Pacific Symposium on Biocomputing 7:18-29 (2002)

SINGULAR VALUE DECOMPOSITION REGRESSION MODELS FOR CLASSIFICATION OF TUMORS FROM MICROARRAY EXPERIMENTS

DEBASHIS GHOSH
Department of Biostatistics, University of Michigan
1420 Washington Heights, Ann Arbor, MI 48109-2029
ghoshd@umich.edu

An important problem in the analysis of microarray data is correlating the high-dimensional measurements with clinical phenotypes. In this paper, we develop predictive models for associating gene expression data from microarray experiments with such outcomes. They are based on the singular value decomposition. We propose new algorithms for performing gene selection and gene clustering based on these predictive models. The estimation procedure using the regression models occurs in two stages. First, the gene expression measurements are transformed using the singular value decomposition. The regression parameters in the model linking the principal components with the clinical responses are then estimated using maximum likelihood. We demonstrate the application of the methodology to data from a breast cancer study.

1 Introduction

DNA biochips have the potential of significantly impacting the study of human disease. By simultaneously gauging the expression of thousands of genes in clinical specimens, a wealth of data points is generated coalescing to form a molecular fingerprint of a disease process. Such experiments have been performed on acute leukemias, lymphomas, breast cancers and cutaneous melanomas.[1,2,3] Obtaining large-scale gene expression profiles of tumors should theoretically allow for the identification of subsets of genes that function as prognostic disease markers or biologic predictors of therapeutic response.

Most primary analyses have utilized hierarchical clustering techniques.[4] However, in many instances, there is external clinical information (such as survival time or tumor type) available. Typically, the investigators use these variables in secondary analyses. For many molecular profiling studies, the scientific goal appears to be finding candidate genes that successfully discriminate between disease classes based on the clinical phenotype. These genes can then be screened for further follow-up studies using immunohistochemical techniques such as tissue microarrays.[5]

Some preliminary work has been put forward correlating gene expression data with clinical outcomes.[6,7] However, these approaches have been univariate and ignore correlations between genes. A problem with joint modeling of gene

effects on clinical outcomes is that the number of genes is typically much larger than the number of samples profiled. In statistical terminology, the dimension space of the predictors is much larger than that of the independent samples. Consequently, it is not possible to calculate regression parameter estimates using traditional statistical procedures.

In this paper, we develop a regression framework based on the singular value decomposition for correlating gene expression data with clinical phenotypes. We explore the use of these models for three goals: prediction, gene selection and clustering. We propose novel algorithms for accomplishing the latter two tasks. While the framework presented here can be generalized, we are motivated by the specific problem of modeling the association between gene expression data with type of tumor. Singular value decomposition has been applied to other areas of microarray data analysis. [8,9,10] In the statistical literature, singular value decomposition analysis is known as principal components analysis; we will use the two terms interchangeably throughout the article. Regression modeling using SVD has been done with great success in other areas of application, such as chemometrics.[11] A complication in the current setting that does not arise in other applications is that the clinical outcome may not be continuous; our proposal here involves using categorical regression models[12] for associating the gene expression measurements with tumor type. We demonstrate the procedure using data from a recently published breast cancer study.[13] Because of space limitations, we refer the interested reader to the following URL for more details regarding this project and the analysis of the breast cancer data:

http://www.sph.umich.edu/~ghoshd/SVD/.

2 Methods

Before describing the regression model for correlating gene expression profiles with tumor phenotype, we introduce some notation. Let $\mathbf{X}_i$ denote the p-dimensional column vector of gene expression measurements for the ith subject, $i = 1, \ldots, n$. Note that p will typically be much larger than n. For $i = 1, \ldots, n$, we define Y_i to be the tumor type for the ith individual; this will take values $0, 1, \ldots, J-1$, where J is the number of tumor types. The class $Y = 0$ will be known as the reference category or reference tumor type. We will assume that the $\mathbf{X}_i$ are standardized across chips to have mean zero and variance one for each gene.

2.1 Regression model and estimation

We formulate the effects of gene expression on tumor type using the following multinomial logistic regression model:

$$\log \frac{P(Y_i = r)}{P(Y_i = 0)} = \beta_{r0}^T \mathbf{X}_i, \tag{1}$$

where $P(A)$ is the probability of the event A, $\mathbf{a}^T$ is the transpose of the vector or a matrix $\mathbf{a}$, and β_{r0} is a p-dimensional vector of unknown regression coefficients, $r = 1, \ldots, J-1$. The model is quite general in that separate gene effects are specified for each of the J(J-1)/2 tumor comparisons. More structure can be imposed by placing constraints on β_{r0} $(r = 1, \ldots, J-1)$. For example, we could set $\beta_{r0} = \beta_0$ for all r. This corresponds to a one-unit change in expression level for any gene having the same effect for discriminating any two tumor classes.

In a typical microarray experiment, it is not possible to estimate the parameters in (1) using standard statistical methods because p is much larger than n. We propose using the singular value decomposition to reduce the dimension of β_{r0}. If we let $\mathbf{X}$ denote the $p \times n$ matrix $[\mathbf{X}_1 \cdots \mathbf{X}_n]$, then the singular value decomposition leads to the following decomposition of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{UDV}, \tag{2}$$

where $\mathbf{U}$ is $p \times n$ matrix, and $\mathbf{D}$ and $\mathbf{V}$ are $n \times n$ matrices. The columns of $\mathbf{U}$ are orthonormal, i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$, the $n \times n$ identity matrix. The diagonal matrix $\mathbf{D}$ contains the ordered eigenvalues of $\mathbf{X}$ on the diagonal elements so that $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$, where $d_1 \geq d_2 \geq d_3 \geq \cdots \geq d_n \geq 0$. We will assume without loss of generality that $d_i > 0$ for $i = 1, \ldots, n$. Finally, $\mathbf{V}$ is the $n \times n$ singular value decomposition factor matrix and has both orthonormal rows and columns. The algorithms used to compute the singular value decomposition are typically iterative and quite computationally efficient.[14]

The effect of the singular value decomposition is to project high-dimensional multivariate data into a lower dimensional subspace. By plugging (2) into (1), we obtain the following model:

$$\log \frac{P(Y_i = r)}{P(Y_i = 0)} = \gamma_{r0}^T \mathbf{W}_i, \tag{3}$$

where γ_{r0} $(r = 0, \ldots, J-1)$ is a $n \times 1$ vector of regression coefficients and $\mathbf{W}_i$ $(i = 1, \ldots, n)$ is the ith column of the $n \times n$ matrix $\mathbf{W} \equiv \mathbf{DV}$. It can be shown that β_{r0} in (1) and γ_{r0} in (3) are linked by the following relationship: $\gamma_{r0} = \mathbf{U}^T \beta_{r0}$.

By transforming the regression model from (3) into (1), we have reduced the dimension of the space for the predictor variables from p to n. This makes the problem computationally tractable, i.e. model (3) can be fit using traditional statistical estimation procedures. We use the method of maximum likelihood to estimate γ_{r0} $(r = 0, \ldots, J-1)$.

2.2 Gene selection and clustering based on SVD regression

Ultimately, we are interested in determining which genes have the greatest ability in discriminating between disease classes defined by the clinical phenotype. This corresponds to ranking the components of β_{r0} $(r = 0, 1, \ldots, J-1)$. It would be desirable if we could backtransform the estimators of γ_{r0} in order to derive estimators of β_{r0} in (1) $(r = 0, 1, \ldots, J-1)$. However, this is not possible because the mapping from β_{r0} to γ_{r0} (defined by $\mathbf{U}$) is a many to one mapping, so the inverse mapping is not well-defined.

Our proposal is to rank the p genes using the vector of gene scores $\mathbf{s}_r = \mathbf{U}\widehat{\gamma}_r$ $(r = 0, 1, \ldots, J-1)$. This gives a measure of the p genes to discriminate between the rth category relative to the reference category. If one were to adopt a Bayesian framework for model (1), one can show that with a suitable choice of prior on the regression parameters, $\mathbf{s}_r$ is asymptotically equivalent to the posterior mode of β_{r0}.[15] However, our interest is in ranking the values of $\mathbf{s}_r$, not in performing formal inference. An advantage of this proposed gene selection scheme relative to previous approaches is that potential correlation between the genes is taken into account.

The variance-covariance matrix of the $\mathbf{s}_r$ $(r = 0, \ldots, J-1)$ can be standardized to yield a correlation matrix, which can then be used as an input in a hierarchical clustering algorithm. The clustering algorithm attempts to find relationships between these discriminating genes and is based on the assumption that mutual coexpression potentially implies a common regulatory mechanism or that the genes might be involved in the same pathway. This clustering procedure utilizes the clinical phenotype information in a sensible fashion. Previous clustering methods have failed to take this external information into account.[4]

2.3 Filtering genes

Typically in microarray experiments, the number of potential predictor genes will be on the order of thousands. In studies involving gene expression, it seems biologically plausible that only a fraction of the set of genes on the chip have real biological activity. Consequently, certain authors have suggested that reducing the initial number of variables under consideration leads to improved

predictive performance.[15,16] With the breast cancer data, we study the use of an initial preprocessing in order to filter out a subset of the original set of genes. We fit an analysis of variance (ANOVA) model of gene expression measurement versus tumor class individually for each gene. For each ANOVA model, we calculate an overall F-statistic; this yields a set of p F-statistics. We then take the M genes with the largest F-statistics as the potential predictor variables in the model. The effect of this variable selection is to eliminate genes whose power in discriminating between tumor types is not significantly above the experimental variability in the gene expression measurements. An empirical study of the effect of M on the predictive performance on the singular value decomposition regression modeling is given in the application to the breast cancer data.

It has been noted in the literature that variable selection is an inherently unstable procedure.[17] This instability will be even more apparent here because of the relatively small values of n. In order to stabilize the performance of the variable selection described in the previous paragraph, we also examined the use of bagging methods.[18] This method involves creating B perturbed versions of the original dataset by resampling from the set of independent samples B times. For each dataset, we rank the genes by the values of the F-statistic. We then compute the average rank of each gene over the B datasets and take those with the M highest averages. We break ties using random jittering.

2.4 Choosing number of principal components

A major issue in the application of singular value decomposition regression modeling to high-dimensional data is determining how many principal components to use in model (3). There are many ways of performing this variable selection.[11] We have employed leave-one-out cross-validation. In this procedure, one sample is removed from the dataset at a time. For a fixed number of principal components, say k, the regression model is fit to the remaining data. Based on the estimated model, the model is used to predict the tumor type of the withheld sample. An error measure is then calculated based on Hamming distance. We repeat this training procedure leaving out each of the other samples from the dataset one at a time; this yields an estimate of the classification error rate. This is done for every possible value of k; the value of k that yields the smallest classification error rate is then chosen. Leave-one-out cross-validation is a popular method in situations with small samples where no test data are available. We note that this is a data-driven rule for selecting the number of principal components to use in the modeling.

3 Application

In this section, we apply the proposed methodology to data from a study of BRCA1- and BRCA2-positive tumors.[13] In this study, 23 biopsy specimens of primary breast tumors were collected. Seven had BRCA1 germ-line mutations, and eight had BRCA2 germ-line mutations. In addition, another eight samples were collected that had neither BRCA1 nor BRCA2 germ-line mutations; these were treated as sporadic cases of breast cancer. The goal of the study was to determine if there were differences in global gene expression profiles that could be used to discriminate the three classes of cancer (BRCA1, BRCA2 and sporadic).

While we will not go into the details of the analysis performed by Hedenfalk et al., we do wish to make two points. First, univariate statistical methods were used in order to determine the ability of genes to discriminate between the tumor types. Second, the analysis of the data was divided into two subgroup analyses. The first subgroup comparison was between BRCA1-positive and sporadic tumors; the second involved comparing BRCA2-positive and sporadic tumors. While this analysis approach seems reasonable in terms of the scientific goals of the study, it is potentially statistically more efficient to incorporate the correlations between the three tumor classes as well as the genes in order to incorporate correlations between genes. In the discussion that follows, we take the sporadic tumor class to be the reference category.

We first focus on the performance of the principal components regression modeling in terms of the classification error rate, defined using Hamming distance. In particular, we look at the effect of varying M. The results are summarized in Figure 1. Based on Figure 1, the optimal number of principal components varies on M; however, it does not appear to be possible to derive a general rule. For example, for $M = 25$, we have one misclassification using the singular value decomposition procedure with 11 principal components in the model. Comparable optimal misclassification rates can be obtained using $M = 1500$ and $M = 3226$. Using cross-validation, the choice of the number of principal components will depend on the particular dataset. We also examined the effect of the bagging variable selection procedure described in the paper (data not shown). The bagging variable selection tends to improve the predictive performance of the singular value regression models; we refer the interested reader to our website for these results.

We now illustrate the ranking and clustering procedures based upon the SVD regression modelling. For the purposes of discussion, we take $M = 100$. Based on Figure 1, the number of principal components for $M = 100$ that minimizes the classification error rate is $k = 2$. We subsequently fit model (3)

with two principal components and estimate the regression parameters using maximum likelihood estimation. Based on fitting the model and the back-transformation described in Section 2.2, we can rank the genes in terms of their ability to discriminate between these three classes of tumors. A ranking of the top 20 genes from the subset of $M = 100$ and their corresponding gene scores for discriminating BRCA1-positive tumors from sporadic tumors is given in Table 1. A similar table of the top genes for discriminating BRCA2-positive tumors from sporadic tumors can be found at the website. Many of the genes on this list overlap with the discriminatory genes found by Hedenfalk et al., but there are also genes that do not make their list.

Finally, we wish to examine potential relationships between the genes in Table 1. One way to do this would be to simply cluster the genes using average linkage hierarchical clustering.[4] We do not present the resulting dendrogram here; it can be found at our website. However, if we now use the estimated variance matrix of the gene scores from the SVD regression model based on two principal components as the basis of the hierarchical clustering, this yields the dendrogram in Figure 2. In particular, we find that there are two distinct groupings with the second dendrogram, but this increase in separation comes at the price of losing the finer substructure between the genes . The reason for this because the estimates of the gene scores are highly correlated. Consequently, most of the off-diagonal entries of the distance matrix used in the hierarchical clustering algorithm are close to one. However, the initial separation between the genes is greater using this method compared to that from performing hierarchical clustering on the gene expression data where the tumor class is not taken into account (data not shown).

4 Discussion

In this article, we have developed a singular value decomposition regression modelling approach for correlating gene expression profiles with tumor class in microarray settings. This methodology is important for determining the diagnostic and predictive ability of microarray technology in clinical settings. While we have focused mainly on a categorical response (tumor type), the ideas in this article can be applied to other types of clinical phenotypes, such as censored failure times, using different regression models in lieu of (1). Singular value decomposition regression models have a rich tradition in other fields of application, but the presence of noncontinuous clinical phenotypes introduce new issues in statistical modelling.

We utilized SVD regression modeling for three purposes. First, predictive models were constructed in the situation where the dimension of predictors is

much larger than that of the independent samples. Second, it provided the basis for ranking genes in terms of their discriminative abilities. Finally, the parameter estimates from the principal components regression method were used to cluster genes. Based on the analysis of the breast cancer data, we found that the SVD regression approach is successful for prediction and variable selection. However, it is problematic for clustering in terms of finding finer structural relationships among genes.

As was mentioned in the Introduction, singular value decomposition regression models have been applied in other disciplines; one unique challenge here is that the outcome measure is not continuous. A major advantage of this method is that it can accommodate the scenario where the number of predictors is larger than the number of independent samples. However, other predictive modelling methods exist in this setting, such as partial least squares and ridge regression.[11] It would be very useful to compare these methods in terms of their predictive modelling capabilities and is a current area of focus of our research. However, it should be noted that it does not appear to be straightforward to develop gene selection and clustering schemes based on partial least squares.

Because gene expression data are highly multivariate, they are inherently complex. This research has also demonstrated that multiple levels of data analysis are needed in order to perform classification of tumors using microarray data.

References

1. Golub T. R. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
2. A. A. Alizadeh *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503–511.
3. M. Bittner *et al* (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**: 536–540.
4. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Nat Acad Sciences* **95**, 14863–14868.
5. J. Kononen *et al.* (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* **4**, 844–847.
6. P. J. Park, M. Pagano and M. Bonetti. (2000). A nonparametric scoring algorithm for identifying informative genes from microarray data. In *Proc Pac Symp Biocomputing.*

7. V. G. Tusher, R. Tibshirani and G. Chu. (2001) Significance analysis of microarrays applied to ionizing response. *Proc Nat Acad Sciences* **98**, 5116–5121.
8. S. Raychaudhuri, J. M. Stuart and R. Altman. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Proc Pac Symp Biocomputing*.
9. N. S. Holter *et al.* (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Nat Acad Sciences* **97**, 8409–8414.
10. O. Alter, P. O. Brown and D. Botstein (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Nat Acad Sciences* **97**, 10101–10106.
11. I. E. Frank and J. H. Friedman (1993). A statistical view of some chemometric regression tools (with discussion). *Technometrics* **35**, 109–135.
12. A. Agresti. *Categorical Data Analysis*. (1990). New York: John Wiley and Sons.
13. I. Hedenfalk *et al.* (2001) Gene expression profiles in hereditary breast cancer *NEJM***344**, 539–548.
14. G. H. Golub and C. F. van Loan. *Matrix Computations*. (1996). Baltimore: John Hopkins University Press.
15. West, M. (2001). Bayesian regression analysis in the "large p, small n" paradigm. Technical Report, Institute of Statistics and Decision Sciences, Duke University.
16. Dudoit, S., Fridlyand, J. and Speed, T. P. (2001). Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report, Department of Statistics, University of California at Berkeley.
17. L. Breiman (1996). Heuristics of instability and stabilization in model selection. *Ann Stat* **24**, 2350–2383.
18. L. Breiman (1996). Bagging predictors. *Mach Learn* **24**, 123–140.

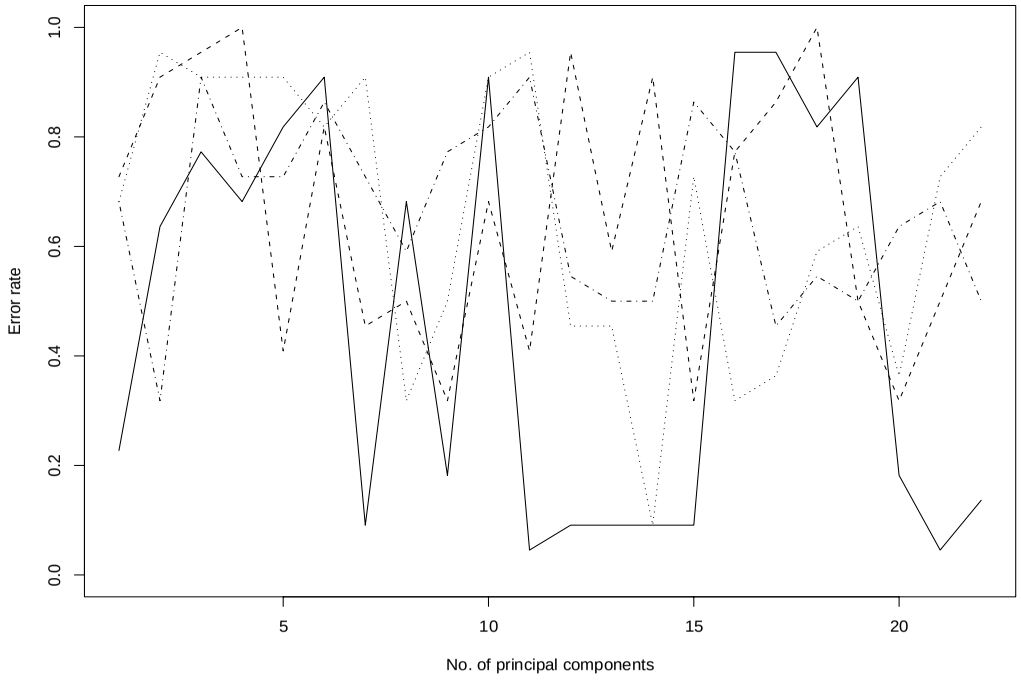

Figure 1: Plot of estimated classification error rates (based on Hamming distance) versus number of principal components. Solid line: $M = 25$; dashed line: $M = 100$; dotted line: $M = 1500$; dashed/dotted line: $M = 3226$.

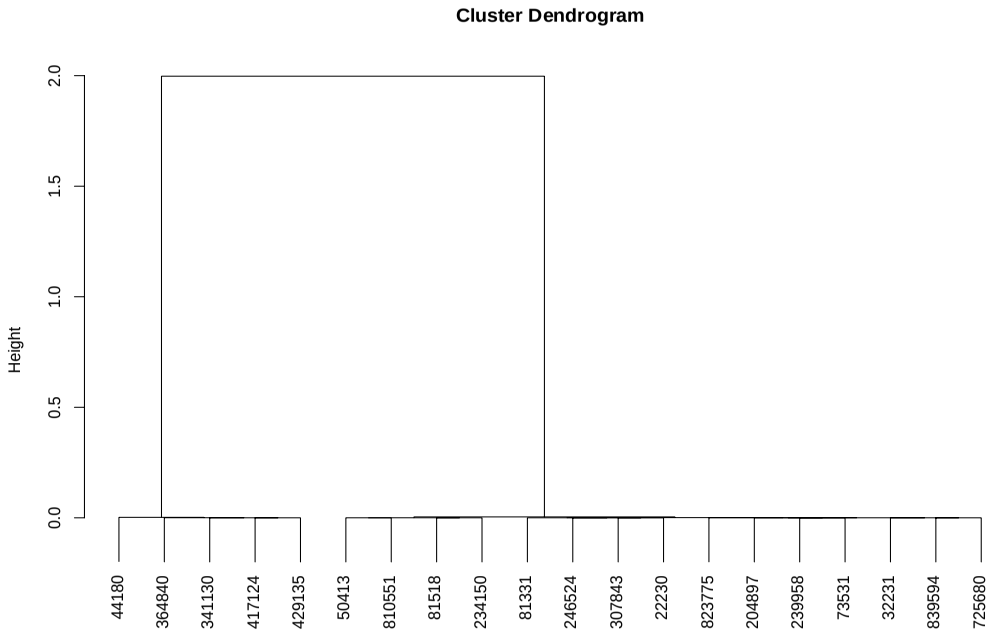

Figure 2: Hierarchical clustering dendrogram of genes from Table 1 based on gene scores. Average linkage clustering used.

Table 1: List of ranked genes and gene scores for discriminating BRCA1-positive tumors from sporadic breast cancer tumors.

Clone	Gene	Score
823775	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3	0.204
364840	ESTs, Moderately similar to mouse Dhm1 protein [M.musculus]	0.194
44180	alpha-2-macroglobulin	0.175
32231	KIAA0246 protein	0.172
81518	apelin; peptide ligand for APJ receptor	0.171
417124	APEX nuclease (multifunctional DNA repair enzyme)	0.167
839594	ribosomal protein L38	0.155
239958	DKFZP586G1822 protein	0.154
234150	myotubularin related protein 4	0.151
73531	nitrogen fixation cluster-like	0.150
204897	phospholipase C, gamma 2 (phosphatidylinositol-specific)	0.148
725860	transcription factor AP-2 gamma (activating enhancer-binding protein 2 gamma)	0.146
246524	CHK1 (checkpoint, S.pombe) homolog	0.144
429135	suppression of tumorigenicity 13 (colon carcinoma) (Hsp70-interacting protein)	0.143
307843	ESTs	0.142
22230	collagen, type V, alpha 1	0.131
50413	armadillo repeat gene deletes in velocardiofacial syndrome	0.130
81331	fatty acid binding protein 5 (psoriasis-associated)	0.129
341130	retinoblastoma-like 2 (p130)	0.128
810551	low density lipoprotein-related protein 1 (alpha-2-macroglobulin receptor)	0.127