Prognostic Molecular Classification of Breast Cancer Based on Features Extracted from a Scale Space

Yingchao Wu, Jeroen de Ridder and Marcel J.T. Reinders
Bioinformatics, Delft University of Technology, The Netherlands
Email: Y.Wu-4@student.tudelft.nl

Abstract

Breast cancer is one of the most prevalent cancers affecting females in the world. In recent years, many cancer researchers have been trying to determine molecular prognosis tools that predict cancer patient treatment response and/or chance of survival. In particular, the determination of gene expression signatures obtained by feature selection methods applied to large microarray datasets has shown potential. The main purpose of this study is to extend these gene signatures and molecular prognostic classifiers by investigating features constructed from a scale-space representation of the microarray data.

Here, we construct a scale space by first mapping all genes to a one-dimensional functional space using protein family information. Next, we applied successive smoothing to the expression values resulting in one scale-space representation of the gene expression data from one sample. At the lowest scale, the scale space contains the original gene expression values, whereas at higher scales meta-features are formed, which are weighted sums of groups of genes.

To test whether a scale-space representation is useful we performed feature selection and classification on a publicly available breast cancer expression dataset. We found that, instead of signatures consisting of single genes, meta-genes (i.e. groups of genes) that exist at higher scales were preferentially selected. We furthermore determined cross-validation errors using seven distinct classifiers (NMC, LDC, QDC, FISHERC, PARZENC, 3NNC, and LOGLC) and found that better performance is obtained using the scale-space representation than with the traditional representation of the gene expression data. As a result, we conclude that the scale-space analysis constitutes a potent way of selecting molecular signatures and is useful for prognostic classification.

Keywords: Breast cancer; Scale space; Classification; Feature selection.

1 Introduction

Breast cancer is one of the most diagnosed human cancers among females, and the chance being diagnosed with breast cancer with increasing over time. However, the mortality rate of it has been a progressive decrease every year. This is the result not only of the technology advances in the field of medical care, but also of the proposition of several molecular biological classification schemes [1] [2]. It also benefit from more accurate and robust classifier is obtained by analyzing multiple cancer datasets [3].

Generally speaking, the information of biological data is numerous and complex. With the development of computer technology, both the large number of molecular data and a variety of clinical information can be explored and analyzed in depth. In this report, we focus on exploring prognostic classification of breast cancer based on scale-space system [4]. **Figure 1** shows different steps

required by direct analysis and scale space based analysis of microarray data.

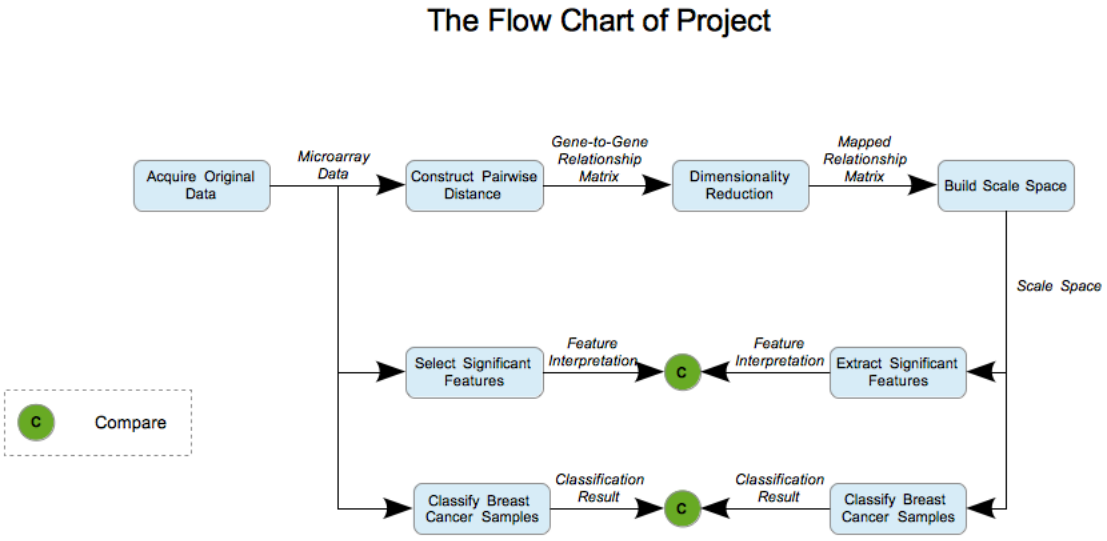

Figure 1: The whole flow-chart

Most of breast cancer classification schemes are based on a large numbers of DNA sequences. The microarray that simultaneously measures the expression levels of a large number of genes is a well-known method to provide these molecular data, and it is used in our study. The main principle of microarray is hybridization between DNA and cDNA sequences. The microarray experience workflow is shown in **Figure 2**. As depicted by the image, probes made of thousands of oligonucleotide are synthesized in the chip. Then, these probes hybridize with the target labeled by a radioactive marker. The hybridization result can be obtained by scanning the radioactive maker. Processing and analyzing the intensity and distribution of hybridization signals reveal the gene expression profiles of target.

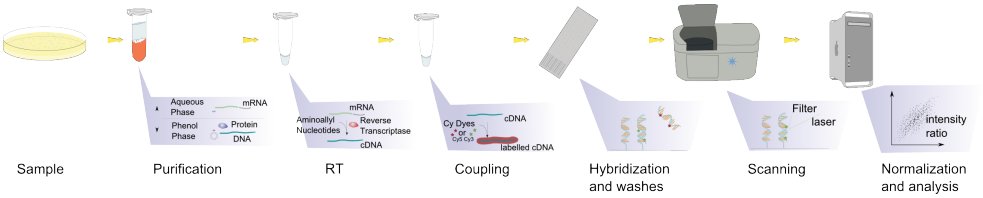

Figure 2: Concrete steps of the microarray experiment. Figure adapted from Wiki.

The idea of scale space is that we can catch different views in the different observing distances. It means it is important to know the particular observing distance where the interesting objects will be shown. Likewise, it is important to know the scale of processed signal for classification problem. The traditional classification is based on the signal of original scale, which includes only the interesting information of original data. However, different information can be obtained by stepping into different scales, and generally, we do not know which scale contains the information of interest. Therefore, we need use information of multiple scales. The scale-space system [4] [5] [6] can be used to reach this goal.

The scale-space system was firstly proposed by Witkin in [4] to apply in image domain. He per-

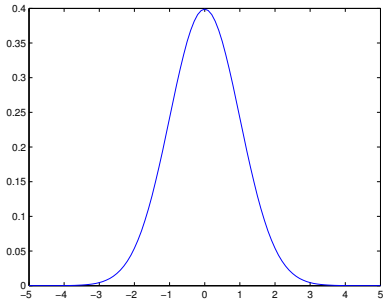

Figure 3: One-dimensional Gaussian function

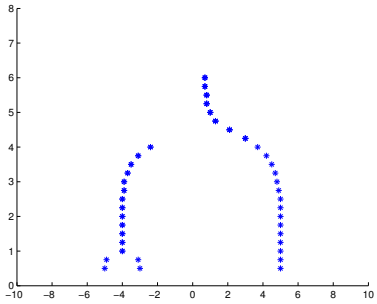

Figure 4: The scale-space tree of one-dimensional signal that includes three pulses

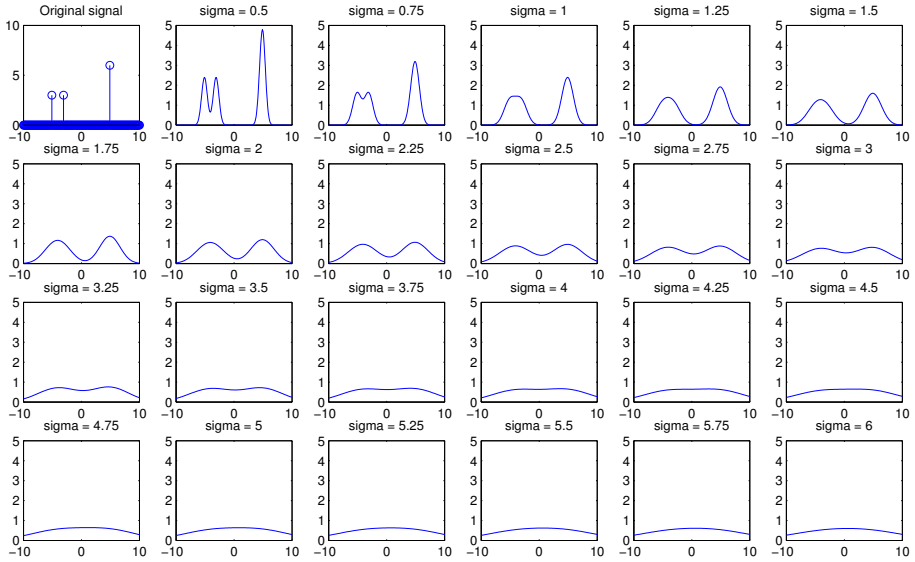

Figure 5: A series of Gaussian blurred results of one-dimensional signal that includes three pulses

formed a Gaussian scale-space framework, which uses the Gaussian function $g(x;t) = \frac{1}{\sqrt{2\pi t}} e^{-(x^2)/2t}$ as a convolution kernel to smooth the original signal, as shown in **Figure 3**. With smoothing the original signal, the trivial information that can be seen as noise will be removed. The parameter t, called scale parameter, determines the width of the Gaussian kernel. By increasing the value of t continuously, the blurred results at different scales can be obtained, as shown in **Figure 5**. Then, the scale-space tree, also known as deep structure, can be extracted by tracking the local extrema in all blurred signals, as shown in **Figure 4**. The deep structure reveals the significant changes of signal when is blurred step by step. It gives a comprehensive view of interesting information in multiple-scales signal.

In recent years, the scale-space system is gradually applied into molecular biological domain. The application of scale-space system within modular biology domains brings several advantages, such as: allowing deeper relationship between biological elements to be revealed. [7] [8] [9], or reducing the data noise in molecular classification. Similar to distance between pixels under image processing context, the distance between biological elements should be defined before establishing the scale-space system. The distance between biological elements can be defined using their spatial relationship, se-

quence or functional similarity [10], or phylogenetic relationship [11]. In our study, we are interested in the genes that are related with breast cancer. When an undefined gene has a closer functional distance with the defined breast cancer gene, it is more likely that this undefined gene is related to breast cancer. Therefore, we select functional similarity to describe the pair-wise distance between genes. The collection of gene distance from the biological data forms the gene distance matrix.

The gene distance matrix, or gene relationship matrix, is a dataset with high dimensionality, containing redundant information, and it is difficult to be visualized. In addition, it is computationally intensive to perform Gaussian smoothing on high-dimensional dataset, also large error can be expected from the end result. To tackle the aforementioned disadvantages, dimensionality reduction to the gene relationship matrix is necessary. We will describe the dimensionality reduction scheme in details in the next section.

The scale-space system can be created after performing dimensionality reduction to the gene relationship matrix. Then the breast cancer metastasis can be prognosed by classification. To increase the accuracy of classification, the feature selection is usually performed to reduce the redundant features in classification. It should be noted that a selection bias has to be corrected before estimating the classification error [12]. In [12] two bias correction methods are proposed, involving either the cross-validation or bootstrap estimation. In our study, we use the cross-validation to perform bias correction.

2 Materials and Methods

In the cancer research, multiple clinical information of patients has to be taken into account together with the molecular data. One of them is 'survival time', which is the number of years patients have survived post to the diagnosis of breast cancer. The survival time of every patient has large variation between each other, ranging from 1 year, to more than 10 years. Therefore, we take 5 year, our first clinical information, as a threshold to classify the patient samples.

On the other hand, there are other factors that affects the research. For instance, the death of patient caused by reasons other than breast cancer, termination of research activities by the patient, or mistakes made by patient in submission the result of research and so on. All the aforementioned situations are considered as the second clinical information, called censoring, to help classify the samples. For example, when the survival time is smaller than five years and not censoring, the sample is labeled by 'poor'. On the other hand, if the survival time is larger than five years and censoring, the patient sample is labeled by 'good'. Other samples are labeled as 'other' any are eliminated from later analysis, obtaining a two-class sample dataset.

The breast cancer sample dataset used is named as 'Miller' [3], which is measured on the Affymetrix platform. It consists of microarray data of 22268 human genes with Entrez gene ID and other clinical information from 247 breast cancer patients. By classifying the sample set with the clinical information described above, 37 samples are labeled by 'poor', while 156 samples are labeled by 'good'.

2.1 Data preprocessing

Sample classes tend to have different size, this is know as class imbalance. When the size of sample classes have significant difference, the classification result is likely to be unreasonable. Since an unlabeled sample has a greater chance to be classified into the class with larger size during training, resulting in a biased testing classification. In 'Miller' dataset, the size of patient samples labeled as 'good' is much more than samples labeled as 'poor'. Thus, before prognostic prediction, we have to remove partial samples from the 'good' class to balance the size difference between the 'good' class and the 'poor' class.

As mentioned in former section, the distances between elements should be established before using scale-space system. In our study, we tried three types of element distance matrices: correlation relationship between genes, functional distance between genes and functional distance between genes and gene families. These matrices are constructed based on the protein-to-family functional distance matrix[1]. Proteins are shown with Ensembl gene ID in [1]. The following are steps to connect the breast cancer dataset 'Miller' with dataset in [1].

I Stage 1: converting EntrezIDs to EnsemblGeneIDs.

(a) step 1: removing genes in our dataset with 'NaN' in original EntrezID list; (Number of genes: $22268 -> 21177$)

(b) step 2: using a web-tool convert EntrezIDs to EnsemblGeneIDs (Gene ID Conversion Tool: *http://david.abcc.ncifcrf.gov/conversion.jsp*);

(c) step 3: building corresponding EnsemblGeneID list for retained genes obtained by step 1;

(d) step 4: removing genes in our dataset with 'unknown' EntrezIDs or EntrezIDs converted unsuccessfully to EnsemblGeneID in step 2. (Number of genes: $21177 -> 20322$)

II Stage 2: by EnsemblGeneID list obtained in last step, getting the corresponding ProteinNumber (defined in protein-to-protein similarity dataset) list.

(a) step 1: building corresponding ProteinNumber list for retained genes in last step;

(b) step 2: removing genes that have not corresponding proteins in protein-to-protein similarity dataset. (Number of genes: $20322 -> 19844$)

Comparing to gene-to-gene correlation (GGC) matrix, the gene-to-gene functional distance (GGD) matrix and gene-to-family functional distance (GFD) matrix are established on functional relationship between biological elements, and hence are more biologically reasonable. Among GGD and GFD, GFD matrix is chosen, because several disadvantages of GGD. For instance, the computational complexity of GGD matrix is quite high, and the dimensional reduction result of GGD obtained by t-Distributed Stochastic Neighbor Embedding method (tSNE) [13] is always a uniform distribution, rendering it useless.

Because one gene can encode multiple types of protein, it is necessary to set criterion on how to calculate the exact distance between two genes based on protein-to-family distance. The formula used is $gfd(i,j) = min(pfd(k_1 : k_t, j))$. For the $gfd(i,j)$, the minimum value of all protein-to-family distances from family j and t proteins k_1 to k_t and related to gene i, is used as minimum estimation of distance between gene i to family j. The specific steps are shown as following.

I Step 1: according to the ProteinNumber list '*protein_retain*', which includes related proteinIDs of every gene, produce an unique protein list '*uni_protein*';

II Step 2: using '*uni_protein*' and proteinID list in protein-to-protein similarity dataset, obtain protein list 'p' and protein-to-family e-value list '*evalue*' related to genes in our data;

III Step 3: using the e-value list '*evalue*', compute the gene-to-family functional distance matrix.

2.2 Dimensionality reduction

When elements are distributed in high-dimensional space, the computational cost is usually expensive. In the mean time, the classification accuracy is low. To overcome this problem, dimensional reduction methods are proposed to map the dataset from the high-dimensional space to a lower dimension. Generally, the original data is mapped to the intrinsic-dimensional space.

One of the dimensional reduction methods is Multi-Dimensional Scaling (MDS), which starts with a matrix of item-to-item similarity or dissimilarity, and then provides a location to each item in a

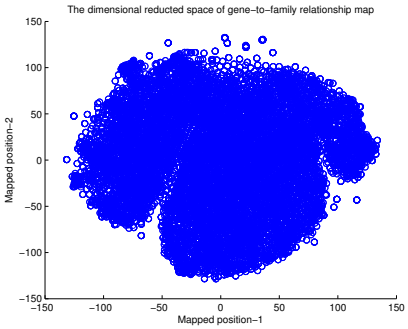

Figure 6: Mapped result of gene-to-family relationship matrix obtained by tSNE

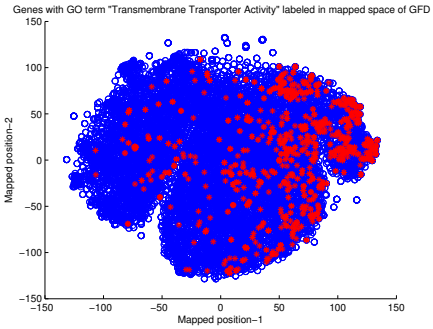

Figure 7: Labeled genes by GO term in mapped result of GFD by tSNE

low-dimensional space such that distance between all items are preserved. Another method is tSNE [13], which is improved version of Stochastic Neighbor Embedding (SNE) [14] method by adding student-t distribution. tSNE is good at capturing much of local structure of the high-dimensional data and revealing global structure.

In the DRToolbox of Matlab, functions '*mds*' and '*tsne*' can fulfilling two dimensional reduction methods mentioned above. The function '*tsne*' firstly performs Principal Component Analysis (PCA) to preprocess the original data, obtaining a initial-dimension dataset for tSNE. A Gaussian kernel with fixed perplexity, which can be seen as the number of effective nearest neighbors, is used to smooth. The normal range of perplexity is [5 50] and a larger or denser data normally need a larger perplexity. We use a perplexity of 30. Comparing with MDS, the mapping result of tSNE contains more obvious distribution characteristic in scatter plot, which can be seen in **Figure 6**. To check whether the mapped result is biologically reasonable, genes with Gene Ontology (GO) term 'Transmembrane Transporter Activity' are labeled in the mapped space with red point, as shown in **Figure 7**. From the labeled result, we can see that genes with same GO term tend to cluster together. Following are the steps to label genes in mapped space by GO term.

I Step 1: using DAVID Functional Annotation Tool, input our EntrezID list of gene set, and get the functional annotation summary.

II Step 2: select related annotation categories in the summary. We choose three categories ('GOTERM-CC-FAT', 'GOTERM-BP-FAT' and 'GOTERM-MF-FAT'), and obtain 1914 GO term records. 1630 genes in our gene set are missed in all records.

III Step 3: by these records, find out the corresponding EntrezID list of GO-term records, which we interest in.

IV Step 4: using these EntrezID lists, label genes with GO term 'Transmembrane Transporter Activity' in the mapped result to visualize it.

2.3 Scale space construction

To reduce the computational time of scale space construction, the difference of expression values of each gene between all patient samples are calculated. Then the difference values are sorted, and the genes corresponding to top 2000 difference values are chosen to build the scale space. As shown in **Figure 8**, chosen genes spread out in two dimensional mapped space of tSNE, which suggests chosen genes have global representative of all genes.

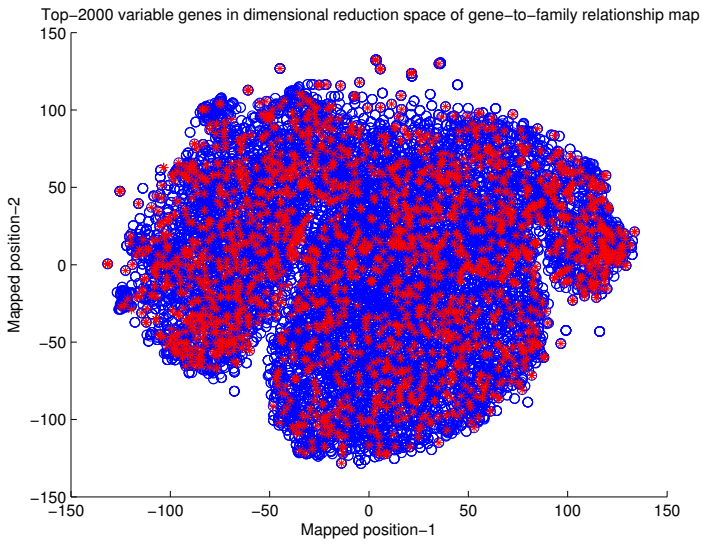

Figure 8: 2000 most variable genes labeled in 2D mapped space obtained by tSNE

Based on mapped result obtained by tSNE and 2000 chosen genes, the Gaussian scale space can be then constructed. The specific steps and mathematics used are shown as following.

I Step 1: combine the reduced distance matrix (d-dimension) with measurement value of all elements, which can be formulated as $G(\mathbf{x}) = \sum_{n=1}^{N} g_n \delta^d(\mathbf{x} - p_n)$, where g_n represents the measurement value of element n, and p_n represents the position of element n. in our case, the measurement value is gene-expression value. All elements are decided positions by mapped matrix, and values by corresponding measurement value.

II Step 2: obtain the scale space representation of all elements in scale-level h by convoluting $G(\mathbf{x})$ with the Gaussian kernel $K(h)$ of width h: $\hat{G}(\mathbf{x}; h) = G(\mathbf{x}) * K(h) = \sum_{n=1}^{N} g_n \exp(-\frac{1}{2}||\frac{\mathbf{x}-p_n}{h}||^2)$. By varying the width of Gaussian kernel, the scale space representation of different scale-level can be obtained.

III Step 3: build the critical curve by linking the local extrema in all convolution results. The set of all critical curves is the scale-space tree.

Based on the position data and expression value data of chosen genes, the scale-space tree with 72 scale levels is obtained for 74 patient samples. In first scale-level, the tree-matrix contains the original position of all genes. Because most of local extrema changes happen in low scales during scale-level increasing, a logarithmic scale of kernel width is used. The scale-space tree for one patient sample is shown in **Figure 9**.

2.4 Feature selection

Before prognosing the breast cancer metastasis, several feature selection methods are applied to reduce the feature dimension and to reduce the classification error. First, Principal Component Analysis (PCA) method is experimented. It can map a high feature-dimension to lower one and preserves the main component. Then, the performance of 'individual' feature selection is also evaluated. It simply chooses d individual best features from all features. However, the top d individual best features does not necessarily form the best feature subset available from the whole feature set. There, another feature selection method, known as 'forward' is explored, which starts with an empty feature subset, and then, one at a time, keeps adding feature that gives best performance considering

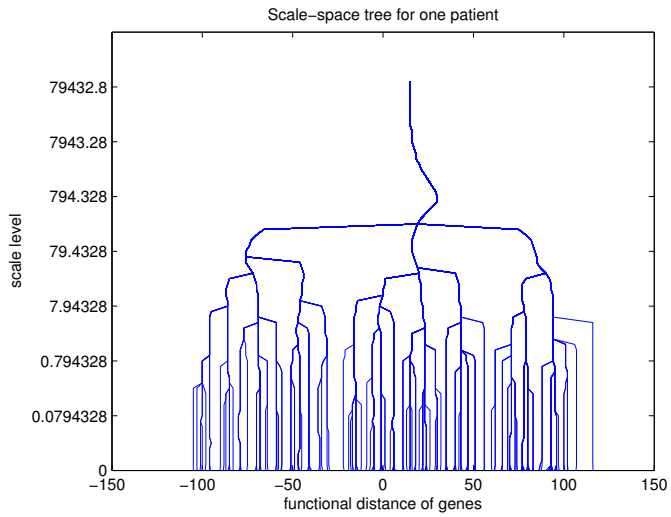

Figure 9: The scale-space tree for first patient with 72 scale-levels

entire chosen feature set.

As aforementioned, when the feature selection and classification is directly applied to original dataset, the classification error estimation is biased and unreasonable. To avoid the selection bias, the function 'crossval_featsel' was constructed. Firstly, the function divides the whole dataset into n parts. In each loop, the $n-1$ parts are used for feature selection and classification training, while the remaining part is used for classification testing with same mapped feature space as training set. The final result is mean error of all classification errors obtained in each loop.

3 Experimental results and discussion

3.1 Explore the prognostic gene of breast cancer based on scale space

To explore the prognostic genes of breast cancer, the scale-space tree of samples in two patient classes are compared and analyzed with density estimation and t-test statistics.

First, using density estimation, we can obtain a general comparison between two sample classes. The whole scale-space tree dataset is split to two subset according to the labels. Then, the probability density estimation of trees for two sample classes can be measured in some specific scale-levels, as shown in **Figure 10**. From the estimation results, we can see that in lower scale levels, the differences between trees in two classes are very small, while there are more obvious differences with an increased scale level. Generally speaking, gene features in higher scale-level tree are more significant for classification than in lower scale level.

Another prognostic analysis is the t-test statistics, which can give more explicit comparison between two sample classes. Firstly, one gene in one scale level can be seen as a feature, thus there are 144000 (2000*72), features in the whole scale space. Comparisons between two patient classes based on these 144000 features are calculated respectively, and the results are 144000 p-values. The gene feature with smaller p-value can be seen as a significant feature. To find out significant features from scale-space tree, a threshold was set to chose genes in single scale level with small p-value, as shown in **Figure 11**. To explore the suitable threshold value, different significant gene sets obtained by different thresholds, where log(p-value) was set to 0.001, 0.005, 0.01, 0.02, 0.05, 0.1, were seen as classification feature set to perform cross-validation, respectively. The best classification error

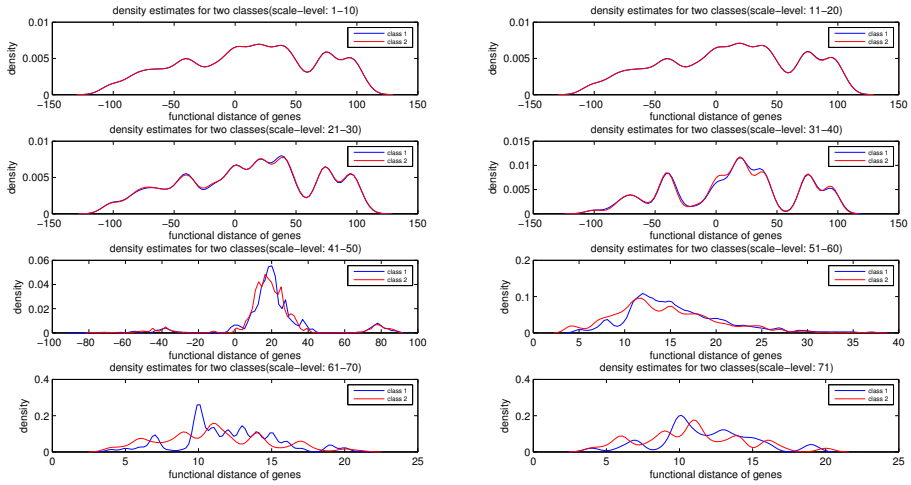

Figure 10: Probability density estimation of scale-space tree in two sample classes

estimation, 0.3515, is given by the threshold of 0.02, which is selected for further research.

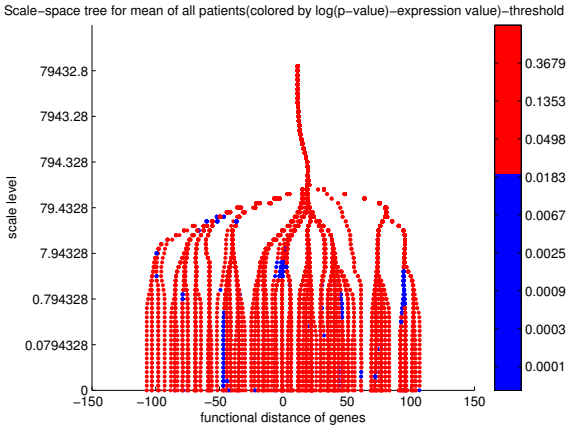

Figure 11: Significant features selected by t-test in mean scale-space tree

As shown in **Figure 11**, hundreds of features with log(p-value) smaller than 0.02 are marked with red color. The gene features, which are marked in a continuous series of scale-levels, are particularly interesting. It should be noted that a marked feature in high scale level represents more than one gene. After removing duplicated genes, 146 genes are marked as significant prognostic genes for breast cancer. To evaluate the prognostic genes chosen by feature selection based on scale space, different feature selection algorithms are performed on original microarray data. The 'Individual', 'Forward' and 'Backward' selection algorithm have been tried. We carried out cross-validation on each gene subset obtained by different algorithm. The subset with highest classification accuracy, 0.3862, selected by 'Forward' selection algorithm based on 200 best features obtained by 'Individual' algorithm is consisted by 62 genes. The gene selection result is shown in **Table 1**. We can see that the prognostic gene selection accuracy of t-test on scale space gene data is slightly higher than 'Forward' feature selection with original gene expression data.

Gene selection	Number of selected genes	Number of breast cancer genes	Effective rate of selection
t-test	146	6	4.1%
'Forward' feature selection	62	2	3.2%

Table 1: The gene selection accuracy of different methods

3.2 Classification results

To evaluate the scale space method used for breast cancer classification analysis, the classification error estimation by cross-validation for different classifiers carried on original gene data and scale space gene data, respectively, are shown in following tables and compared.

3.2.1 Classification based on feature representation of data

In image processing domain, the classification can be utilized for three representations: feature, dissimilarity and pixel. Feature representation uses the feature to represent the dataset. This feature is collected from the dataset through measurement of the digit image. While the pixel representation uses the pixel value as the feature to represent the digit. In dissimilarity representation, the distance between images is used instead of the feature. In biological field, feature and dissimilarity both can be used as representation for classifying.

The error estimation for seven classifiers (NMC, LDC, QDC, FISHERC, PARZENC, 3NNC, and LOGLC) using feature representation of 2000 chosen gene expression data and scale space built by 2000 chosen gene are shown in **Table 2**. For each classifier, the smallest error is boldface in the table. We can see that all classifiers obtained higher accuracies for scale space gene data except for Parzen Classifier (PARZENC). The smallest classification error is 0.3706, obtained by 3-Nearest Neighbor Classifier (3NNC) carried on scale space gene expression data. In next section, we tried increasing the accuracy by reducing the number of features.

Classifier	NMC	LDC	QDC	FISHERC	PARZENC	3NNC	LOGLC
Classification error (Original gene expression data)	0.4120	0.4220	0.3997	0.4649	**0.3880**	0.4514	0.4668
Classification error (Scale space position data)	**0.4068**	0.4072	**0.3818**	**0.3920**	0.5000	0.3766	**0.3939**
Classification error (Scale space expression data)	0.4080	**0.3909**	0.3826	0.4090	0.5000	**0.3706**	0.4060

Table 2: Classification error based on feature representation of original gene expression data and scale space data

3.2.2 Classification based on feature representation after feature extraction

As mentioned in Section Materials and Methods, we know that an increase in the number of features is not equivalent to the improvement of classification accuracy. To avoid high error caused by large amount of features, one feature extraction method was used before classification: principal component analysis (PCA). The error estimation for seven classifiers after PCA for original gene

expression data and scale space gene data are shown in **Table 3**.

From this table, we can see that most of classification accuracies are slightly increased after feature extraction comparing to **Table 3**. Furthermore, all classifiers get smallest classification error on scale space data except Nearest Mean Classifier (NMC), and the highest classification accuracy, 0.3594, was obtained by 3NNC on scale space gene position data.

Classifier	NMC	LDC	QDC	FISHERC	PARZENC	3NNC	LOGLC
Classification error (Original gene expression data)	**0.4006**	0.3930	0.3981	0. 4031	0.4350	0.4019	0.3906
Classification error (Scale space position data)	0.4075	0.3894	**0.3688**	0.3894	0.5000	**0.3594**	**0.3875**
Classification error (Scale space expression data)	0.4022	**0.3825**	0.3875	**0.3825**	**0.4037**	0.3787	0.4050

Table 3: Classification error based on original gene expression data after PCA and scale space gene data after PCA

3.2.3 Classification based on dissimilarity representation of data

Another exploration of classification is that the feature replaced by dissimilarity as the representation elements in classification. In our study, we used the CityBlock distance as measurements between different elements. The error estimation for seven classifiers (NMC, LDC, QDC, FISHERC, PARZENC, 3NNC, and LOGLC) using dissimilarity representation based on 2000 chosen gene expression data and scale space built by these genes are shown in **Table 4**. Comparing with the classification errors shown in **Table 3**, there is no obvious improvement between classification based on dissimilarity representation and feature representation.

Classifier	NMC	LDC	QDC	FISHERC	PARZENC	3NNC	LOGLC
Classification error (Original gene expression data)	0.4140	0.4057	**0.3810**	0.4057	0.4123	**0.3810**	0.4157
Classification error (Scale space position data)	**0.4040**	**0.3957**	0.3850	**0.3957**	0.4023	0.3831	0.4088
Classification error (Scale space expression data)	0.4190	0.3993	0.4003	0.4010	**0.3920**	0.3887	**0.3973**

Table 4: Classification error based on dissimilarity representation of 2000 chosen original data and scale space data built by 2000 chosen genes

4 Conclusion

Technology advance based on microarray gene expression shows significant improvement for investigating breast cancer. In this report, we research on the scale-space system applied on prognostic gene

selection and classification method. Comparing to the traditional feature selection on original gene data, the t-test based on scale space built by microarray data and gene-to-protein family functional similarity has slightly better performance. In classification problem, the scale space method, as important multi-scale analysis method of information processing has obvious advantages, comparing to the traditional single-scale method. We can see the improvement through the classification errors in **Table 3** and **Table4**. Moreover, we found that comparing to other classifiers, the 3-Nearest Neighbor Classifier (3NNC) gives the better classification accuracy in our project. To improve the accuracy, principal component analysis is useful method that can be performed before classification. However, in our study, the feature replaced by dissimilarity representation is an acceptable but not necessary method for improving the classification.

At present, the scale space theory and its applications are mainly confronted by the following issues. Firstly, The differential of scale space has a large amount of computation. So the computing time is always too long. This disadvantage limits the application of scale space method in real-time processing, motion detection and other fields. How to design the fast and stable algorithm for scale space is an important problem. Secondly, it is a core issue in application of scale space that how to combine with the background of problem to determine the optimal scale.

Supplementary Information

The supplemental document contains additional methods experimented, figures obtained during the study but not included in this report and Matlab codes.

Acknowledgements

I would like to express my deepest appreciation to all those who provided me kind support and help during my thesis project. A special gratitude and thanks to my supervisor Jeroen de Ridder for his patient guidance and constant supervision through entire process of this master thesis. Furthermore I would like to acknowledge with much appreciation my professor Marcel J.T. Reinders, who provided me the opportunity to do this interesting project as well as all supports on the way. My appreciation also goes to Sepideh Babaei, who provided the data used in this project and helped me in processing them. Last but not least, I would like to thank my parents, friends and members of Pattern Recognition and Bioinformatics Group for the support and encouragement given to me.

References

[1] S. Babaei, E. van den Akker, J. de Ridder, and MJ. Reinders, "Integrating protein family sequence similarities with gene expression to find signature gene networks in breast cancer metastasis," *PRIB*, vol. 7036, pp. 247259, 2011.

[2] AE. Teschendorff, A. Naderi, NL. Barbosa-Morais, SE. Pinder, IO. Ellis, S. Aparicio, JD. Brenton, and C. Caldas, "A consensus prognostic gene expression classifier for er positive breast cancer," *Genome Biology*, vol. 7, pp. R101, 2006.

[3] MH. van Vliet, F. Reyal, MH. Horlings, MJ. Vijver, MJ. Reinders, and LF. Wessels, "Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability," *BMC Genomics*, vol. 9, pp. 375, 2008.

[4] A.P. Witkin, "Scale-space filtering," *Proceedings of the Eighth International Joint Conferences on Artificial Intelligence*, vol. 2, pp. 1019–1022, 1983.

[5] F. Meyer and P. Maragos, "Morphological scale-space representation with levelings," *Proceedings of the Second International Conference on Scale-Space Theories in Computer Vision*, pp. 187–198, 1999.

[6] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[7] M.Y. Park, T. Hastie, and R. Tibshirani, "Averaged gene expressions for regression," *Biostatistics*, vol. 8, no. 2, pp. 212–227, 2007.

[8] J. de Ridder, J. Kool, A. Uren, J. Bot, L. Wessels, and M. Reinders, "Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes," *Bioinformatics*, vol. 23, no. 13, pp. 33–41, 2007.

[9] M. Ceccarelli, A. d'Acierno, and A. Facchiano, "A scale space approach for unsupervised feature selection in mass spectra classification for ovarian cancer detection," *BMC Bioinformatics*, vol. 10, pp. S9, 2009.

[10] G. Lerman and B.E. Shakhnovich, "Defining functional distance using manifold embeddings of gene ontology annotations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 27, pp. 11334–11339, 2007.

[11] D.F. Robinson and L.R. Foulds, "Comparing of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

[12] C. Ambroise and GJ. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proc Natl Acad Sci USA*, vol. 99, no. 10, pp. 6562–6566, 2002.

[13] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[14] G. Hinton and S. Roweis, "Stochastic neighbor embedding," *in Advances in neural information processing systems*, vol. 15, pp. 833–840, 2003.