34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Identification of Genes for Complex Diseases by Integrating Multiple Types of Genomic Data

Hongbao Cao, *Member, IEEE,* Shufeng Lei, Hong-Wen Deng, and Yu-Ping Wang, *Senior Member, IEEE*

***Abstract*—Combining multi-types of genomic data for integrative analyses can take advantage of complementary information and thus can have higher power to identify genes/variables that would otherwise be impossible with individual data analysis. Here we proposed a sparse representation based clustering (SRC) method for integrative data analyses, and applied the SRC method to the integrative analysis of 376821 SNPs in 200 subjects (100 cases and 100 controls) and expression data for 22283 genes in 80 subjects (40 cases and 40 controls) to identify significant genes for osteoporosis (OP). Comparing our results with previous studies, we identified some genes known related to OP risk, as well as some uncovered novel osteoporosis susceptible genes ('DICER1', 'PTMA', etc.) that may function importantly in osteoporosis etiology. In addition, the SRC method identified genes can lead to higher accuracy for the identification of osteoporosis subjects when compared with the traditional T-test and Fisher-exact test, which further validate the proposed SRC approach for integrative analysis.**

I. INTRODUCTION

During the past few years, various clustering techniques have been developed to identify subsets of genes significant for diagnosis or classification[1]-[7]. For example, Soneson et al. used Canonical Correlation Analysis (CCA) for joint analysis of gene expression and copy number variations (CNVs) [2]. Berger et al. developed a generalized singular value decomposition (GSVD) to locate genes with both high variations across genes and high correlation across samples between gene expression changes and CNVs [4]. These methods demonstrated limited success[2][4]. Due to different nature, structure and format of diverse sets of genomic data, multiple genomic data integration is challenging[2]-[4]. In this work, we employed the sparse representation based clustering (SRC) method for gene selection using joint analysis of two different types of genomic data: gene expression data and SNP data, based on multiple characteristics extracted from genomic data. Sparse representation or compressive sensing (CS) is a novel statistical method recently developed in statistics and applied mathematics, which has found many successful applications in many disciplines. For example, Wright *et al.* proposed a CS based method for face recognition, which showed better accuracy and resistance to noise [8]. We have developed and applied the SRC method for chromosome image classification and showed improved accuracy [9].

To validate our method, we apply it to the study of osteoporosis, which is a major public health problem over the world [10][11]. However, specific genetic factors contributing to development of osteoporosis are largely uncharacterized.

The paper is organized as follows. We first briefly describe the two data sets we tested on (SNP data and gene expression data) and the SRC model we proposed. Then we applied the method to gene selection through integrative analysis of both data sets. For the purpose of comparison with individual data analysis, we also performed the study on each data type. To demonstrate the advantage of integrative approach, we compared the selected genes using the SRC method with previously reported osteoporosis susceptive genes [5][12]. To further validate the selected genes, we applied the method to the classification of osteoporosis patients with the selected gene expression and/or SNP data. Results showed that SRC method is able to better locate genes significant for the diagnosis of osteoporosis patients than that from single data sets. In addition, our proposed SRC method gives better diagnosis results when compared with T-test and Fisher-exact test. In particular, we identified two new potential osteoporosis related genes (e.g., *'DICER1', 'PTMA'*) through joint data analysis. Those genes cannot be located with single data set but show significant roles in osteoporosis etiology from studies published, which suggests that integrated data analysis can provide new insights into the identification of disease susceptive genomic markers.

II. MATERIAL AND METHODS

In this section, we first describe the data used in our study (*Section A*). Then we present the SRC model (*Section B*), the feature selection method (*Section C*) and the SRC based gene/variable shaving algorithm (*Section D*).

*Research supported by NIH Foundation.

H. Cao. is with the Department of Biomedical Engineering, Tulane University, New Orleans, USA (e-mail: hcao3@tulane.edu)

S. Lei is with Laboratory of Molecular and Statistical Genetics, Hunan Normal University, Changsha, P. R. China(e-mail: leishufeng@yahoo.com)

H. Deng is with Department of Biostatistics, Tulane University, New Orleans, USA (e-mail: hdeng2@tulane.edu).

Y. Wang is with the Department of Electrical Engineering and Department of Biostatistics, Tulane University, New Orleans, USA (e-mail: wyp@tulane.edu, Tel: 504-988-1341).

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

A. Data

We applied the SRC method to an integrative analysis of two data sets (i.e., gene expression data set and a SNP data set) from osteoporosis study. We describe the data sets as follows.

The gene expression data was from female osteoporosis subjects with extremely low (N=40) (cases) vs. high (N=40) (controls) bone mineral density (BMD)[13]. In the present study we selected circulating monocytes as our target cells because circulating monocytes serve as progenitors of osteoclasts [14][15], and secrete osteoclastogenic cytokines, such as IL-1, IL-6, and TNF-α [16][17].

The SNP data set was from osteoporosis vs. health subjects, which were recruited with the purpose of identifying genetic factors underlying osteoporosis via genome-wide association study in a total of 1000 random female subjects (age: 50.3+18.3 years) [13]. We selected the bottom 100 and top 100 subjects of the BMD phenotypic distribution as cases and controls, respectively. A total of 376821 eligible SNPs were used in final analysis. In addition, we randomly selected 70 cases and 70 controls as training data for gene selection, and the rest 30 cases and 30 controls were used as an independent testing data set.

To perform joint data analysis, we generate a combined data set from the two single data sets, as shown in Fig. 1. For each gene, the feature vector contains two sub-vectors corresponding to gene expression and SNP data, which will be used as the input to our SRC method for joint data analysis.

B. SRC Model

Figure 2 shows the diagram of the proposed SRC model. $\boldsymbol{Y} = \{\boldsymbol{y}_i\} \in R^{m\times p}$ consists of the feature vectors extracted from the 'Data', where $\boldsymbol{y}_i \in R^{m\times 1}$ is the feature vector extracted for the i th gene/variable; m is the number of features; $i = 1,2,\dots,p$, and p is the number of gene/variables. Each column of $\boldsymbol{Y}$ is normalized to be within the range of [0,1]. $\boldsymbol{A} = [\boldsymbol{A}_1,\dots,\boldsymbol{A}_c] \in R^{m\times n}$ is the characteristic matrix that we will design to separate the data into c groups. In each group $\boldsymbol{A}_i \in R^{m\times n_i}$ contains n_i samples, and $n = \sum n_i$. The 'SRC clustering' is to cluster each y_i according to the characteristic matrix $\boldsymbol{A}$. The design of characteristic matrix $\boldsymbol{A}$ and detailed description of 'SRC clustering' algorithm can be found in [18].

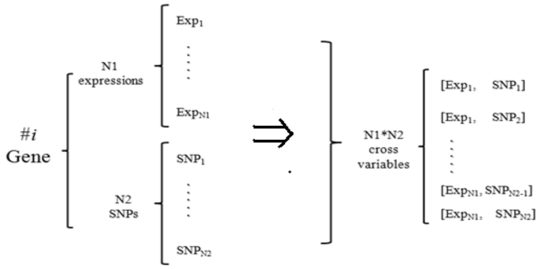


Figure1 An illustration of the combination of two different types of data for the ith gene withN1 expression vectors and N2 SNP vectors.

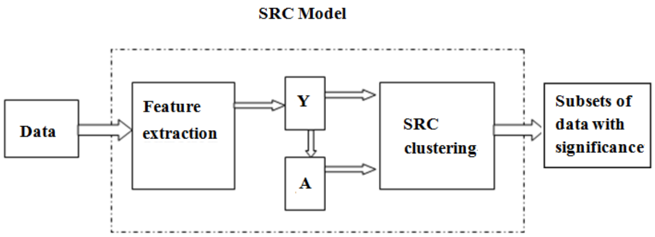


Figure 2 Diagram of SRC model for the data analysis using multi-features

C. Features selection

In this work, we extract five features for each gene/variable (gene expressions or SNPs):

$$[std_0, std_1, |\bar{X}_0 - \bar{X}_1|, |corr|, 1 - \|a\|_2]^T \qquad (1)$$

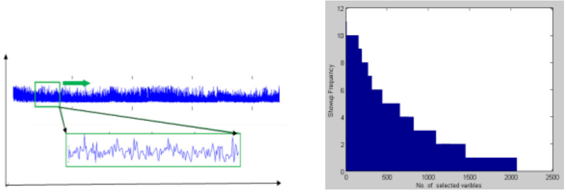
where $\bar{X}_0$ and $\bar{X}_1$, std_0 and std_1 are the means and standard deviation of control and case group respectively; $corr$ is the Pearson correlation coefficient between each gene expression (SNP) data and the healthy status ('1' for patients, and '0' for controls); and a is the normalized amplitude of vector $[std_0, std_1, |\bar{X}_0 - \bar{X}_1|, |corr|]^T$.

D. The SRC Based Gene/Variable Shaving

When the data set is very large, which is always the case for genomic data, a window is applied and the gene selection is performed within the window (Figure 3 (a)) to account for local variations in the data, with a Fisher-Yates Shuffling algorithm [19] to reduce bias. Those selected with highest frequencies will be the ones that are most significant (3 (b)). Please see ***Algorithm 1*** for details.

Algorithm 1: SRC based gene shaving algorithm.

1. Set the window length, window sliding step length, and starting point;
2. For the *l*-th iteration, perform gene selection within a window and record the selected genes;
3. Move the window from the starting point with the pre-set step length, and repeat step 2 until the window reaches the end of the data.
4. Shuffle the data with Fisher-Yates Shuffling algorithm; repeat step 2 – step 3.
5. Compare the gene list generated by all *l* iterations with that generated by previous *l*-1 iterations; if the gene list change is smaller than a pre-set threshold, exit; otherwise, go to step 4.



(a) (b)

Figure 3 The SRC based gene shaving with a sliding window (a) Gene selection was performed within each sliding window; (b) Bar plot of the selected genes with different frequencies.

III. Results

One goal of this work was to study whether integrative analysis approaches with our proposed SRC algorithm can lead to better identification of susceptible genes and diagnosis of complex diseases such as osteoporosis. To validate the selected genes, we compared our selected gene lists with those previously reported. In addition, we tested if the selected genes can result in better diagnosis of osteoporosis.

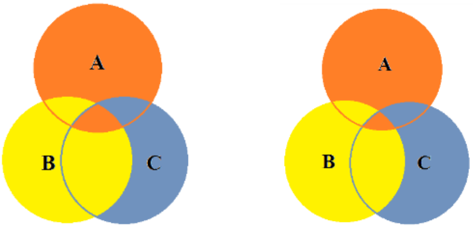
A. *Comparison of selected genes*

To show the differences between integrative analysis using the SRC and individual analysis with both SRC and traditional feature selection methods (e.g., T-test and Fisher-exact test), we compared the first 500 gene expressions and 1000 SNPs selected by different methods using the Venn diagram, as shown in Figure 4. The intersection between individual analysis using the SRC method and T-test for the gene expression selection is about 45%; the intersection is about 39% for the selected SNPs using individual analysis for the SRC method and Fisher-exact test; and the intersection between combined analysis using SRC method and separated analysis is below 10%.

When compared to the previous osteoporosis study, the SRC based variable selection method located osteoporosis susceptive genes that reported before [14][19] such as *'ESRRA', 'CALM1', 'CALM1', 'SPARC', 'LRP1', 'THSD4', 'CRHR1', 'HSD11B1', 'THSD7A', 'BMPR1B', 'ADCY10', 'PRL', 'CA8'*, *et. al.*. Specifically, there are some significant genes that were not identified by individual data analysis, such as 'DICER1', 'PTMA' *etc*.. Evidence shows that those genes may be associated with the osteoporosis disease [21]-[26].

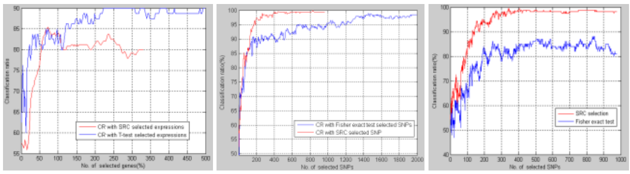
B. *Further validation of the selected variables*

We used the SRC classifier proposed by us [9] followed by the leave one out (LOO) cross validation to further validated the selected. Classification ratio (CR), defined as the number of correctly classified samples over total number of samples, was used for the test accuracy. When using selected gene expression data alone to identify the osteoporosis patients, we got the highest (86.25%) with 73 expression data; while for t-test method, we got the highest CR of 90% with 225 gene expressions (Figure 5 (a)). For the SNP data set, we got the highest CR (100%) with 883 SNPs using the SRC, while the highest CR 96.5% with 1460 SNPs using Fisher-exact test (Figure 5 (b)). When testing on the independent SNP data set, the classification ratio reached as high as 98.33% with the SRC method, while the highest CR was only 88.33% with Fisher-exact test (Figure 5 (c)).



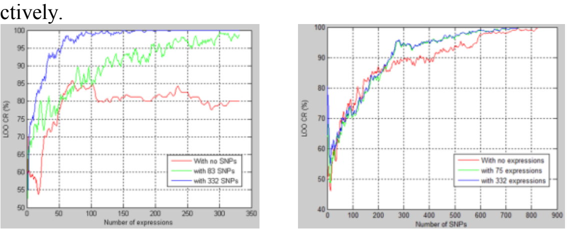
(a) (b)

Fig. 4 Comparison of selected variables (expressions/SNPs) with integrative analysis to those with individual analysis using the Venn Diagram: (a) comparison of the first 500 gene expressions selected with integrative analysis by the SRC method (A), individual analysis using SRC method (B) and using T-test (C) respectively; (b) comparison of the first 1000 SNPs selected with integrative analysis with the SRC method (A), individual analysis with the SRC method (B) and Fisher- exact test (C) respectively.



(a) (b) (c)

Figure 5 Comparison of classification results with different variable selection methods. (a) LOO validation results for the expression data by SRC method and t-test method respectively. (b) LOO validation results for the SNP data by SRC method and Fisher-exact test method respectively. (c) Testing on the independent SNP data set, with SRC method and Fisher exact test respectively.



(a) (b)

Figure 6 Using selected variables from both data sets for the classification of osteoporosis patients. (a) Classification accuracy using gene expressions along with N=0, 83, 332 selected SNPs employed for the classification. (b) Classification accuracy using SNPs along with N=0, 75, 332 selected gene expressions employed for the classification.

In addition, we compare the classification accuracy using both sub-vectors with that using one sub-vector as shown in Figure 6, which demonstrates higher identification accuracy using complementary information from both data sets.

IV. Discussion and Conclusion

In this work, we proposed an SRC based gene/variable selection method for the integrative analysis of multi-type genomic data and applied it to the identification of genes associated with osteoporosis diseases. The SRC method demonstrates two advantages: 1. Different from other analysis methods, the SRC method employs multi-features extracted from diverse data sets rather than the original raw data, facilitating the integration of data with different formats and structures. 2. The SRC method outperforms several currently used significance test methods such as the T-test and Fisher-exact test, by employing a more sophisticated clustering approach.

When compared with previously reported osteoporosis susceptible genes, the SRC based gene shaving method not only identified genes that were previously reported [12], but also new susceptive genes (*'DICER1', 'PTMA' et. al.*). Evidences [21]-[26] have shown that those genes play a significant role in the etiology of osteoporosis. In particular, it should be noticed that those genes cannot be identified with the analysis of single data sets, which indicates the advantage of integrative analysis of multiple data sets

When we compared the selected gene list with that selected by t-test and Fisher-exact test (Figure 4(a) and (b)), it can be seen that the variables (SNPs/expressions) selected by the SRC method are quite different (>50% in the number). However, the integrative analysis with the SRC method selects two sub-vectors simultaneously, resulting in better accuracy even with one set of the data for classification(Figure 5 and Figure 6), because of the use of complementary information in the SRC method. For example, using the SNP data, the SRC based method can give the highest CR of 100% vs. 96.5% of using Fisher exact test with less number of SNPs (see Figure 5 (b)). When using both types of data for the cross validation, the CR of using combined data sets with the SRC method is much higher than that of using single type of data (Figure 5 (c)), demonstrating the significance of integrative data analysis with the SRC method. In addition, when performing blind test on an independent SNP data set (30 cases 30 controls), the CR can be as high as 98.33% with the SRC method; while with Fisher-exact test selected SNPs, the highest classification ratio is only 88.33%, showing the advantage of the SRC method.

In the integrative analysis, the gene expression and SNP data were combined in terms of each gene. Therefore, the integrative analysis uses joint information from two complementary data rather than from a single type of data, which can lead to the increase of reliability in gene identification. Besides the significance discussed above, the integrative analysis employed in this work can be generalized to include more types of data. We are currently testing the method for the integration of multiple genomic data from TCGA database for improved diagnosis of cancers such as the leukemia.

REFERENCES

[1] Yang, H.H., Liu, J.Y., Sui, J., Pearlson, G. and Calhoun, V.D., (2010). A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia, Neurosci., doi: 10.3389/fnhum.2010.00192

[2] Soneson, C., Lilljebjörn, H., Fioretos, T., Fontes,M., (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis, BMC Bioinformatics 11, 191

[3] Cao, K.A. L., Martin, P.G., Robert-Granie, C., Besse, P., (2009). Sparse canonical methods for biological data integration: application to a cross platform study. BMC Bioinformatics, 10, 34.

[4] Berger, J. A., Hautaniemi, S., Mitra,S. K. and Astola, J., (2006). Jointly Analyzing Genes Expression and Copy Number Data in Breast Cancer using Data Reduction model. IEEE T. Comput. B.I.3(1), 2-16.

[5] Liu, Y.J., Shen, H., Xiao, P., Xiong, D.H., Li, L.H., Recker, R.R., Deng, H.W., (2005) Molecular Genetic Studies of Gene Identification for Osteoporosis: A 2004 Update, Journal of Bone and Mineral Research, 21(10), 1551-1535.

[6] Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R., (2005). A Method for Calling Gains and Losses in Array CGH Data, Biostatistics6, 45-58.

[7] Hautaniemi,S.,Ringner,M.,Kauraniemi,P.,Autio,R.,Edgren,H.,Yli-Harja,O.,Astola,J.,Kallioniemi,A., and Kallioniemi, O.P., (2004) A Strategy for Identifying Putative Causes of Gene Expression Variation in Human Cancers. J. Franklin Inst. 341, 77-88.

[8] Loo, L.W.M., Grove, D.I., Williams, E.M., Neal, C.L., Cousens, L.A., Schubert, E.L., Holcomb, I.N., Massa, H.F., Glogovac, J., Li,C.I., et al., (2004). Array Comparative Genomic Hybridization Analysis of Genomic Alterations in Breast Cancer Subtypes. Cancer Research64, 8541-8549.

[9] Cao, H., Wang, Y.P., M-Fish Image Analysis with Improved Adaptive Fuzzy C-Means Clustering Based Segmentation and Sparse Representation Classification", in Proc. BICoB, 2011, pp.167-171.

[10] Melton, L.J., Chrischilles, E.A., Cooper, C., Lane, A.W., Riggs, B.L., (2005) How many women have osteoporosis? JBMR Anniversary Classic. J. Bone Miner Res. 20, 886-892.

[11] Cummings, S.R., Nevitt, M.C., Browner, W.S., Stone, K., Fox, K.M., Ensrud, K.E., Cauley, J., Black, D., and Vogt, T.M., (1995) Risk factors for hip fracture in white women. Study of Osteoporotic Fractures Research Group. N. Engl. J. Med. 332, 767-773.

[12] Xu, X.H., Dong, S.S, Guo, Y., Yang, T.L., Lei, S.F., Papasian, C.J., Zhao, M., and Deng, H.W., (2010). Molecular Genetic Studies of Gene Identification for Osteoporosis: The 2009 Update, Endocr Rev.31, 447–505.

[13] Xiong, D.H., Liu, X.G., Guo, Y.F., Tan, L.J., Wang, L., Sha, B.Y., Tang, Z.H., Pan, F., Yang, T.L., Chen, X.D., et al., (2009) Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups. Am. J. Hum. Genet. 84(3), 388-398.

[14] Udagawa, N., Takahashi, N., Akatsu, T., Tanaka, H., Sasaki, T., Nishihara, T., Koga, T., Martin, T.J. and Suda, T., (1990). Origin of osteoclasts: mature monocytes and macrophages are capable of differentiating into osteoclasts under a suitable microenvironment prepared by bone marrow-derived stromal cells. Proc Natl. Acad. Sci. U S A87(18), 7260–7264.

[15] Fujikawa, Y., Quinn, J.M., Sabokbar, A., McGee, J.O., Athanasou, N.A., (1996). The human osteoclast precursor circulates in the monocyte fraction. Endocrinology137(9), 4058–4060.

[16] Cohen-Solal, M.E., Graulet, A.M., Denne, M.A., Gueris, J., Baylink, D., de Vernejoul, M.C., (1993). Peripheral monocyte culture supernatants of menopausal women can induce bone resorption: involvement of cytokines. J. Clin. Endocrinol. Metab.77(6), 1648–1653.

[17] Pacifici, R., (1996). Estrogen, cytokines, and pathogenesis of postmenopausal osteoporosis. J. Bone Miner Res. 11(8), 1043–1051.

[18] Hongbao Cao and Y. Wang, Integrated Analysis of Gene Expression and Copy Number Data using Sparse Representation Based Clustering Model, in Proc. BICoB, 2011, pp.172-177.

[19] Fisher, R.A., Yates, F., (1948). Statistical tables for biological, agricultural and medical research (3rd ed.). (OCLC 14222135London: Oliver & Boyd), pp. 26–27.

[20] Nagaraja, A.K., Andreu-Vieyra, C., Franco, H.L., Ma, L., Chen, R., Han, D.Y., Zhu, H.F., Agno, J.E., Gunaratne, P.H., DeMayo, F.J. and matzuk, M.M., (2008). Deletion of Dicer in somatic cells of the female reproductive tract causes sterility. Mol. Endocrinol 22(10), 2336-2352.

[21] Mizoguchi, F., Izu, Y., Hayata, T., Hemmi, H., Nakashima, K., Nakamura, T., Kato, S., Miyasaka, N., Ezura, Y., Noda, M., (2010). Osteoclast-specific Dicer gene deficiency suppresses osteoclastic bone resorption. J. Cell Biochem. 109(5), 866-875.

[22] Sugatani, T., and Hruska, K.A., (2009). Impaired micro-RNA pathways diminish osteoclast differentiation and function. J. Biol. Chem. 284(7), 4667-4678.

[23] Sugatani, T., Vacher, J. and Hruska, K.A., (2011) A microRNA expression signature of osteoclastogenesis. Blood117(13), 3648-3657.

[24] Wang, X., Kua, H.Y., Hu, Y., Guo, K., Zeng, Q., Wu, Q., Ng, H.H., Karsenty, G., de Crombrugghe, B., Yeh, J., Li, B., (2006).p53 functions as a negative regulator of osteoblastogenesis, osteoblast-dependent osteoclastogenesis, and bone remodeling. J. Cell. Biol. 172(1), 115-125.

[25] Dumble, M., Gatza, C., Tyner, S., Venkatachalam, S., Donehower, L.A., (2004). Insights into aging obtained from p53 mutant mouse models. Ann. N. Y. Acad. Sci. 1019, 171-177.

[26] Tyner, S.D., Venkatachalam, S., Choi, J., Jones, S., Ghebraniousk, N., Igelmann, H., Lu, X., Soron, G., Cooper, B., Brayton, C., et al., (2002). p53 mutant mice that display early ageing-associated phenotypes. Nature415(6867), 45-53.