Mathematical Biosciences and Engineering

MBE, 17(3): 2453–2469.
DOI: 10.3934/mbe.2020134
Received: 28 February 2018
Accepted: 17 February 2019
Published: 25 February 2019

http://www.aimspress.com/journal/MBE

Research article

Computational methods for recognition of cancer protein markers in saliva

Ying Sun[1,2], Wei Du[3], Lili Yang[4], Min Dai[1], Ziying Dou[1], Yuxiang Wang[1], Jining Liu[1] and Gang Zheng[1,*]

[1] Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China
[2] Information Technology Research Base of Civil Aviation Administration of China, Civil Aviation University of China, Tianjin 300300, China
[3] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China
[4] Department of Obstetrics, The First Hospital of Jilin University, Changchun 130012, China

* **Correspondence:** Email: kenneth_zheng@vip.163.com; Tel: +86-022-60216678.

Abstract: In recent years, many studies have supported that cancer tissues can make disease-specific changes in some salivary proteins through some mediators in the pathogenesis of systemic diseases. These salivary proteins have the potential to become cancer-specific biomarkers in the early diagnosis stage. How to effectively identify these potential markers is one of the challenging issues. In this paper, we propose novel machine learning methods for recognition cancer biomarkers in saliva by two stages. In the first stage, salivary secreted proteins are recognized which are considered as candidate biomarkers of cancers. We picked up 557 salivary secretory proteins from 20379 human proteins by public databases and published literatures. Then, we present a training set construction strategy to solve the imbalance problem in order to make the classification methods get better accuracy. From all human protein set, the proteins belonging to the same families as salivary secretory proteins are removed. After that, we use SVC-KM method to cluster the remaining proteins, and select negative samples from each cluster in proportion. Next, the features of proteins are calculated by tools. We collect 24 protein properties such as sequence, structure and physicochemical properties, a total of 1087 features. An innovative procedure based on the local samples is proposed for selecting the appropriate features, in order to further improve the performance of SVM classifier. Experimental results show that the average sensitivity, specificity and accuracy of salivary secretory protein recognition using selected 32 features

in training set are 97.09%, 98.10%, 97.61%, respectively. The use of these methods can improve the accuracy of recognition by solving the problems of unbalanced sample size and uneven distribution in training set. In the second stage, we apply the best model to dig out the salivary secreted proteins from 58 reported cancer markers, and get a total of 42 proteins which are considered to be used for salivary diagnosis. We analyze the gene expression data of three types of cancer, and predict that 33 genes will appear in saliva after they are translated into proteins. This study provides an important computational tool to help biologists and researchers reduce the number of candidate proteins and the cost of research. So as to further accelerate the discovery of cancer biomarkers in saliva and promote the development of saliva diagnosis.

Keywords: salivary secretory protein; cancer biomarker; feature selection based on local samples; SVC-KM; computational methods

1. Introduction

In many cases, the diagnosis can only be made when the cancer cells metastasize to the surrounding tissues or when the whole body deteriorates [1]. At this moment, the traditional treatment methods for most patients are invalid. Some advanced diagnostic methods such as X-ray fluoroscopy and clinical biopsy of early thoracic cancers have improved the diagnostic ability of cancers. However, they can't meet the needs of early detection of cancer due to the lack of obvious specific symptoms in the early stage of cancer. Early diagnosis of cancer is of great importance for cancer control and prevention. Now it is generally believed that the occurrence and development of cancer is related to gene mutation [2,3], so molecular level detection method can detect the existence of cancer earlier. Cancer biomarkers are substances produced directly by tumor cells or by non-tumor cells induced by tumor cells. The detection of cancer biomarkers can judge the diagnosis, pathogenesis and prognosis of tumor.

Up to now, there are a lot of cancer markers in tissue samples, both in types and quantity, but they are not suitable for diagnosis of cancer. Some cancer markers in the blood have been used in the early diagnosis of cancer. Through the blood samples of physical examination, patients with early cancer can be found. There are four main types of cancer marker in body fluids: carcinoembryonic antigen (PSA [4], p53 [5], AFP [6], CEA [7]), enzymes (NSE [8]), hormones (in situ hormones, ectopic hormone HCG [9]), and glycoproteins. Most of them are secreted proteins which can appear in various body fluids, such as blood, saliva, urine, milk, sweat, etc. So far, most of the early marker studies focused on serum markers [10]. The main blood cancer markers provided by Chinese physical examination centers include: detection of alpha-fetoprotein (AFP), detection of carcinoembryonic antigen (CEA), ovarian cancer marker carcinoma(CA-125) [11], detection of breast cancer marker carcinoma (CA-153) [12], detection of non-small cell lung cancer and other cancers marker cytokeratin (CYFRA21-1) [13] and so on. The cancer markers are not lacking, however the sensitivity and specificity of a single marker is often not high enough to reach the clinical requirements. In theory and practice, simultaneous determination of multiple markers is advocated to improve sensitivity and specificity. Therefore, many researchers use machine learning algorithms to analyze human omics data, so as to find a combination of markers that can be used for specific cancer diagnosis. Around 2013, research on identifying cancer markers by computational methods reached its peak. A series of

biomarkers are selected by supervised or unsupervised computing methods based on transcriptome, proteomic data of genes and microRNAs. They can distinguish normal samples from cancer samples very efficiently. The algorithms include unsupervised dynamic hierarchical self-organization algorithm [14], a hybrid method combining genetic algorithms and SVM [15], binary state pattern clustering model [16], semi-supervised genetic learning model [17], general feature selection algorithm are based on linear support vector machine and reverse elimination method [18], knowledge-guided multi-scale independent component analysis method [19].

However, due to the strict conditions of blood collection, the collection process will bring pain to patients. These markers are not suitable for the observation of the development and prognosis of cancer, especially the follow-up observation after taking medicine. In recent years, saliva diagnosis has gradually attracted the attention of researchers and medical practitioners. Compared with serum samples, saliva collection is safe, convenient, non-invasive, without the risk of blood-borne disease transmission, painless and easy to accept. Compared with urine samples, saliva has the advantage of real-time sampling. Saliva detection has aroused great interest, and some preliminary results have been achieved. Many studies have supported that in the pathogenesis of systemic diseases, cancer tissues can make disease-specific changes in some salivary proteins through some mediators. These salivary proteins have the potential to become disease-specific biomarkers. How to effectively identify these potential markers is one of the challenging issues.

We insist that not all proteins are likely to be cancer markers in body fluids. So far, most of the cancer markers in body fluids are secreted proteins. This is related to the way they enter the body fluids. Therefore, we will search for cancer markers in body fluids in two steps. The first step is to identify candidate set of cancer biomarkers. The second step is to establish a model to predict which biomarkers in candidate set can enter the body fluid. Previous studies have shown that this method is effective. Cui et al. [20] first proposed the use of computational methods to predict proteins secreted by humans into the blood. They identified 85 features related to the process of protein secretion into the blood from 1521 protein features. These features are used to train the support vector machine, and good classification results are obtained. This study was subsequently applied to the identification of biomarkers for gastric cancer [21]. Firstly, differentially expressed genes in cancer tissues and adjacent tissues were identified by gene expression data. Then, the proteins encoded by these genes were classified by classifier. Finally, five proteins in blood were found as blood biomarkers for gastric cancer detection by mass spectrometry. Hong et al. [22] used a similar calculation method to predict the secreted proteins in urine. They found 18 valid features, combined with gene expression data, and obtained six candidate proteins for urinary biomarkers of gastric cancer, five of which were detected by Western blot. Wang et al. [23] predicted the source of salivary protein, and combined with the gene expression data of breast cancer, 31 candidate salivary protein markers were predicted. We proposed a framework for recognition of salivary secretory proteins [24], and achieved good results in the identification of markers of head and neck squamous cell carcinoma.

The main difficulties of these studies are as follows: Firstly, positive samples in training set can be obtained through literature collection, while negative samples are relatively difficult to select. Whether the negative samples are representative or not affects the recognition accuracy of the model. Secondly, the proteins in training set are also different. Although these proteins can be detected in the body fluid, the mechanism of their entry into the body fluid is not identical. They may be distributed far from each other in the classification space, which further restricts the accuracy of model prediction.

In order to solve these problems, an improved recognition framework is proposed in this paper.

Firstly, the essential information of human proteins, including UniProt id, sequence information, structure information, physical and chemical properties, are widely collected as the features to classify proteins. Then, through literature search, the secretory proteins present in saliva are collected as positive samples in the training set. In order to overcome the class distribution imbalance problem, an improved support vector clustering method (SVC-KM) is used to help select non-salivary secretory protein samples. Finally, we use feature selection method based on local samples to select more effective features for classifier training. The new framework can solve the problem of unbalanced sample size and uneven distribution in training set in order to improve the accuracy of recognition.

The rest of this paper is organized as follows: Section 2 presents our proposed method for salivary secretory proteins recognition. The section starts with an overview of the framework followed by the training set construction method using SVC-KM. A feature selection method based on local samples integrating SVM classifier is discussed, which makes two-class decisions for salivary secretory proteins recognition. Section 3 introduces the collection procedure of proteins and their properties. Section 4 discusses our experimental results. Section 5 concludes this paper with a summary of our work.

2. Salivary secretory proteins recognition model

2.1. Framework of the proposed method

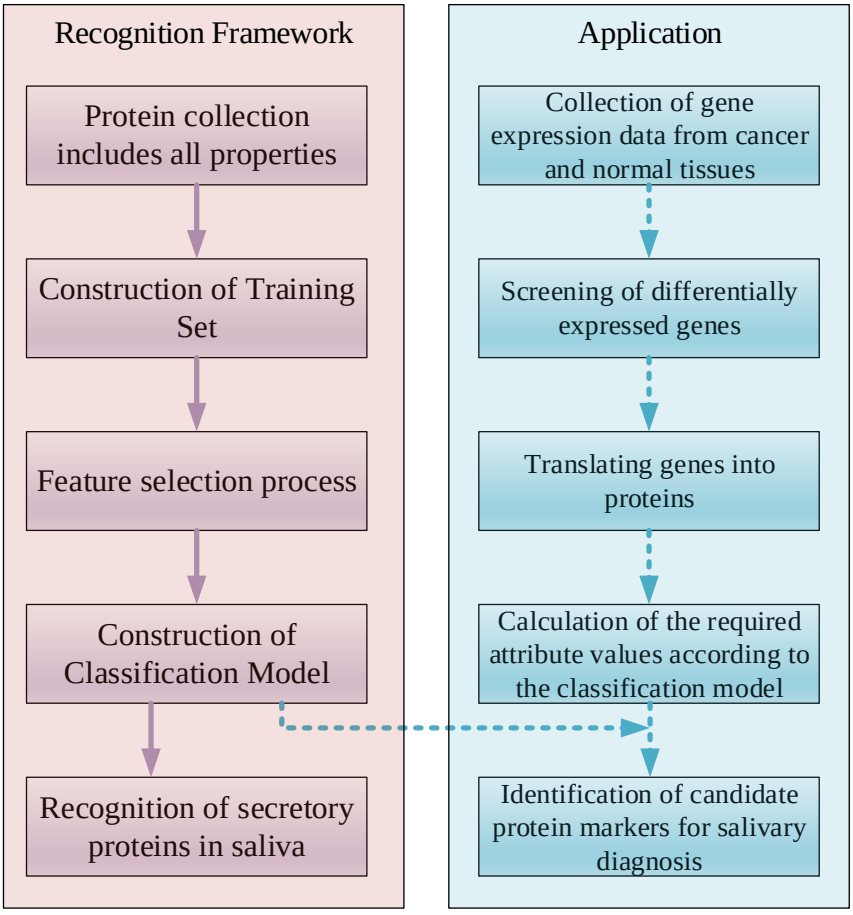

Figure 1. Flow chart of recognition and application of salivary secretory proteins.

Recognition of cancer biomarkers in saliva can be divided into two steps. The first step is to identify candidate set of cancer biomarkers. The second step is to establish a model to predict which biomarkers in candidate set can appear in saliva. The first step has been fully studied, and a large number of literatures have reported the identification methods of cancer biomarkers. While the second step is rarely reported. Therefore, our research focuses on the implementation of the second step. The framework of our proposed method and its application are shown in Figure 1.

As far as we know, most salivary protein markers enter saliva through autocrine and paracrine. Therefore, when construct the model to predict which biomarkers in candidate set can appear in saliva, we use the information of salivary proteome and secretory proteome to easily get the positive set of salivary secretory proteins. However, it is difficult to collect negative samples, owing to we are not sure which proteins must not appear in saliva. In the next, we will introduce our strategy to collect negative set samples, and the improvement of the feature selection method to solve uneven distribution in training set.

2.2. Clustering method for construction of negative sample set

As mentioned above, it is very difficult to collect negative samples. There are two common solutions. One is to construct a one-class classifier so that no negative samples need to be collected. The other is to try to collect more reasonable negative samples. The one-class classifier often requires a very large positive sample size, while secreted proteins have been reported to appear in saliva are obviously insufficient. Therefore, we attempt to solve this problem by the second way.

In previous studies [24], we excluded proteins belonging to the same family as salivary secretory proteins from human proteins. The remaining proteins were randomly selected to form a negative sample set. This brings uncertainty to the model. It cannot guarantee that the selected proteins can well describe the distribution of negative sample sets. In this paper, we propose a clustering method to guide the selection of proteins in negative sample sets.

In 2001, Ben-Hur et al. [25] proposed a kernel-based unsupervised clustering method, Support Vector Clustering (SVC). In this method, the sample points X are mapped from the sample space to the high-dimensional feature space by using the non-linear function Φ. In this feature space, a minimum hypersphere is found which can envelop all the sample points. The formulization is as the following equation:

$$\min R^2 \qquad s.t. \quad \left\|\Phi(\mathbf{x}_j) - \mathbf{a}\right\|^2 \le R^2 + \xi_j, \quad j = 1,\ldots,N \tag{1}$$

where $\Phi(\mathrm{x}_j)$ is the image of the sample points in the feature space, $\|\cdot\|^2$ is the Euclidean distance, $\mathbf{a}$ is the center of the hypersphere, R is the radius of the minimum hypersphere, and ξ_j is the relaxation variable. By transforming the objective function into a quadratic programming problem, the Eq.2 can be obtained:

$$\begin{cases} \max\limits_{\alpha} \quad W = \sum\limits_{j} \alpha_j \left\langle \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_j) \right\rangle - \sum\limits_{i,j} \alpha_i \alpha_j \left\langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \right\rangle \\ \qquad s.t.\ 0 \le \alpha_j \le C, \sum\limits_{j} \alpha_j = 1,\ j = 1, ..., N \end{cases} \tag{2}$$

where α_i is a lagrange operator and C is a constant. Here, the inner product of the sample point images is represented by the Gauss kernel function:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \right\rangle = e^{(-q\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2)} \tag{3}$$

where q is the kernel width. The distance from the image of the sample to the center of the hypersphere in the feature space can be expressed as:

$$R^2(\mathbf{x}) = \left\|\Phi(\mathbf{x}) - \mathbf{a}\right\|^2 = K(\mathbf{x}, \mathbf{x}) - 2\sum_{j} \alpha_j K(\mathbf{x}_j, \mathbf{x}) + \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{4}$$

The quadratic programming problem solved by the original SVC algorithm can theoretically obtain the global optimal solution. But it runs very slowly.

Therefore, we use SVC-KM algorithm to cluster proteins. The specific steps are as follows:

(1) A sample set containing N sample points is set up, and the sample points are mapped into the high-dimensional space by using the non-linear transformation to find the smallest hypersphere that can envelop almost sample points.

(2) The proteins outside the hypersphere are eliminated at this stage. K-means algorithm is used to classify the remaining proteins and adjust the value of k to obtain the optimal classification results. For previously eliminated proteins, assign them to the nearest cluster label.

We extract proteins proportionally from the optimal classification results. Because the number of samples in each cluster is different, we choose samples with a certain probability according to formula 5, which is related to the number of samples in the cluster.

$$c_j = \left[N_{pos} * \frac{m_j}{N_{negc}} \right] \quad j = 1, ..., K \tag{5}$$

where N_{pos} represents the number of proteins in the salivary secretion protein set, m_j represents the number of samples contained in the jth cluster, and N_{negc} represents the candidate set of non salivary secretory proteins. There are K clusters in total. Finally, the number of proteins selected from the jth cluster is c_j. The number of proteins selected in this way will be very close to the number of positive samples.

2.3. Feature extraction and selection

Feature selection method is a very important method in the field of machine learning. By selecting feature subset, we can improve the accuracy of the model, reduce the complexity of the model and the running time. Traditional methods are based on the samples in the whole training set to extract features, rarely considering the impact of abnormal samples and sample distribution. Among the problems to be solved in this paper, the unbalanced distribution of samples is a very prominent problem. The author proposes a method to improve the feature selection effect of filter by sample localization. For each test

sample, only the nearest sample is used for information feature selection. By using localized samples, the influence of abnormal samples and sample distribution on feature selection results can be effectively overcome. As an example, the flow chart of the algorithm is shown in Fig. 2 to find the nearest neighbor of a test sample in the positive sample set.

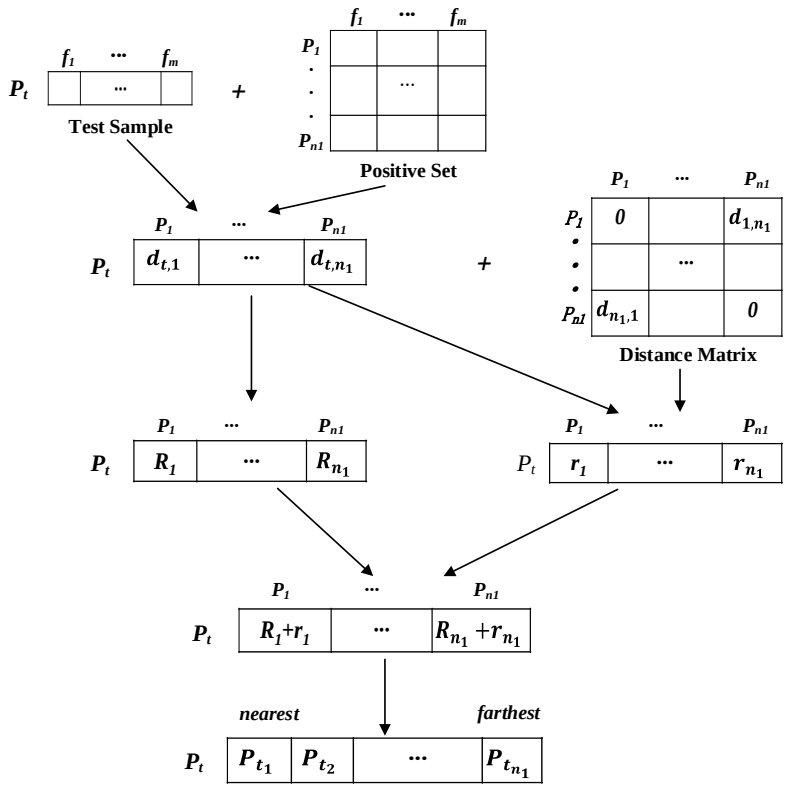

Figure 2. Flow chart of local samples acquisition method for a test sample in positive set.

In Figure 2, P_t represents a protein sample from the test set, P_j ($j = 1, \ldots, n1$) represents a protein sample from the positive set, f_i represents a feature of the protein, and. By formula 6, the Euclidean distance between the test sample and the training sample can be calculated.

$$dist(P_t, P_j) = \sqrt{\sum_{i=1}^{m} (P_{ti} - P_{ji})^2} \tag{6}$$

After obtaining the *dist* between the test sample and each training sample, we use an ascending numerical sort, stored in R_j. Similarly, for each training sample P_j, we can also get the ranking of P_t in P_j neighbors, stored in r_j. Finally, we can get the comprehensive ranking according to the sum of the two rankings.

According to above process, for each test sample, k nearest neighbors can be selected from each category to form the final training set. Generally, we use proportional parameter ratio instead of k. By adjusting ratio, we can get different number of local samples. After neighbors obtained, we use SVM method to classify the samples.

3. Data preparation

3.1. Protein collection

In order to collect salivary secretory proteins, we search for secretory proteins and salivary proteins in public databases and published literature respectively. The source and the number of corresponding proteins is detailed in Table 1. The two collections of proteins have 557 in common, which are used to form the positive sample set.

Table 1. Sources and quantities of proteins collected [24].

Secretory proteins	Num.	Proteins in saliva	Num.
SPD [26]	2194	Sys-body Fluid [29]	2161
LOCATE [27]	3376	Hu et al. [30]	331
UniProt [28]	1847	Denny et al. [31]	1166
Elimination duplicates	4312	Elimination duplicates	1987

Since some databases are no longer maintained, the data in this table from our earlier paper [24]. At present, there is no clear report on non-salivary secretory proteins. Therefore, we use protein family information to construct negative sample set by exclusion method. Protein family domain database (Pfam) [32] is a widely used protein family database. In this database, proteins and protein families are many-to-many relationships, that is, a protein may belong to one or more protein families, a protein family contains several to hundreds of different proteins. How to select excellent representatives of non-salivary secretory proteins is a challenging problem. The selection method proposed in this paper is shown in Figure 3. We obtained all human proteins from the Uniprot database and their family information in the Pfam database, and screened them according to the following steps: (1) excluding salivary secretory proteins (dataset 1) from the human protein collection; (2) excluding all proteins (dataset 2) contained in the salivary secretory proteins family; (3) further excluding the proteins (dataset 3) in the protein families, which the dataset 2 proteins are belong to, that are not excluded before; (4) clustering the remain proteins by SVC-KM based on the collected features; (5) selecting non-salivary secretory proteins from each cluster by ratio.

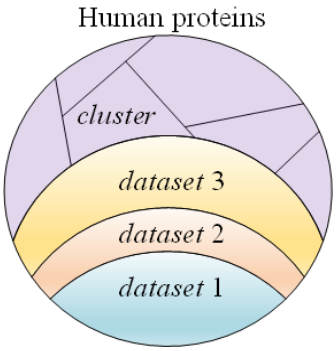

Figure 3. Sketch of negative sample set selection strategy.

3.2. Computation of protein properties

We convert protein properties into data array representation, and each property may correspond to multiple values. Detailed properties list, dimension corresponding to properties and acquisition tools are detailed in Table 2.

Table 2. List of proteins properties and tools.

Type	Name of the properties（Num. of the features）	Tools
General sequence features	Sequence length (1),	Fldbin
	Amino acid composition (20),	Profeat
	Di-peptides composition (400),	
	Normalized Moreau-Broto autocorrelation (90),	
	Moran autocorrelation (90),	
	Geary autocorrelation (90),	
	Sequence order (160),	
	Pseudo amino acid composition (80)	
Physicochemical properties	Hydrophobicity (21),	Profeat
	Normalized Van der Waals volume (21),	
	Polarity (21),	
	Polarizability (21),	
	Charge (21),	
	Secondary structure (21),	
	Solvent accessibility (21),	
	Unfoldability (1),	Fldbin
	Fldbin charge (1),	
	Longest disordered regions (1),	
	Isoelectric point (1),	
	Molecular weight (1),	
Domains/Motifs	Log P BBTM/Non-BBTM protein ratio (1),	
	Twin-arginine signal peptide (1),	
	Transmembrane domains (1),	
	Singal peptide (1)	

These properties can be classified into four types, including (1) sequence characteristic information, such as amino acid composition information; (2) domain properties, such as secondary structure information of proteins; (3) independent folding units and functional domains within the tertiary structure of proteins, such as transmembrane domain and signal peptide sequence; (4) physicochemical properties, such as polarity and hydrophobicity. There are 24 properties mentioned in Table 2, and after converting them into array representation by computational method, there are a total of 1087 features.

4. Experiments results and discussion

4.1. Negative sample set

The UniProt is the central for the collection of functional information on proteins, with accurate, consistent and rich annotation. In the latest version, there are a total of 20379 human proteins in Swiss-Prot, which means that all records are with information extracted from literature and curator-evaluated computational analysis. As mentioned in section 3.2, 557 proteins were collected as positive samples in the training set. After the exclusion method, there are 4062 proteins left. Then, we use SVC-KM method to cluster these proteins into 24 clusters. From each cluster, we select about one-seventh of the samples size nearest to the center of the cluster and finally there are 579 proteins to form the negative sample set.

4.2. *Feature selection results based on local samples*

Direct use of all 1087 features of proteins to train classifiers is not very effective, because there are many problem-independent features and noises, which will affect the effectiveness of classifiers. Therefore, these feature elements are filtered by *t-test* method. The threshold of significance level is *p-value* <= 0.05, and 486 of 1087 feature are selected. Then, we use feature selection method based on local samples to select these 486 features. In order to ensure the number of local samples, the proportion ratio of local samples is defined as 1/4. Then, combined with *t-test*, all features are ranked according to the *p* value of the features. Finally, SVM is used to evaluate the results of feature selection. When the average accuracy reaches the maximum, the feature set is used to construct the final recognition model. Accuracy, sensitivity, and specificity are used to evaluate the effect of the model. The formulas are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{9}$$

while, *TP* represents the number of the correctly recognized positive samples, *TN* represents the number of the correctly recognized negative samples, *FP* represents the number of negative samples recognized as positive samples, and *FN* represents the number of positive samples recognized as negative samples.

For different feature combinations, we mainly use accuracy as the measurement standard. This process is based on the idea of SVM-RFE: we remove one feature from the feature set at a time, if and only if the removed feature set obtains the highest average accuracy. When the number of features reaches dozens, the average accuracy no longer changes significantly. When the number of features is less than 32, the average accuracy begins to decline. Therefore, we select 32 features to form the feature set of the model. When the accuracy of classifier reaches the maximum during the feature selection procedure based on local samples, there are 32 features as shown in Table 3.

We train the SVM classifier with 1087 features, 486 features and 32 features respectively. 100 times of 10-fold cross validation to test the effectiveness of classifier. The average sensitivity, specificity, and accuracy are shown in Table 4.

Table 3. List of selected protein features and properties.

Features (Num. of Dimensions)
Amino acid composition (2)
transmembrane domain (1)
Di-peptides composition (6)
Moran autocorrelation (3)
Sequence order (4)
Secondary structure (3)
normalized Van der Waals volume (1)
Pseudo amino acid composition (5)
Polarizability (3)
Signal peptide (1)
Polarity (1)
Solvent accessibility (2)

Table 4. Performance evaluation of classifier using different features.

No. of features	Average *Sensitivity* (%)	Average *Specificity* (%)	Average *Accuracy* (%)
1087	64.73	72.71	68.82
486	88.73	93.61	91.23
32	97.09	98.10	97.61

The experimental results show that the average sensitivity, specificity and accuracy of salivary secretory protein recognition using selected features have been improved. These features provide useful information, including amino acid composition, transmembrane domain, Di-peptides composition, Moran autocorrelation, sequence order, secondary structure, normalized Van der Waals volume, pseudo amino acid composition, polarizability, signal peptide, polarity, solvent accessibility. Among them, transmembrane domain is one of the most important features for recognizing secreted proteins. Most of the proteins secreted through endoplasmic reticulum contain signal peptides. According to the information of signal peptides, these proteins are transported (transferred) to different places.

4.3. Prediction of cancer biomarkers by proteins

In practice, the user can input the ID of one protein (or one group of proteins) into the model, and the model will show whether this protein (or a group of proteins) can be a salivary secreted protein or not. If this protein is a salivary secreted protein, this means it can be used as a candidate cancer protein marker for saliva diagnosis, which needs to be further confirmed by biological experiments.

we collected 58 known cancer protein markers (shown as Table 5), and identified 42 of them through the model, which may be used for salivary diagnosis. The 42 salivary secretion proteins are shown in Table 5 with the bold font. These proteins will be the first choice for biologists to test.

Table 5. List of known cancer biomarkers.

No.	UniProt ID	Protein Name	Disease Name
1	P01033	TIMP-1	cancer; cardiovascular diseases; diabetes
2	P54108	CRISP-3	Sjogren's syndrome
3	P02766	transthyretin (TTR)	familial amyloidotic polyneuropathy（FAP）
4	Q12794	Hyaluronidase (HAse)	head and neck squamous cell carcinoma (HNSCC)
5	Q8WWA0	lactoferrin	Sjogren's syndrome
6	Q01469	epidermal fatty acid-binding protein	rheumatoid arthritis
7	P01034	Cystatin-C	Sjogren's syndrome
8	P16562	Cysteine-rich secretory protein 2	Sjogren's syndrome
9	Q2M3T9	Hyaluronidase-4	head and neck squamous cell carcinoma (HNSCC)
10	P54107	Cysteine-rich secretory protein 1	Sjogren's syndrome
11	Q12891	Hyaluronidase-2	Sjogren's syndrome
12	P05305	Endothelin-1(ET-1)	oral lichen planus or oral cancer in remission
13	O43820	Hyaluronidase-3	Sjogren's syndrome
14	P61626	Lysozyme C	Sjogren's syndrome
15	Q02747	Guanylin	pleomorphic adenoma warthin tumors
16	P05231	Interleukin-6	Sjogren's syndrome
17	Q16661	Guanylate cyclase activator 2B	pleomorphic adenoma warthin tumors
18	P01037	cystatin SA-I	oral squamous cell carcinoma
19	P43080	Guanylyl cyclase-activating protein 1	pleomorphic adenoma warthin tumors
20	Q9HC47	Cutaneous T-cell lymphoma-associated antigen 1	lymphomas
21	P21217	Le^b antigen	pancreatic cancer
22	P09228	salivary cystatin S	discriminate plaque resistant, periodontitis
23	P01036	salivary cystatin S	discriminate plaque resistant, periodontitis
24	P04637	Cellular tumor antigen p53	head and neck squamous cell carcinomas
25	P31947	14-3-3 protein	rheumatoid arthritis
26	P05109	Calgranulin-A	rheumatoid arthritis
27	P29622	Kallikrein	Sjogren's syndrome
28	P60568	Interleukin-2	Sjögren's syndrome
29	P12104	Fatty acid-binding protein	rheumatoid arthritis
30	P63104	14-3-3 protein zeta/delta	rheumatoid arthritis
31	P02144	Myoglobin	acute myocardial infarction (AMI)
32	P52209	6-phosphogluconate dehydrogenase	rheumatoid arthritis

Continued on next page

No.	UniProt ID	Protein Name	Disease Name
33	**Q04917**	**14-3-3 protein eta**	**rheumatoid arthritis**
34	**P61981**	**14-3-3 protein gamma**	**rheumatoid arthritis**
35	**P10645**	**Chromogranin-A**	**sleep bruxism**
36	**P06731**	**Carcinoembryonic antigen (CEA)**	**Colorectal cancer**
37	**P01266**	**Thyroglobulin (Tg)**	**Papillary and follicular thyroid cancer**
38	**P01137**	**TGFβ**	**Malignant tumors**
39	**P07288**	**Prostate specific antigen**	**Prostate cancer**
40	**P02771**	**Alpha-foetoprotein (AFP)**	**Hepatocellular carcinomas (HCC)**
41	**P15941**	**Cancer antigen 15-3 (CA15-3)**	**Breast cancer**
42	**Q969X2**	**Cancer antigen 19-9 (CA 19-9)**	**Pancreatic cancer; Bladder cancer**
43	P38398	BRCA-1	Breast cancer
44	P51587	BRCA-2	Breast cancer
45	Q8WXI7	Mucin-16	pelvic masses; malignant pelvic tumors; malignant ovarian tumors; epithelial ovarian cancer
46	P30044	Peroxiredoxin-5	rheumatoid arthritis
47	P01583	Interleukin-1 alpha	periodontal disease
48	O15511	Actin-related protein 2/3 complex subunit 5	upper-aerodigestive-tract cancer
49	Q10981	Galactoside 2-alpha-L-fucosyltransferase 2	pancreatic cancer
50	P01584	Interleukin-1 beta	periodontal disease
51	P02647	Apolipoprotein A-I	rheumatoid arthritis
52	P22894	Matrix metalloproteinase-8	oral disease, rheumatoid arthritis (RA)
53	P62258	14-3-3 protein epsilon	rheumatoid arthritis
54	P10415	Apoptosis regulator Bcl-2	Burkitt's lymphoma
55	P25054	APC gene	Adenocarcinoma, squamous cell carcinoma of the stomach, pancreas, thyroid and ovary
56	P07339	Cathepsin D	breast cancer
57	P31151	psoriasin	pulmonary involvement in systemic sclerosis
58	P80511	Protein S100-A12	rheumatoid arthritis

4.4. Prediction of cancer biomarkers by gene expression data

If a user has cancer and control gene expression data samples. He should recognize differentially expressed genes first, then translate differentially expressed genes into proteins. Put the features of these proteins into our model to get the prediction results.

In order to acquire the cancer and control set of gene expression data used in this subject, we select the data collection and download it in the GEO gene expression database. The quality control of the gene expression data set of the needed tumor is carried out. Try to select the sample in the data set was

labeled and clearly divided data sets as experimental research samples, and the Matrix Data file in the data set is downloaded to prepare for the upcoming work. After downloading the required tumor gene expression data, manual classification is performed according to the sample of health and cancer in order to complete the subsequent experimental work. Ordinary preprocessing involves the following stages: if the difference between the maximum and minimum values is less than a specific value or the difference multiple (maximum/minimum value) is less than a specific value, then filtering excludes genes from the data set, where the maximum and minimum values are for a particular gene. Highest and lowest expression.

Perform a 2-based logarithmic transformation of each gene expression value; data normalization is used to eliminate systematic differences between samples. Due to the characteristics of high-dimensional small samples of gene expression data, the data distribution will have a large standard deviation, and the data needs to be normalized by logarithmic transformation. Can have a great impact on subsequent experimental research.

We use T-test and restrain fold change to remove the irrelevant genes and redundant genes from each dataset. Organize the genes in the new data set according to the level of gene expression. Then mapping the heat map to analyze the gene expression levels of different genes in tumor samples. The gene expression level Heat maps with classification results are shown as Figure 4.

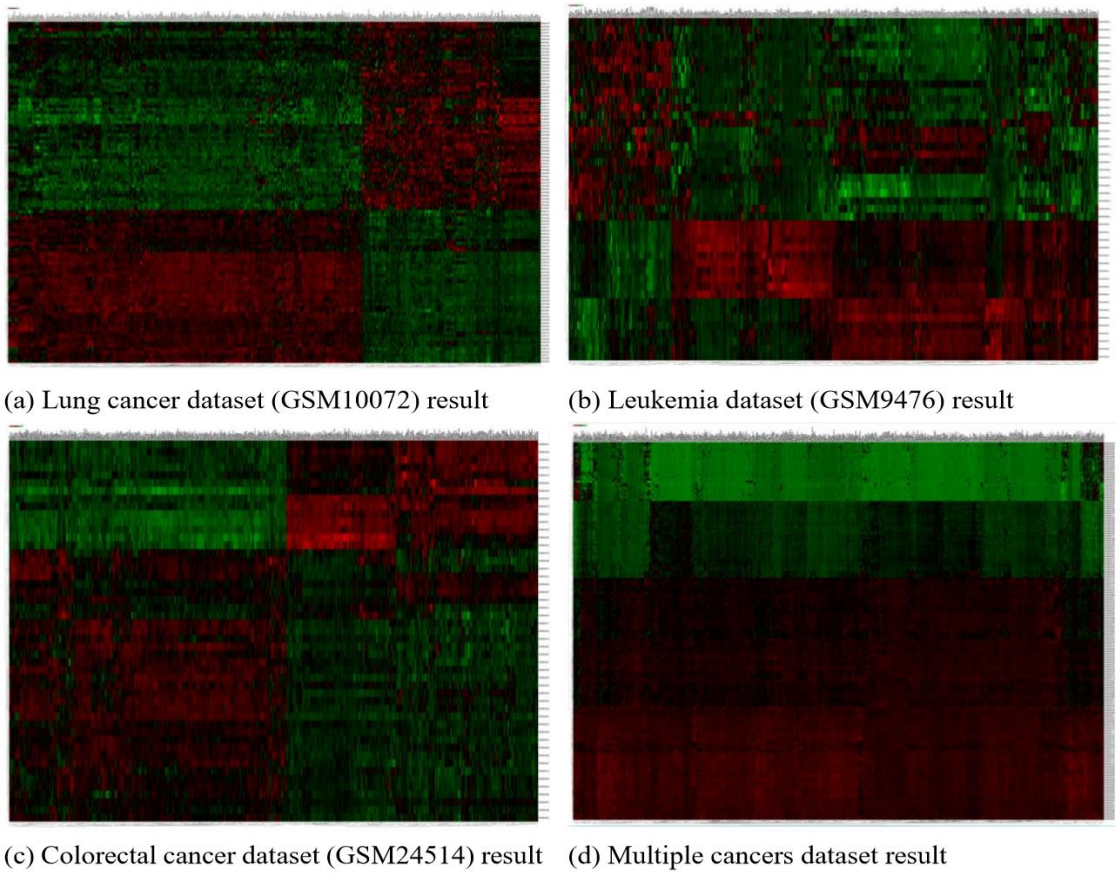

Figure 4. Sketch of negative sample set selection strategy.

For each dataset, we choose the top 20 differentially expressed genes. The selected gene symbols are list in Table 6. Using these genes to classify the healthy and cancer samples, the classification

accuracy can reach 99.07% for lung cancer, 100% for Leukemia, 93.88% for colorectal cancer, 98.81% for multiple cancers, respectively. Then, we use our model to identify the salivary secreted proteins as candidate cancer biomarkers, which are showed in bold in Table 6. There are total 80 genes from four datasets. After removing the repeat value, 75 different genes are left. For gene OR7E47P, we didn't find the protein information. By our model, 33 of them are identified as salivary secretion proteins.

Table 6. List of Selected feature gene symbol for each dataset.

NO.	Gene symbol			
	Lung cancer	**Leukemia**	**Colorectal cancer**	**Multiple cancers**
1	**CD36**	**KCNH2**	**GUCA2A**	**ARHGAP6**
2	FHL1	MYH10	PKM2	**CD93**
3	JAM2	GMPR	ALPI	**ABCA3**
4	**CLEC3B**	RTN1	GTF2IRD1	**HEG1**
5	AGTR1	AURKA	MCM7	**KCNK3**
6	LDB2	AHR	**UBE2C**	GPM6A
7	LMO2	PBK	MTHFD2	GPM6B
8	S1PR1	RFC4	SYNM	**NTRK2**
9	SASH1	ALDH1A1	C20orf20	ADH1B
10	**FIGF**	PTGS2	RRM2	**RAMP2**
11	PGC	TSPYL5	FANCG	**OLFML2A**
12	**GIMAP6**	**PLXNC1**	**SLCO4A1**	**LRRC32**
13	**DNASE1L3**	TUBG1	BUB1B	THBD
14	**C14orf132**	**SLC15A2**	**BACE2**	LDB2
15	HOXA5	SPAG5	VSNL1	PDLIM2
16	RRM2	**PRTN3**	EIF4EBP1	**BCHE**
17	**AGR2**	FOXM1	KPNA2	OR7E47P
18	FOXF1	REXO2	**CLDN8**	TNNC1
19	**THBD**	**RHCE**	**CXCL9**	**EMCN**
20	**CACNA2D2**	**TFR2**	FHL1	**CACNA2D2**

5. Conclusion

In this work, methods based on machine learning are proposed to identify cancer biomarkers that can appear in saliva. We have improved the existing methods in two respects： To build a training set, SVC-KM method is used to help select non-salivary secretory protein samples based on protein family information. The feature selection method based on local samples is used for feature selection. Using these machine learning algorithms to get the features for training and classification, it is expected that a higher rate of salivary secretory proteins recognition can be achieved. Furthermore, we have predicted some proteins and genes in saliva, which can provide reference for the research of biology and medicine workers.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61872418), Natural

Science Foundation of Jilin Province (20180101050JC), Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (NO. CAAC-ITRB-201603).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. R. Ruddon, Cancer Biology, *Oxford University Press*, 2007.
2. Y. Wang, S. Liang, Y. Tian, J. Zhao, W. Du, Y. Liang, et al., Using machine learning to measure relatedness between genes: a multi-features model, *Sci. Rep.*, **9** (2019), 1–15.
3. S. Liang, A. Ma, S. Yang, Y. Wang, Q. Ma, A review of Matched-pairs feature selection methods for gene expression data analysis, *Comput. Structur. Biotechnol. J.*, **16** (2018), 88–97.
4. A.W. Partin, J. Yoo, H. B. Carter, J. D. Pearson, D. W. Chan, J. I. Epstein, et al., The use of prostate specific antigen, clinical stage and Gleason score to predict pathological stage in men with localized prostate cancer, *J. Urol.*, **150** (1993), 110–114.
5. M. Hollstein, D. Sidransky, B. Vogelstein, C. C. Harris, P53 mutations in human cancers, *J. Sci.*, **253** (1991), 49–53.
6. K. E. Stuart, A. J. Anand, R. L. Jenkins, Hepatocellular carcinoma in the United States: prognostic features, treatment outcome, and survival, *Cancer Interdiscipl. Int. J. Am. Cancer Soc.*, **77** (1996), 2217–2222.
7. P. Kuusela, C. Haglund, P. J. Roberts, Comparison of a new tumour marker CA 242 with CA 199, CA 50 and carcinoembryonic antigen (CEA) in digestive tract diseases, *British J. Cancer*, **63** (1991), 636–640.
8. J. Schneider, H. G. Velcovsky, H. Morr, N. Katz, K. Neu, E. Eigenbrodt, Comparison of the tumor markers tumor M2-PK, CEA, CYFRA 21-1, NSE and SCC in the diagnosis of lung cancer, *Anticancer Res.*, **20** (2000), 5053–5058.
9. L. A. Cole, J. M. Sutton, Selecting an appropriate hCG test for managing gestational trophoblastic disease and cancer, *J. Reproduct. Med.*, **49** (2004), 545–553.
10. J. A. Ludwig, J. N. Weinstein, Biomarkers in cancer staging, prognosis and treatment selection, *Nat. Rev. Cancer*, **5**(2005), 845–856.
11. G. J. Rustin, M. Marples, A. E. Nelstrop, M. Mahmoudi, T. Meyer, Use of CA-125 to define progression of ovarian cancer in patients with persistently elevated levels, *J. Clin. Oncol.*, **19** (2001), 4054–4057.
12. H. Zheng, R. C. Luo, Diagnostic value of combined detection of TPS, CA153 and CEA in breast cancer, *J. First Milit. Med. Univers.*, **25** (2003), 1293.
13. H. Q. Zhang, R. B.Wang, H. J. Yan, W. Zhao, K. L. Zhu, S. M. Jiang, et al., Prognostic significance of CYFRA21-1, CEA and hemoglobin in patients with esophageal squamous cancer undergoing concurrent chemoradiotherapy, *Asian Pacific J. Cancer Prevent.*, **13** (2012), 199–203.
14. A. Hsu, S. L. Tang, S. Halgamuge, An unsupervised hierarchical dynamic self-organising approach to cancer class discovery and marker gene identification in microarray data, *Bioinformatics*, **19** (2003), 2131–2140.
15. J. J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, et al., Multiclass cancer classification and biomarker discovery using GA-based algorithms, *Bioinformatics*, **21** (2005), 2691–2697.

16. B. J. Beattie, P. N. Robinson, Binary state pattern clustering: A digital paradigm for class and biomarker discovery in gene microarray studies of cancer, *J. Comput. Biol*., **13** (2006), 1114–1130.
17. C. Harris, N. Ghaffari, Biomarker discovery across annotated and unannotated microarray datasets using semi-supervised learning, *BMC Genomics*, **9**(2008), S7.
18. T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, *Bioinformatics*, **26** (2010), 392–398.
19. L. Chen, J. Xuan, C. Wang, I. M. Shih, Y. Wang, Z. Zhang, et al., Knowledge-guided multi-scale independent component analysis for biomarker identification, *BMC Bioinformatics*, **9** (2008), 416.
20. J. Cui, Q. Liu, D. Puett, Y. Xu, Computational prediction of human proteins that can be secreted into the bloodstrea, *Bioinformatics*, **24** (2008), 2370–2375.
21. J. Cui, Y. Chen, W. C. Chou, L. Sun, L. Chen, J. Suo, et al., An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer, *Nucleic Acids Res.*, **39** (2011),1197–1207.
22. C. S. Hong, J. Cui, Z. Ni, Y. Su, D. Puett, F. Li, et al., A computational method for prediction of excretory proteins and application to identification of gastric cancer markers in urine, *PloS One*, **6** (2011), e16875.
23. J. Wang, Y. Liang, Y. Wang, J. Cui, M. Liu, W. Du, et al., Computational prediction of human salivary proteins from blood circulation and application to diagnostic biomarker identification, *PloS One*, **8** (2013), e80211.
24. Y. Sun, W. Du, C. Zhou, Y. Zhou, Z. Cao, Y. Tian, et al., A Computational Method for Prediction of Saliva-Secretory Proteins and its Application to Identification of Head and Neck Cancer Biomarkers for Salivary Diagnosis, *IEEE Transact. Nanobiosci.*, **14** (2015),167–174.
25. A. Ben-Hur, D. Horn, H. T. Siegelmann, V. Vapnik, A support vector method for clustering, *Adv. Neural Inform. Process. Syst.*, **13** (2001), 367–373.
26. Y. Chen, Y. Zhang, Y. Yin, G. Gao, S. Li, Y. Jiang, et al., SPD—a web-based secreted protein database, *Nucleic Acids Res.*, **33** (2005), D169–D173.
27. J. Sprenger, J. Lynn Fink, S. Karunaratne, K. Hanson, N. A. Hamilton, R. D. Teasdale, LOCATE: A mammalian protein subcellular localization database, *Nucleic Acids Res.*, **36** (2007), D230–D233.
28. M. Magrane, Uniprot knowledgebase: A hub of integrated protein data, *Database*, **2011** (2011).
29. S. J. Li, M. Peng, H. Li, B. S. Liu, C. Wang, J. R. Wu, et al., Sys-bodyfluid: A systematical database for human body fluid proteome research, *Nucleic Acids Res.*, **37** (2009), 907–912.
30. S. Hu, J. A. Loo, D. T. Wong, Human saliva proteome analysis and disease biomarker discovery, *Expert Rev. Proteom.*, **4** (2007), 531–538.
31. P. Denny, F. K. Hagen, M. Hardt, L. Liao, W. Yan, M. Arellanno, et al., The proteomes of human parotid and submandibular/sublingual gland salivas collected as the ductal secretions, *J. Proteom. Res.*, **7** (2008), 1994–2006.
32. S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, et al., The Pfam protein families database in 2019, *Nucleic Acids Res.*, **47** (2019), D427–D432.

AIMS Press