Computational Prediction of Human Salivary Proteins from Blood Circulation and Application to Diagnostic Biomarker Identification

Jiaxin Wang[1], Yanchun Liang[1], Yan Wang[1], Juan Cui[2], Ming Liu[1], Wei Du[1*], Ying Xu[1,2*]

1 Key Laboratory for Symbolic Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China, **2** Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, Georgia, United States of America

Abstract

Proteins can move from blood circulation into salivary glands through active transportation, passive diffusion or ultrafiltration, some of which are then released into saliva and hence can potentially serve as biomarkers for diseases if accurately identified. We present a novel computational method for predicting salivary proteins that come from circulation. The basis for the prediction is a set of physiochemical and sequence features we found to be discerning between human proteins known to be movable from circulation to saliva and proteins deemed to be not in saliva. A classifier was trained based on these features using a support-vector machine to predict protein secretion into saliva. The classifier achieved 88.56% average recall and 90.76% average precision in 10-fold cross-validation on the training data, indicating that the selected features are informative. Considering the possibility that our negative training data may not be highly reliable (i.e., proteins predicted to be not in saliva), we have also trained a ranking method, aiming to rank the known salivary proteins from circulation as the highest among the proteins in the general background, based on the same features. This prediction capability can be used to predict potential biomarker proteins for specific human diseases when coupled with the information of differentially expressed proteins in diseased *versus* healthy control tissues and a prediction capability for blood-secretory proteins. Using such integrated information, we predicted 31 candidate biomarker proteins in saliva for breast cancer.

Citation: Wang J, Liang Y, Wang Y, Cui J, Liu M, et al. (2013) Computational Prediction of Human Salivary Proteins from Blood Circulation and Application to Diagnostic Biomarker Identification. PLoS ONE 8(11): e80211. doi:10.1371/journal.pone.0080211

Editor: Luonan Chen, Chinese Academy of Sciences, China

Received August 5, 2013; **Accepted** September 29, 2013; **Published** November 12, 2013

Funding: This work is supported by the Natural Science Foundation of China (61272207), the Science-Technology Development Projects of Jilin Province of China (20120730, 20130522111JH, 20130522114JH), the Ph.D. Program Foundation of MOE of China (20120061110094, 20120061120106), and the Postdoctoral Science Foundation (2012M520678). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: weidu@jlu.edu.cn (WD); xyn@bmb.uga.edu (YX)

Introduction

Human blood has long been used as an information source for detection of human diseases such as liver enzymes for detecting hepatitis, white-blood cell counts for infection detection and prostate-specific antigen (PSA) for diagnosing prostate cancer. In comparison, human saliva has not been used for the same purposes nearly as much. Recent large-scale proteomic analyses have revealed that human saliva is also rich in proteins [1], some of which come from the blood circulation and hence can potentially serve as a general information pool for disease biomarker identification. This study is on the development of a computational method for identification of the distinct features of salivary proteins that come from circulation and an application of the identified features to predict proteins that can get into saliva from circulation.

The earliest work on using salivary proteins as disease biomarkers of distal organs can be traced back to 1986 when the Kallikreina salivary biomarkers for detection of breast cancer and gastrointestinal cancer were published [2]. Since then, a number of salivary proteins have been found to have elevated levels in patients of specific cancer types compared to the healthy population such as PSA for prostate cancer [3], c-tumor protein erbB-2 and p53 for breast cancer [4]. While a few salivary proteins have been found to be relevant to specific diseases, there has not been a general and effective approach for identifying disease markers in saliva, to the best of our knowledge.

The current understanding about how biomolecules can move from circulation into saliva can be summarized as

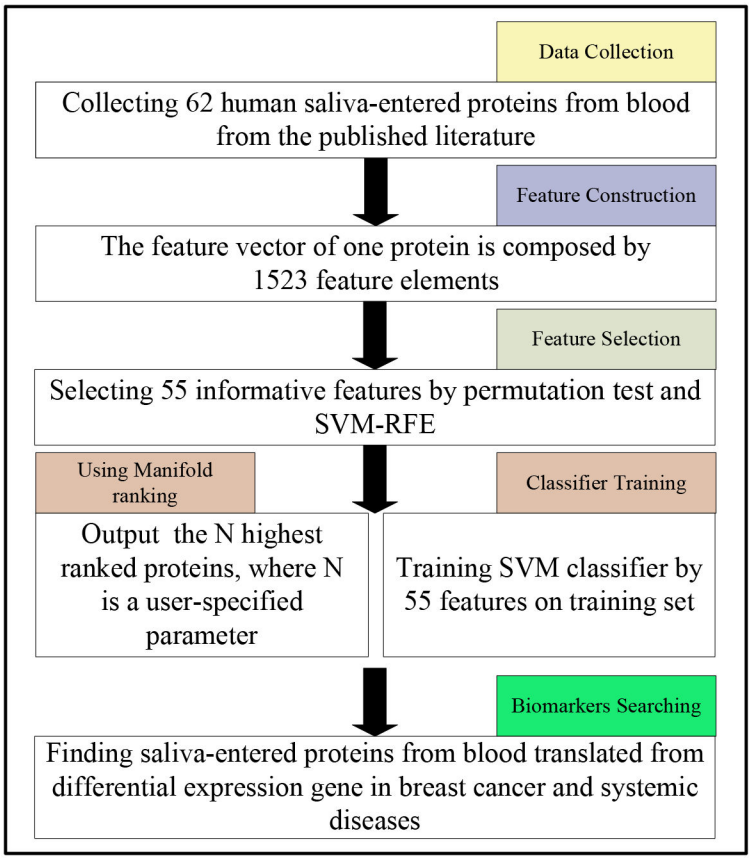

Figure 1. A flowchart of the approach.

doi: 10.1371/journal.pone.0080211.g001

follows. Three mechanisms have been identified for biomolecules to travel from circulation into saliva [5,6]: active transportation for various proteins such as secretory IgA and immunoglobulin E, passive transportation for drugs and steroids, and ultrafiltration for small polar molecules such as creatinine. The basis of our prediction method is that some of the disease-associated proteins in circulation can get into saliva through one of these three mechanisms, hence making it possible for us to identify them in saliva even for diseases of distal organs.

Two large datasets for salivary proteins are publicly available. One consists of 1,166 proteins and 657 of them are also found in human blood [1]. Another one has approximately 2,000 proteins and 26% of them are also found in blood [6]. We hypothesize that salivary proteins are secreted by the salivary glands either from circulation or in response to the biomolecules that get into the glands from circulation. In this study, we focus on proteins that come from the circulation and leave the prediction work of proteins secreted by salivary glands in response to blood proteins that get into the glands as a future study.

We have collected 62 human salivary proteins coming from circulation from the published literature, which have been experimentally detected by multiple salivary proteomic studies, and used them as the initial positive training data. We then expanded this dataset by including additional proteins based on Pfam family information [7]. A total of 261 proteins are selected at the end as the positive training data. We then identified a set of proteins that are deemed not to be able to get into saliva, totaling 6,816, and used them as the negative training data. We then examined a number of sequence and structure-based features to identify those with discerning power between the two sets of proteins. Using these features, we have trained a classifier using a support vector machine (SVM) to predict proteins that can travel to saliva from circulation via salivary glands. In addition, we have also trained a ranking method aiming to rank the known blood-originated salivary proteins the highest among the background proteins, knowing that our negative training data may not be the most reliable. The flowchart of the approach is shown in Figure 1.

We believe that this prediction capability can serve as a general tool for predicting proteins that can travel from circulation to saliva. Hence when applied in conjunction with capabilities for predicting proteins that may be present in circulation of patients of a specific disease, this capability can suggest candidate biomarkers in saliva for that disease. Using this tool along with gene-expression data of breast cancers and a prediction tool for blood-secretory proteins [8], we predicted 31 candidate proteins in breast cancer patients' saliva.

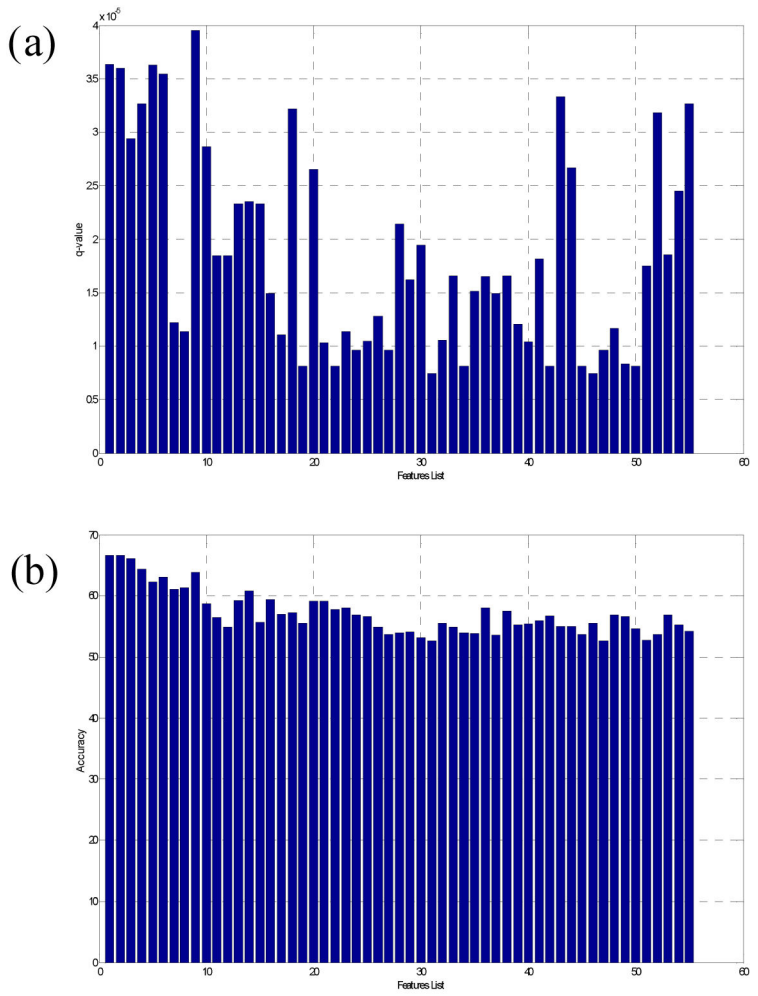

Figure 2. The q-value and accuracy of the 55 selected features.
doi: 10.1371/journal.pone.0080211.g002

Results

Features of blood-originated salivary proteins

With the aim of training a SVM-based classifier and rank the predicted proteins, we have examined a total of 34 protein features (see Table S1 and Material and Methods), represented as a feature vector of 1,523 dimension. We then trained a classifier with a linear kernel using these features calculated on proteins in both the positive and negative training sets, aiming to derive a classifier that can best distinguish the positive from the negative samples. We then checked which feature elements are relevant to the final classification performance by using a feature selection procedure, and removed all the irrelevant ones, giving rise to 55 final feature elements. Then a manifold ranking method [8] is trained based on the selected feature elements with the performance given in Table S2. We have assessed the contributions by the 55 feature elements to the classification accuracy, using a statistical significance q-value [9], and found that the q-values for the 55 feature elements are less than 4.0E-5, as shown in Figure 2(A). We have also compared the classification performance based on the 55 features *versus* the top 10 features, and noted that there is a clear difference in performance, as shown in Figure 2(B). The following features are the most important ones to our classification accuracy, ranked in the decreasing order of their contribution to the classification results: radius, Moran autocorrelation, hydrophobicity, Geary autocorrelation, amino acid composition, normalized Moreau-Broto autocorrelation, dipeptide composition, secondary structure composition and polarity. This observation is consistent with our general understanding of secretory proteins and salivary proteins. For example, the diffusion coefficient is inversely proportional to the molecular radius [10].

Performance of the SVM model

Based on the 55 selected feature elements, we trained a classifier and evaluated the performance using 10-fold cross validation by repeating the prediction 100 times to derive a performance distribution of the classifier. On the training data,

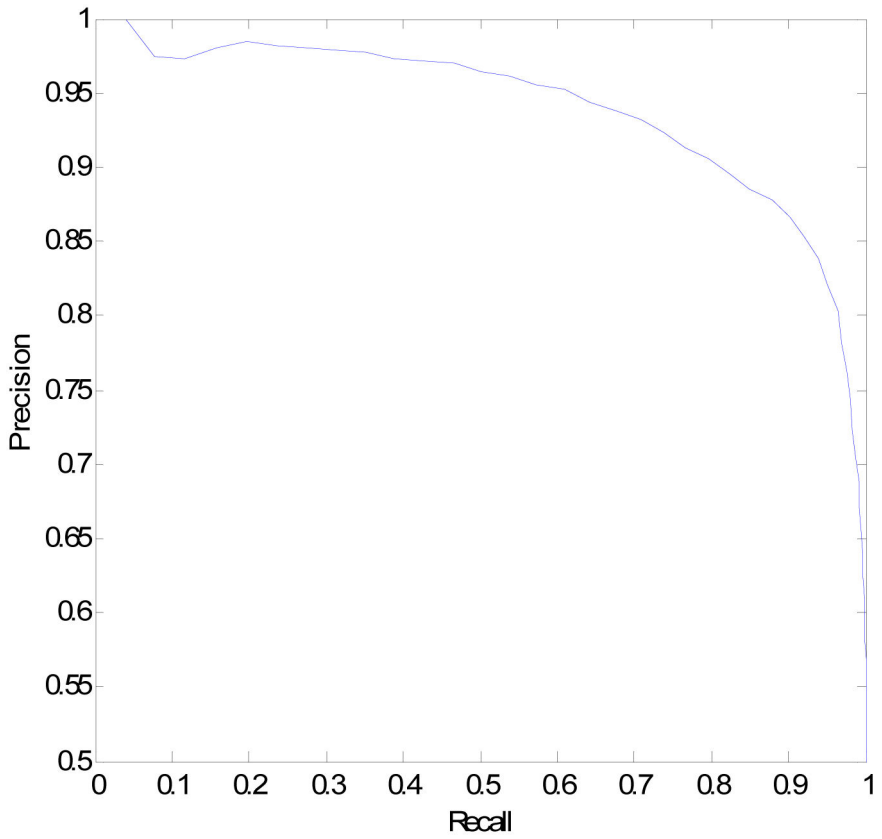

Figure 3. The recall-precision curve.

doi: 10.1371/journal.pone.0080211.g003

the classifier achieved an average recall and precision at 88.56% and 90.76%, respectively. We applied the general recall-precision curve shown in Figure 3 to the training data with 10-fold cross validation to examine the prediction precision at each recall level. The AUC of the recall-precision curve is 80.96%.

We also used 41 of the 62 collected proteins as the training data that have been reported in the literature before 2000. Out of the 62 collected proteins, 21 are used as the testing data, which have been reported after 2000. On these 21 salivary proteins, our model predicted 14 (76.19%) to be salivary proteins from blood.

Predicting and ranking the known salivary proteins

We have run the trained classifier on all 20,209 human proteins in the UniProt database [11], among which 5,456 are annotated as secretory proteins according to the Uniprot, SPD [12] and LOCATE [13] databases, and 1,823 have been detected in saliva from previous experiments [1,5,14]. We predicted 2,498 of the Uniprot proteins as salivary proteins from circulation, accounting for 12.36% of the 20,209 Uniprot proteins. Of the 2,498 proteins, 239 (13.11%) are among the 1,823 proteins that have been previously identified in saliva experimentally.

We have also ranked the human proteins in UniProt using a manifold ranking method as done in our previous work [8]. 62 known salivary proteins coming from circulation were assessed in terms of their ranking to be salivary proteins. By using the 62 proteins as positive dataset, 27 (43.55%) of the 62 proteins and 136 salivary proteins are ranked among the top 1,000 proteins. By using the expanding 261 proteins as positive dataset, 34 (54.84%) of the 62 proteins are ranked among the top 1,000 proteins. Among these 1,000 proteins, 155 are known to be salivary proteins identified from other sources (see Table S3 for the list of the protein names; and also see Material and Methods). While we do not know if the prediction of the remaining 845 proteins being salivary proteins is correct or not, we suspect that some of them are indeed salivary proteins. For example, protein Endothelin-1 (P05305), ranked the 728th, has been implicated in cancer[15], and could be a good salivary biomarker for OSCC development in oral lichen planus patients [16]. Tissue inhibitor of metalloproteinases 1 (TIMP-1) (P01033), ranked the 434th, has been identified as a potential biomarker in diseases such as cancer, cardiovascular diseases and diabetes. Moreover, this protein has been reported to be a salivary protein [17].

After the training of our classifier, we did another round of literature search for additional salivary proteins that have been associated with human diseases and do not overlap with our training data. Overall 47 salivary proteins are found, shown in Table S4. These proteins are relevant to different diseases such as periodontal disease [18,19], oral squamous cell

Table 1. Comparison of the ranking result with human saliva biomarkers for many sorts of diseases.

Total protein number	Known salivary biomarker number	Top number	Salivary biomarker included in top number	P-vlaue
20209	47	500	1	0.367
20209	47	1000	3	0.211
20209	47	1500	6	0.076
20209	47	2000	8	0.088
20209	47	2500	11	0.026
20209	47	3000	12	0.038
20209	47	3500	13	0.052
20209	47	4000	13	0.123
20209	47	4500	15	0.082
20209	47	5000	16	0.098
20209	47	5500	19	0.017
20209	47	6000	20	0.020

doi: 10.1371/journal.pone.0080211.t001

carcinoma [20,21], Sjögren's syndrome [22-24], breast cancer [25-29], malignant pelvic tumors, and malignant ovarian tumors [30]. We found that 3 (6.38%) of these 47 proteins are ranked among the top 1,000, 8 (17.02%) among the top 2,000 and 12 (25.53%) among the top 3,000, as shown in Table 1. The p-values for having such rankings if assuming that the ranking is random are 0.211, 0.088 and 0.038, respectively.

We then carried out a pathway enrichment analysis among the top 1,000 ranked proteins, using DAVID [31] against the Gene Ontology, KEGG [32], BBID [33] and BIOCARTA [34] databases to gain an understanding about the cellular functions and subcellular locations of these predicted salivary proteins, using the whole set of human proteins as the background. We noted that the most significantly enriched biological processes are immune response, antigen processing and presentation, cell adhesion, defense response, response to wounding, and inflammatory response. In addition, the most significantly enriched cellular components are extracellular region, membrane and MHC protein complex which all make biological sense (see Table S5).

Application to breast cancer for identification of salivary biomarkers

Based on a public transcriptomic dataset collected on breast cancer and matching control samples (see Materials and Methods), we identified 1,502 consistently differentially expressed genes in breast cancer *versus* control tissue samples. We then used the gene expression data as an approximate protein-expression data here; and applied our trained classifier to these proteins and predicted 248 of them to be blood secretory using a prediction tool for blood secretory proteins that we previously developed [8]. Out of these proteins, we predicted 31 are movable to saliva. Table 2 provides the detailed information of these 31 proteins as candidate salivary biomarkers for breast cancer.

Table 2. Proteins as candidate salivary biomarkers for breast cancer.

Gene symbol	UniProt ID	Manifold ranking	Fold change
F10	P00742	195	0.667
CFD	P00746	227	0.593
TIMP2	P16035	241	0.573
CCL14	Q16627	297	0.595
FBLN1	P23142	324	0.663
FBLN5	Q9UBX5	336	0.542
EFEMP2	O95967	363	0.622
IGF1	P05019	394	0.525
EFEMP1	Q12805	439	0.474
AZGP1	P25311	440	1.563
WISP2	O76076	613	0.581
CLEC3B	P05452	720	0.570
CD93	Q9NPY3	724	0.586
LEPR	P48357	1034	0.638
FABP5	Q01469	1072	0.633
IL6R	P08887	1111	0.551
ALCAM	Q13740	1115	1.764
MCAM	P43121	1119	0.527
CFB	P00751	1148	1.647
PDCD6	O75340	1382	1.562
BCHE	P06276	1416	0.665
DMBT1	Q9UGM3	1531	0.632
CD163	Q86VB7	1539	0.628
NCAM1	P13591	1681	0.640
LTF	P02788	1693	1.583
SRPX	P78539	1959	0.556
FBN1	P35555	2192	0.625
CFH	P08603	2400	0.550
VWF	P04275	2518	0.537
CD99	P14209	2867	0.623
TF	P02787	2907	0.578

doi: 10.1371/journal.pone.0080211.t002

As of now, very little data is available regarding salivary proteins that can be indicative of breast cancer. The only data we can get hold of is the salivary proteins considered by Streckfus et al. to be informative for diagnosing breast cancer [27]. Their predicted list consists of 37 proteins given in Table S6. We have compared our prediction of 31 proteins with this list, 4 of the 31 proteins are in their list [27-29], as shown in Table 3, which has a p-value at 2.89e-7. The relatively low level of overlap between the two sets of predictions is not particularly surprising, which is consistent with previously published studies by different groups on blood biomarkers for different cancers. This is possibly caused by the differences in detailed conditions under which the biological samples, i.e., cancer tissues and saliva, are collected, as well as the less-

Table 3. Prediction Proteins used as salivary biomarkers for the detection breast cancer.

Not included in the training positive dataset				
Accession	**Protein Name**	**Ratio**	**P-value**	**Blood Secretory**
Q01469	Epidermal fatty acid-binding protein	0.633	0.000257	Yes
P02788	Lactotransferrin	1.583	0.000244	Yes
Included in the training positive dataset				
Accession	**Protein Name**	**Ratio**	**P-value**	
P02787	Transferrin	0.578	0.000013	Yes
P25311	Zinc-alpha-2-glycoprotein	1.563	0.000940	Yes

doi: 10.1371/journal.pone.0080211.t003

than-perfect prediction methods employed, on top of the overall very challenging nature of the problem.

We have also carried out a pathway and subcellular location enrichment analysis similar to that in the above. We noted that the most enriched biological processes by these 31 proteins are response to wounding, acute inflammatory response, cell adhesion, biological adhesion and immune response, which are all known to be involved in the development of or in defense of cancer. Besides, the most enriched cellular locations are extracellular region and cell surface (Table S7). The most enriched pathways are complement and coagulation cascades, and the second enriched pathways are cell adhesion molecules (CAMs) (Table S8).

Discussion and Conclusion

A reliable prediction capability for proteins that can travel from circulation to saliva will represent a highly useful tool as it can provide a candidate list of biomarkers specific to a particular disease. This will allow targeted searches for effective biomarkers in saliva using antibody-based techniques, in comparison with the traditional search strategies by direct comparisons among proteomic data collected from saliva samples of multiple patients and healthy controls, which have proved to be ineffective in searches for biomarkers in blood [8,35] and urine [36]. Here we demonstrated that it is possible to develop one such tool, which by no means represents the possibly most reliable tool for such a prediction. The key contribution of work is the proof of principle that we can possibly identify distinguishing features between proteins that can move to saliva from circulation and proteins that cannot get into saliva. In addition the identified features can also provide useful information to the mechanism studies of how proteins move between blood and saliva. In the future study, we hope that our method could be used in conjunction with the technology platforms for saliva diagnostics, and identify the definitive disease-associated salivary biomarkers.

Materials and Methods

Collecting salivary proteins coming from blood and generating negative training data

There is no existing dataset about proteins that can move from circulation to saliva. Proteins that have been found in both salivary proteome and blood proteome cannot serve this purpose since some of the salivary proteins may not come from circulation, instead are secreted from the salivary glands in response to other biomolecules that get into the glands from circulation. Therefore, we collected proteins that can move from circulation to saliva and have been experimentally validated and reported in the literature, such as IgA [6,37], albumin and Zn-alpha2-glycoprotein [38]. 62 such proteins are found from the literature and used as the positive training data, shown in Table S9. Considering the relatively small size of this positive training dataset, we added additional proteins from the same Pfam families of these 62 proteins with sequence similarities lower than 30% to our training set, assuming that proteins in the same Pfam family have the same properties in getting into saliva. To avoid the issue of over-representing any particular family, we limit to have at most five additional members per family, specifically the most distant five members of each of the 62 proteins. This gives rise to a total of 261 proteins, which are used as the positive training data.

Generating the negative training data is a challenge since our information about which proteins are movable or not is clearly incomplete at this point. We employed a method similar to that proposed by Cui et al. [35] by choosing proteins from the Pfam families not containing any proteins that have been detected in saliva. For each such family, we choose five members as the negative training data. In addition, we keep only those with at least five peptides in the Plasma Proteome Project (PPP) database [39], the largest human plasma protein database. As a result, 6,816 proteins are selected as the negative dataset.

Feature construction

To train a classifier for proteins that are movable from circulation to saliva, we consider the following features, which can be grouped into four categories: (i) general sequence features such as sequence length, amino acid composition and di-peptide composition; (ii) physicochemical properties such as hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charges, solubility, unfoldability and disordered regions; (iii) domains/motifs such as signal peptides, transmembrane domains and twin-arginine signal peptides motif (TAT); and (iv) structural properties such as secondary structural content and radius of gyration, totaling 34 features, represented by 1,523 feature elements. The details of these features are provided in Table S1.

Feature selection and classification

For each protein, we calculated a feature vector of 1,523 dimensions defined above. We first trained a classifier using all the 1,523 feature values on the training data, and then applied a two-stage feature-selection procedure to remove those irrelevant and redundant features. A permutation test and q-

value [9] are used to identify and remove the irrelevant features. 10,000 permutations are generated and used to calculate the statistical significance on the relevance of individual feature elements to the prediction accuracy. Then, we used the approach proposed by Storey and Tibshirani [9] to calculate the q-value, which is used to control the False Discovery Rate (FDR) [40], in terms of the p-value obtained from the permutation test. We used 0.005 as the q-value cutoff to remove less relevant features, giving rise to 1,087 retained feature elements. In the second step, an improved feature selection method (SVM-RFE) that considers dependence relationships among features [41] is applied to rank these features. Then we went through an iterative classification and feature removal procedure to have kept only 55 feature elements, which give essentially the same classification result as using the larger feature set.

A SVM-based classifier is trained on the training data using the 55 feature elements for each protein, and the output is 1 or -1 representing if the input protein is movable to saliva or not. The following parameters are used to evaluate the prediction performance: recall, precision and the area under curve (AUC) of the recall-precision curve [42], defined as follows:

$$\text{recall} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (2)$$

where *TP* is the number of true positives, *FP* refers to the number of false positives, and *FN* is the number of false negatives.

A method for ranking predicted salivary proteins

We have also ranked the predicted salivary proteins using the manifold ranking algorithm as in our previous work [8]. The essence of a manifold ranking algorithm [43,44] can be intuitively explained as follows: the problem is defined on two datasets, a true sample set and a background set. Our goal is to rank the individual members of the background set according to their relevance to the true samples. A weighted graph is used to represent the combined true and the background set, with each sample represented as a node of the graph and each pair of nodes being represented as an edge with a weight defined as the similarity between the two nodes in the feature space. Then an evidence propagation process starts, in which each true sample propagates its presence to its neighboring nodes to increase their relevance to the true sample set, where the increased relevance is valued proportionally to the corresponding edge weight in the graph. An overall relevance score of each node is summed over all the scores propagated to it from all the relevant true samples, by which elements in the background set can be ranked at the end. For our problem, the true sample set is the same as the positive training dataset defined in the previous section, and the background contains all the 20,209 human proteins in UniProt minus the positive set.

Identification of genes differentially expressed in breast cancer

The microarray gene expression datasets GSE15852 for 43 paired samples of breast cancer and adjacent normal tissues are downloaded from the GEO database of NCBI [45]. For these samples, we applied t test and fold-change to identify differentially expressed genes in cancer *versus* control samples. The expression fold changes of each gene can be calculated using the following formula:

$$fc_i = \frac{1}{m}\sum_{j=1}^{m}\frac{c_{ij}}{n_{ij}} \quad (3)$$

where fc_i is the ratio of the gene expression value on cancer sample versus control sample of gene i. c_{ij} is the expression value of gene *i* of cancer sample in patient *j*, and n_{ij} is the expression value of gene *i* of normal sample in patient j. m = 43 is the sample number. The fc_i value is greater than one for up-regulated genes and less than one for down-regulated genes. To identify differentially expressed genes, we choose 1.5 as the threshold of fold change (1/1.5 for down-regulation). Then we can obtain the differentially expressed genes between cancer samples *versus* control samples.

P-value calculation for comparison of the ranking result with human saliva biomarkers

We calculated the statistical significance p-value assuming the underlying distribution for our problem follows a hypergeometric distribution [46], i.e., the probability of selecting *s* tails in *n* draws without replacement from a finite population of size *N* coins each with an equal probability in selecting a head *versus* a tail containing exactly *S* tails, calculated as follows:

$$P\left(x=s\right) = \frac{C(S,s)\cdot C(N-S,n-s)}{C(N,n)} = \frac{\binom{S}{s}\binom{N-S}{n-s}}{\binom{N}{n}} \quad (4)$$

Where C(a, b)=a!/[b!(a−b)!], *N* is the number of human proteins, *n* is the number of the selected top proteins, *S* is the number of proteins used as salivary biomarkers, and *s* is the number of proteins that are among the 47 known salivary biomarkers and among the top *n* predicted candidate proteins. *N* is 20,209 and *S* is 47. Table 1 shows the p-values, for different *s* and *n*.

Supporting Information

Table S1. A list of initial features for prediction of salivary proteins from blood circulation.
(XLS)

Table S2. Features of blood-originated salivary proteins as selected by recursive feature elimination method.
(XLS)

Table S3. A list of top 1000 blood-originated salivary proteins that ranked by manifold ranking method.
(XLS)

Table S4. Salivary proteins that have been associated with human diseases.
(XLS)

Table S5. Result of GO enrichment analysis among the top 1,000 ranked proteins.
(XLS)

Table S6. A list of candidate up- and down-regulated salivary proteins in breast cancer.
(XLS)

Table S7. Result of GO enrichment analysis among the 31 predicted proteins.
(XLS)

Table S8. Result of Pathway enrichment analysis among the 31 predicted proteins.
(XLS)

Table S9. A list of proteins that can move from circulation to saliva and have been experimentally validated and reported in the literature.
(XLS)

Acknowledgments

We acknowledge the assistance of our teachers and students at JLU and UGA. In particular, we thank Chi Zhang (Computational Systems Biology Lab of the UGA) for his assistance in statistical analyses in this project.

Author Contributions

Conceived and designed the experiments: JXW YCL YW WD YX. Performed the experiments: JXW WD. Analyzed the data: JXW YW YX. Contributed reagents/materials/analysis tools: JC ML. Wrote the manuscript: JXW YCL YW WD YX.

References

1. Denny P, Hagen FK, Hardt M, Liao L, Yan W et al. (2008) The proteomes of human parotid and submandibular/sublingual gland salivas collected as the ductal secretions. J Proteome Res 7: 1994-2006. doi:10.1021/pr700764j. PubMed: 18361515.
2. Jenzano JW, Courts NF, Timko DA, Lundblad RL (1986) Levels of glandular kallikrein in whole saliva obtained from patients with solid tumors remote from the oral cavity. J Dent Res 65: 67-70. doi: 10.1177/00220345860650011201. PubMed: 3455701.
3. Turan T, Demir S, Aybek H, Atahan O, Tuncay OL et al. (2000) Free and total prostate-specific antigen levels in saliva and the comparison with serum levels in men. Eur Urol 38: 550-554. doi: 10.1159/000020354. PubMed: 11096235.
4. Streckfus C, Bigler L, Tucci M, Thigpen JT (2000) A preliminary study of CA15-3, c-erbB-2, epidermal growth factor receptor, cathepsin-D, and p53 in saliva among women with breast carcinoma. Cancer Invest 18: 101-109. doi:10.3109/07357900009038240. PubMed: 10705871.
5. Wong DT (2006) Salivary diagnostics powered by nanotechnologies, proteomics and genomics. J Am Dent Assoc 137: 313-321. PubMed: 16570464.
6. Pfaffe T, Cooper-White J, Beyerlein P, Kostner K, Punyadeera C (2011) Diagnostic potential of saliva: current state and future applications. Clin Chem 57: 675-687. doi:10.1373/clinchem. 2010.153767. PubMed: 21383043.
7. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L et al. (2002) The Pfam protein families database. Nucleic Acids Res 30: 276-280. doi: 10.1093/nar/30.1.276. PubMed: 11752314.
8. Liu Q, Cui JA, Yang QA, Xu Y (2010) In-silico prediction of blood-secretory human proteins using a ranking algorithm. Bmc Bioinforma 11: 250-. PubMed: 20465853.
9. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445. doi:10.1073/pnas. 1530509100. PubMed: 12883005.
10. Brandtzaeg P (1971) Human secretory immunoglobulins. II. Salivary secretions from individuals with selectively excessive or defective synthesis of serum immunoglobulins. Clin Exp Immunol 8: 69-85. PubMed: 4099989.
11. UniProt C (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71-D75. doi: 10.1093/nar/gkr981. PubMed: 22102590.
12. Chen Y, Zhang Y, Yin Y, Gao G, Li S et al. (2005) SPD--a web-based secreted protein database. Nucleic Acids Res 33: D169-D173. doi: 10.1093/nar/gni168. PubMed: 15608170.
13. Sprenger J, Lynn Fink J, Karunaratne S, Hanson K, Hamilton NA et al. (2008) LOCATE: a mammalian protein subcellular localization database. Nucleic Acids Res 36: D230-D233. PubMed: 17986452.
14. Li SJ, Peng M, Li H, Liu BS, Wang C et al. (2009) Sys-BodyFluid: a systematical database for human body fluid proteome research. Nucleic Acids Res 37: D907-D912. doi:10.1093/nar/gkn849. PubMed: 18978022.
15. Grant K, Loizidou M, Taylor I (2003) Endothelin-1: a multifunctional molecule in cancer. Br J Cancer 88: 163-166. doi:10.1038/sj.bjc. 6700750. PubMed: 12610497.
16. Cheng YS, Rees T, Jordan L, Oxford L, O'Brien J et al. (2011) Salivary endothelin-1 potential for detecting oral cancer in patients with oral lichen planus or oral cancer in remission. Oral Oncol 47: 1122-1126. doi:10.1016/j.oraloncology.2011.07.032. PubMed: 21868280.
17. Holten-Andersen L, Jensen SB, Bardow A, Harslund J, Thaysen-Andersen M et al. (2008) Identifying sources and estimating glandular output of salivary TIMP-1. Scand J Clin Lab Invest 68: 548-554. doi: 10.1080/00365510701883180. PubMed: 18609089.
18. Herr AE, Hatch AV, Throckmorton DJ, Tran HM, Brennan JS et al. (2007) Microfluidic immunoassays as rapid saliva-based clinical diagnostics. Proc Natl Acad Sci U S A 104: 5268-5273. doi:10.1073/ pnas.0607254104. PubMed: 17374724.
19. Christodoulides N, Mohanty S, Miller CS, Langub MC, Floriano PN et al. (2005) Application of microchip assay system for the measurement of C-reactive protein in human saliva. Lab Chip 5: 261-269. doi: 10.1039/b414194f. PubMed: 15726202.
20. Hu S, Arellano M, Boontheung P, Wang J, Zhou H et al. (2008) Salivary proteomics for oral cancer biomarker discovery. Clin Cancer Res 14: 6246-6252. doi:10.1158/1078-0432.CCR-07-5037. PubMed: 18829504.
21. St John MAR, Li Y, Zhou XF, Denny P, Ho CM et al. (2004) Interleukin 6 and interleukin 8 as potential biomarkers for oral cavity and oropharyngeal squamous cell carcinoma. Arch Otolaryngol Head Neck Surg 130: 929-935. doi:10.1001/archotol.130.8.929. PubMed: 15313862.
22. Rhodus N, Dahmer L, Lindemann K, Rudney J, Mathur A et al. (1998) s-IgA and cytokine levels in whole saliva of Sjogren's syndrome patients before and after oral pilocarpine hydrochloride administration: a pilot study. Clin Oral Investig 2: 191-196. doi:10.1007/ s007840050069. PubMed: 10388393.
23. Streckfus C, Bigler L, Navazesh M, Al-Hashimi I (2001) Cytokine concentrations in stimulated whole saliva among patients with primary Sjogren's syndrome, secondary Sjogren's syndrome, and patients with primary Sjogren's syndrome receiving varying doses of interferon for

symptomatic treatment of the condition: a preliminary study. Clin Oral Investig 5: 133-135. doi:10.1007/s007840100104. PubMed: 11480812.
24. Ben-Chetrit E, Fischel R, Rubinow A (1993) Anti-SSA/Ro and anti-SSB/La antibodies in serum and saliva of patients with Sjogren's syndrome. Clin Rheumatol 12: 471-474. doi:10.1007/BF02231773. PubMed: 8124907.
25. Streckfus C, Bigler L, Dellinger T, Dai X, Kingman A et al. (2000) The presence of soluble c-erbB-2 in saliva and serum among women with breast carcinoma: a preliminary study. Clin Cancer Res 6: 2363-2370. PubMed: 10873088.
26. Streckfus C, Bigler L (2005) The use of soluble, salivary c-erbB-2 for the detection and post-operative follow-up of breast cancer in women: the results of a five-year translational research study. Adv Dent Res 18: 17-24. doi:10.1177/154407370501800105. PubMed: 15998939.
27. Streckfus CF, Mayorga-Wark O, Arreola D, Edwards C, Bigler L et al. (2008) Breast cancer related proteins are present in saliva and are modulated secondary to ductal carcinoma in situ of the breast. Cancer Invest 26: 159-167. doi:10.1080/07357900701783883. PubMed: 18259946.
28. Bigler LR, Streckfus CF, Dubinsky WP (2009) Salivary biomarkers for the detection of malignant tumors that are remote from the oral cavity. Clin Lab Med 29: 71-85. doi:10.1016/j.cll.2009.01.004. PubMed: 19389552.
29. Zhang L, Xiao H, Karlan S, Zhou H, Gross J et al. (2010) Discovery and preclinical validation of salivary transcriptomic and proteomic biomarkers for the non-invasive detection of breast cancer. PLOS ONE 5: e15573. doi:10.1371/journal.pone.0015573. PubMed: 21217834.
30. Chen DX, Schwartz PE, Li FQ (1990) Saliva and serum CA 125 assays for detecting malignant ovarian tumors. Obstet Gynecol 75: 701-704. PubMed: 2179784.
31. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W et al. (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4: P3. doi:10.1186/gb-2003-4-5-p3.
32. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30. doi:10.1093/nar/28.7.e27. PubMed: 10592173.
33. Becker KG, White SL, Muller J, Engel J (2000) BBID: the biological biochemical image database. Bioinformatics 16: 745-746. doi:10.1093/bioinformatics/16.8.745. PubMed: 11099263.
34. Nishimura D (2001) BioCarta. Biotech Software & Internet Report: The Computer Software. J for Scient 2: 117-120.
35. Cui J, Liu Q, Puett D, Xu Y (2008) Computational prediction of human proteins that can be secreted into the bloodstream. Bioinformatics 24: 2370-2375. doi:10.1093/bioinformatics/btn418. PubMed: 18697770.
36. Hong CS, Cui J, Ni Z, Su Y, Puett D et al. (2011) A computational method for prediction of excretory proteins and application to identification of gastric cancer markers in urine. PLOS ONE 6: e16875. doi:10.1371/journal.pone.0016875. PubMed: 21365014.
37. Chiappin S, Antonelli G, Gatti R, De Palo EF (2007) Saliva specimen: a new laboratory tool for diagnostic and basic investigation. Clin Chim Acta 383: 30-40. doi:10.1016/j.cca.2007.04.011. PubMed: 17512510.
38. Schenkels LC, Veerman EC, Nieuw Amerongen AV (1995) Biochemical composition of human saliva in relation to other mucosal fluids. Crit Rev Oral Biol Med 6: 161-175. doi:10.1177/10454411950060020501. PubMed: 7548622.
39. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P et al. (2006) The PeptideAtlas project. Nucleic Acids Res 34: D655-D658. doi:10.1093/nar/gkj040. PubMed: 16381952.
40. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Stat Methodol): 289-300.
41. Du W, Sun Y, Wang Y, Cao Z, Zhang C et al. (2013) A novel multi-stage feature selection method for microarray expression data analysis. Int J Data Min Bioinform 7: 58-77. doi:10.1504/IJDMB.2013.050977. PubMed: 23437515.
42. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13: 1443-1471. doi:10.1162/089976601750264965. PubMed: 11440593.
43. Zhou DY, Weston J, Gretton A, Bousquet O, Scholkopf B (2004) Ranking on data manifolds. Adv Neural Inf Process Syst 16 16: 169-176.
44. He J, Li M, Zhang HJ, Tong H, Zhang C (2006) Generalized manifold-ranking-based image retrieval. IEEE Trans Image Process 15: 3170-3177. doi:10.1109/TIP.2006.877491. PubMed: 17022278.
45. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37: D885-D890. doi:10.1093/nar/gkn764. PubMed: 18940857.
46. Rice JA (2007) Mathematical statistics and data analysis. Belmont, CA: Thomson/Brooks/Cole.