999) Learning the Parts of Objects by Non-negative Matrix Factorization, Nature, **401**, 788–791.

Li,S.Z., Hou,X.W., Zhang,H.J. and Cheng,Q.S. (2001) Learning spatially localized parts-based representation. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Volume I*, Hawaii, December 2001, pp. 207–212.

Monti,S. *et al*. (2003) Consensus Clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn*., **52**, 91–118.

Nguyen,D.V. and Rocke,D.M. (2002) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.

Paatero,P. and Tapper,U. (1994) Positive matrix factorization: A non-negative factor model with optimal utilization of errorest mates of data values. *Environmetrics*, **5**, 111–126.

Pauca,P., Shahnaz,F., Berry,M. and Plemmons,R. (2004) Text Mining using Non-Negative Matrix Factorizations. In: *Proceedings of the Fourth SIAM International Conference on Data Mining*. Lake Buena Vista, Florida, USA April 22–24, 2004.

Perou,C.M. *et al*. (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Pomeroy,S.L., Tamayo,P., Gaasenbeek,M., Sturla, L.M., Angelo,M., McLaughlin,M.E., Kim,J.Y., Goumnerova,L.C., Black,P.M., Lau,C., Allen,J.C., Zagzag,D., Olson,J.M., Curran,T., Wetmore,C., Biegel,J.A., Poggio,T., Mukherjee,S., Rifkin,R., Califano,A., Stolovitzky,G., Louis,D.N., Mesirov,J.P., Lander,E.S. and Golub,T.R. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Ross,D.T. *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.

Shahnaz,F., Berry,M., Pauca,P. and Plemmons,R. (2004) Document Clustering using Nonnegative Matrix Factorization. Journal on Information Processing & Management, In Press.

Slonim,D.K., Tamayo,P., Mesirov,J.P., Golub,T.R. and Lander,E.S. (2000) In: *Proceedings of the Fourth International Conference on Computational Molecular Biology*, Tokyo, Japan, RECOMB 2000, pp. 263–272.

Tamayo,P. *et al*. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tan,Y.X. *et al*. (2004) Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput. Biol. Chem.*, **28**, 235–244.

Wang,Y., Jia,Y., Hu,C. and Turk,M. (2004) Fisher non-negative matrix factorization for learning local features. In: *Asian Conference on Computer Vision*, Jeju, Korea January 27–30, pp. 806–811.

Xu,W., Liu,X. and Gong,Y. (2003) Document-clustering based on non-negative matrix factorization. In: *Proceedings of SIGIR'03*, July 28-August 1, Toronto, CA, pp. 267–273.