Published in final edited form as:
J Clin Periodontol. 2013 February ; 40(2): 131–139. doi:10.1111/jcpe.12037.

Discovery of biomarker combinations that predict periodontal health or disease with high accuracy from GCF samples based on high-throughput proteomic analysis and mixed-integer linear optimization

Richard C. Baliban[*], **Dimitra Sakellari**[#], **Zukui Li**[*], **Yannis Guzman**[*], **Benjamin A. Garcia**[&], and **Christodoulos A. Floudas**[*]

[*]Department of Chemical and Biological Engineering, Princeton University, Princeton, USA

[#]Department of Preventive Dentistry, Periodontology and Implant Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece

[&]Department of Molecular Biology, Princeton University, Princeton, USA

Abstract

Aim—To identify optimal combination(s) of proteomic based biomarkers in gingival crevicular fluid (GCF) samples from chronic periodontitis (CP) and periodontally healthy individuals and validate the predictions through known and blind test sets.

Materials and Methods—GCF samples were collected from 96 CP and periodontally healthy subjects and analyzed using high-performance liquid chromatography, tandem mass spectrometry, and the PILOT_PROTEIN algorithm. A mixed-integer linear optimization (MILP) model was then developed to identify the optimal combination of biomarkers which could clearly distinguish a blind subject sample as healthy or diseased.

Results—A thorough cross-validation of the MILP model capability was performed on a training set of 55 samples and greater than 99% accuracy was consistently achieved when annotating the testing set samples as healthy or diseased. The model was then trained on all 55 samples and tested on two different blind test sets, and using an optimal combination of 7 human proteins and 3 bacterial proteins, the model was able to correctly predict 40 out of 41 healthy and diseased samples.

Conclusions—The proposed large-scale proteomic analysis and MILP model led to the identification of novel combinations of biomarkers for consistent diagnosis of periodontal status with greater than 95% predictive accuracy.

Keywords

periodontitis; gingival crevicular fluid; tandem mass spectrometry; biomarkers; mixed-integer linear optimization

Corresponding author: Christodoulos A. Floudas, Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, 08544, USA. Tel: (609) 258-4595, Fax: (609) 258-0211, floudas@titan.princeton.edu.

Introduction

A biological marker (biomarker) is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacologic responses to a therapeutic intervention" (Biomarkers Definition Group, 2001). These objective characteristics can be applied as diagnostic, staging, or prognostic tools for a disease and also assist in monitoring clinical response to treatment. In addition, biomarkers could serve as "surrogate clinical endpoints," and could predict clinical benefits in therapeutic intervention trials. A substantial body of literature refers to the identification of biomarkers in gingival crevicular fluid (GCF) discussing the complexity of molecules that GCF contains, the possibility of non-invasive collection, and the potential for site-specific analysis (Champagne et al. 2003, Heitz-Mayfield 2005, Loos & Tjoa 2005, Buduneli & Kinane 2011). Despite the considerable number of related investigations, some of which have resulted in the emergence of diagnostic kits, no tests are currently available for accurate and reliable diagnosis or prognosis in clinical periodontology.

The introduction of novel high-throughput technologies such as proteomic approaches in host-derived clinical samples such as serum or saliva is an innovative approach that could greatly enhance current knowledge of the proteins involved in health or disease (Loo et al. 2010). Data from large -scale proteomic analysis have been reported for saliva, and it has been suggested that the salivary proteome could assist in the assessment of periodontal disease activity (Zhang et al. 2009). Although saliva is easy-to-collect and can reflect the whole-mouth situation, it is considered as a "surrogate" fluid for GCF and its versatile biochemistry require cautious interpretation of findings (Chapple 2009). Currently, limited data in the literature exist for large-scale proteomic analysis of GCF. These reports refer to periodontal health (Carneiro et al. 2012), periodontal health and disease (Bostanci et al. 2010, Baliban et al. 2012a), periodontal patients at maintenance phase (Ngo et al. 2010), or investigated changes during the inflammatory process in an experimental gingivitis model (Grant et al. 2010), and have shown an abundance of mainly host-derived proteins in clinical samples. Recent reports that apply large-scale genomic, transcriptomic, and metabolomic analysis to investigate periodontal conditions, although restricted in number, have demonstrated both the absolute need for integration of these tools and the sometimes unexpected findings that data analysis can generate (Barnes et al. 2009, Schaefer et al. 2010, Jonnson et al. 2011).

By using high-performance liquid chromatography (HPLC), tandem mass spectrometry (MS/MS) and the PILOT_PROTEIN protein identification method (DiMaggio & Floudas 2007a,b, DiMaggio et al. 2008, Baliban et al. 2010, Baliban et al. 2012b), 432 human (120 new) and 30 bacterial proteins were identified in GCF samples from 12 chronic periodontitis (CP) and 12 periodontally healthy individuals (Baliban et al. 2012a). A number of human and bacterial proteins, most of which have not been extensively investigated before, were identified as candidate biomarkers for periodontal health or disease. The candidate biomarkers listed in Baliban et al. (2012a) focused on proteins that were found only in periodontally healthy or chronic periodontitis (CP) subjects. Though these proteins show strong evidence to indicate their selection as biomarkers, it is not necessarily essential that these proteins are the only candidates. In fact, there were several human proteins reported that frequently appeared in the healthy subjects but also appeared a few times in the CP subjects, or vice versa. By selecting a combination of proteins to act as biomarkers for the diagnosis of a GCF sample, it is possible that the unexpected presence of a protein in a sample (e.g., a health-related protein found in a CP sample) would be outweighed by the presence of other proteins that validate the sample diagnosis. The fundamental question that needs to be addressed is what the minimal subset of proteins would be required to distinguish a GCF sample as healthy or diseased based solely on the presence or absence of those proteins. A detailed understanding of the key human and bacterial proteins for such an analysis would include the number of human proteins needed, the number of bacterial proteins needed, the number of proteins related to periodontal health, the number of proteins related to periodontal disease, and the relative importance of each of the proteins.

The aim of the present study is to identify and validate optimal combinations of protein biomarkers in GCF samples using a comprehensive proteomic analysis. The PILOT_PROTEIN protein identification methods (DiMaggio & Floudas 2007a,b, DiMaggio et al. 2008, Baliban et al. 2010, Baliban et al. 2012b) and the recently developed webtool (http://pumpd.princeton.edu) were used to generate a complete list of human and bacterial proteins for each GCF sample, and a novel mixed-integer linear optimization (MILP) model was formulated to identify the combination of biomarkers which would predict whether a sample derives from a periodontally healthy or CP subject. The model was first tested using a thorough cross-validation of 55 samples. Subsequently, two blind test sets, containing a total of 41 samples, were used to test the proposed combination(s) of human and bacterial biomarkers.

Materials and Methods

Subject sample collection

Forty- five periodontally healthy and fifty-one periodontally diseased and non-previously treated subjects participated in the present study. All participants were patients of the Department of Preventative Dentistry, Periodontology, and Implant Biology, Dental School, Aristotle University, Thessaloniki, Greece or were personnel of the Dental School. Demographic and clinical data for participants are presented in Table 1.

All subjects were systematically healthy, non-smokers and not taking medication known to affect periodontal tissues. Subjects reporting antibiotic intake during the previous six months and pregnant or lactating women were excluded from the present study. Subjects were considered to be periodontally healthy or to have chronic periodontitis as previously described (Armitage 1999, Baliban et al. 2012a). Care was taken to include age-matched individuals across the two groups. All subjects signed an informed consent, and the study was conducted according to the protocol outlined by the Research Committee, Aristotle University of Thessaloniki, Greece, and approved by the Ethical Committee of the School of Dentistry.

Clinical recording and sampling and gingival crevicular fluid sample collection

Clinical recordings were performed by a calibrated examiner (DS) using a manual Williams probe (POW, Hu-Friedy, Chicago, IL). The examiner has reproducible recordings (Pearson’s test r=0.971) as determined in 10% of her weekly registrations. Parameters assessed included probing depth, recession, and bleeding on probing at six sites of all teeth present in the dentition. For clinical parameters, the statistical analysis of the data was carried out with the statistical package SPSS (14.0 version). Indicators of Descriptive Statistics were used, including mean and standard deviation for each group with the patient as the observational unit. Differences in clinical parameters were sought by applying a Mann-Whitney test, with a significance level of 0.05 (Table 1).

Gingival crevicular fluid samples

Each participant contributed with one pooled GCF sample from four pre-selected sites. For periodontitis cases, the sample was taken from sites which displayed probing depth >6 mm and <8 mm. For periodontally healthy individuals, the samples were taken from the mesiobuccal sites of first molars. GCF samples were obtained as previously described

(Sakellari et al. 2008). The samples were immediately placed in Eppendorf tubes containing 100 μL of 100 mM ammonium bicarbonate, frozen in liquid nitrogen, and stored at −80 °C. GCF samples were collected prior to the clinical measurements and were discarded when visibly contaminated with blood. Samples were lyophilized for 5 hours at −55 °C and 0.03 mbar in an ALPHA 1-4 (Martin Christ, Gefriertrocknungsan-langen GmbH) lyophilizer. Prior to lyophilization, they were vortexed for 20 min and centrifuged for 10 min at 8,161.4 g while all filter strips were discarded.

Sample preparation and mass spectrometry data analysis

The lyophilized protein samples were reconstituted in water, prepared, and digested with trypsin as previously described (Baliban et al. 2012a). One aliquot of each digested sample (containing approximately 10 μg of protein) was extracted for MS/MS analysis using a hybrid quadrupole ion trap-Orbitrap mass spectrometer as previously described (Baliban et al. 2012a). Peptides were separated by RP-HPLC using a gradient from 2% to 45% Buffer B (Buffer A, 0.1 M acetic acid; Buffer B, 70% acetonitrile in 0.1 M acetic acid) at a flow rate of 200 nL/min for 110 min. The Orbitrap instrument was operated in data-dependent mode using a resolution of 30,000 to obtain a full MS spectrum followed by seven MS/MS spectra obtained in the ion trap. All MS/MS spectra were processed using the on-line version of the PILOT_PROTEIN protein identification methods (DiMaggio & Floudas 2007a,b, DiMaggio et al. 2008, Baliban et al. 2010, Baliban et al. 2012b) and the recently developed webtool (http://pumpd.princeton.edu) using a subset of the Swissprot database derived from the Homo sapiens taxonomy and all bacterial taxonomies (Baliban et al. 2012a). Search tolerances included a value of 0.1 Da for the precursor ion and 0.5 Da for the fragment ion. Searches were performed using a maximum of 2 missed cleavages and a static cysteine modification of 57 Da due to the iodoacetimide treatment. The false discovery rate utilized in this study was 2% and was calculated using a reverse-sequence decoy database. Positive identification of a protein was allowed for one peptide if that particular amino acid sequence could not be associated with another protein in the database. All other positive protein identifications required at least two annotated peptides. Note that additional MS-based techniques including GCF sample fractionation and LC-MS/MS protein targeting can be applied for large-scale proteomic studies. However, an established protocol for GCF sample analysis will be utilized within this study to ensure that a valid comparison can be made with previous results (Baliban et al. 2012a).

Mathematical model for prediction of periodontal status

Scoring function—To predict the periodontal status of a GCF sample based on a list of identified proteins within the sample, the following score measure is proposed:

$$S_i=\sum_i W_i A_{ij} Y_i \qquad (1)$$

In Equation (1), the matrix A contains the protein information in the sample: $A_{ij} = 1$ if protein i exists in sample j, and $A_{ij} = 0$ otherwise. The y_i are binary parameters denoting whether protein i is selected as biomarker ($y_i = 1$) or not ($y_i = 0$), and the w_j are weight parameters for each protein. Upon determination of the y_i and w_{i}, the periodontal status of a sample can be inferred from Equation (1) based on the value of S_j as follows:

1. If $S_j \geq 1$, then the sample is periodontally healthy
2. If $S_j \leq -1$, then the sample is chronic periodontitis (diseased)
3. If $-1 < S_j < 1$, then no conclusion will be made (ambiguous).

Optimally selecting biomarker proteins and determining model parameters— To determine the parameters y_i and w_i in Equation (1), a mixed-integer linear optimization model is proposed to optimally train on a GCF data set with known healthy/diseased status and known protein identification results. Given a data set with known diseased/healthy status, and the proteins inside the samples A_{ij}, the goal of the model will be to find a set of parameters (y_i, w_i) that will minimize the prediction error and maximize the prediction accuracy.

To describe the proposed mathematical model, the following indices, sets, parameters and variables are defined:

Index	
i	protein
j	sample
Set	
H	healthy sample set
D	diseased sample set
Parameters	
A_{ij}	equal to 1 if protein i is in sample j; 0 otherwise
W_i^L , W_i^U	lower and upper bounds on weight variables for protein i
U	constant (U=50)
Variables	
y^i	binary, equal to 1 if protein i is selected as biomarker; 0 otherwise
z_j	binary, equal to 1 if sample j is diseased; 0 otherwise (i.e., healthy)
w^i	continuous, weight of protein i
μ_j^+, μ_j^-	continuous, score/threshold slack variables for sample j

The objective of the optimization model (Equation 2) is to find the combination of biomarkers that will minimize the score error when a wrong prediction is made and maximize the score margin when a correct prediction is made.

$$\min \sum_i \left(\mu_j^+ - 0.1\mu_j^-\right) \quad (2)$$

Score error and score margin are the absolute difference between prediction score and the threshold, and are represented by the slack variables $\mu^+{}_j$ and $\mu^-{}_j$, respectively. The score constraints for healthy and diseased samples are represented by Equations (3) and (4), respectively.

$$\sum_i w_i A_{ij} - 1 + \mu_j^+ - \mu_j^- = 0 \quad \forall j \in H \quad (3)$$

$$\sum_i w_i A_{ij} + 1 - \mu_j^+ + \mu_j^- = 0 \quad \forall j \in D \quad (4)$$

$$0 \le \mu_j^+, \mu_j^- \le 10 \quad (5)$$

Equation (5) is used to bound the slack variables such that the optimal solution of the problem is finite. The following constraints (Equation 6) restrict the total number of proteins used in the prediction model:

$$\sum_i y_i \le N \quad (6)$$

Equations (7) and (8) try to balance the number of human and bacterial proteins selected:

$$\sum_{i \in I_{human}} y_i \le 0.5 \sum_i y_i \quad (7)$$

$$\sum_{i \in I_{Bacteria}} y_i \le 0.5 \sum_i y_i \quad (8)$$

To identify a superset of proteins that can act as candidate biomarkers, all of the proteins identified previously by Baliban et al. (2012a) are preprocessed based on their prevalence in 55 training samples with known periodontal status. For each protein, the following parameters are calculated:

HT_i:	Number of times that protein *i* appears in healthy samples
DT_i:	Number of times that protein *i* appears in diseased samples
Hf_i:	Relative frequency of that protein *i* appears in healthy samples
$Hf_i = HT_i / (HT_i + DT_i)$	
Df_i:	Relative frequency of that protein *i* appears in diseased samples
$Df_i = DT_i / (HT_i + DT_i)$	

Given a cut-off value for each of the four metrics (e.g., $HT_i \ge 7$, $Hf_i \ge 0.5$, $DT_i \ge 7$, $Df_i \ge 0.5$), only a certain subset of proteins will be considered as candidate biomarkers, and they are further assigned to the following sets:

I_{DB}	Set of bacterial proteins found from chronic periodontitis samples (total = 13)
I_{DH}	Set of human proteins found from chronic periodontitis samples (total = 136)
I_{HB}	Set of bacterial proteins found from periodontally healthy samples (total = 12)
I_{HH}	Set of human proteins found from periodontally healthy samples (total = 86)

Based on the above sets, the following constraints are applied to enforce that at least one protein is selected from each set:

$$\sum_{i \in I_{HH}} y_i \ge 1, \quad \sum_{i \in I_{HB}} y_i \ge 1, \quad \sum_{i \in I_{DH}} y_i \ge 1, \quad \sum_{i \in I_{DB}} y_i \ge 1 \quad (9)$$

Notice that in Equations (3) and (4), the scoring function did not contain the binary variable y_i. Equation (10) links the binary variable to the weight variable and enforces that $w_i = 0$ if a protein is not selected (i.e., $y_i = 0$). This formulation is used to change the model from a nonlinear to a linear state, which allows for the determination of the global optimal solution. The weights of the proteins are also bounded by the following constraints:

$$y_i w_i^L \le w_i \le y_i w_i^U \quad \forall i \qquad (10)$$

The lower and upper bounds (W_i^L, W_i^U) are set as [0.5, 2] for proteins in healthy samples (i.e., $i \in I_{HH} \cup I_{HB}$) and [−2, −0.5] for proteins in diseased samples (i.e., $i \in I_{DH} \cup I_{DB}$).

Since the slack variables, μ^{-}_{j}, have negative weight in the objective, in the optimal solution the absolute weights of different proteins tend to reach the same upper bounds, thus the importance of different proteins cannot be reflected. To avoid this problem, the following constraints are introduced to bound the sum of the weight parameters:

$$\sum_{i \in I_{HH} \cup I_{HB}} w_i \le \sum_{i \in I_{HH} \cup I_{HB}} y_i \qquad (11)$$

$$\sum_{i \in I_{HH} \cup I_{HB}} w_i \ge - \sum_{i \in I_{HH} \cup I_{HB}} y_i \qquad (12)$$

To avoid generation of ambiguous predictions on the training set, Equations (13) and (14) are applied to enforce that all the training samples must be predicted either as healthy or diseased. For example, if $z_j = 1$, then Equation (13) is redundant and Equation (14) becomes $\sum_i w_i A_{ij} \le -1$ (i.e., the sample is diseased); if $z_j = 0$, then Equation (14) is redundant and Equation (13) becomes $\sum_i w_i A_{ij} \ge 1$ (i.e., the sample is healthy).

$$\sum_i w_i A_{ij} \ge 1 - z_j U \quad \forall j \qquad (13)$$

$$\sum_i w_i A_{ij} \le \left(1 - z_j\right) U - 1 \quad \forall j \qquad (14)$$

A complete mixed integer linear optimization model consisting of Equations (2) - (14) is formulated and can be solved to global optimality using CPLEX (ILOG 2010) to identify the value of the parameters y_i and w_i. Integer cuts (Equation 15) are used to find all solutions of the model that have the same globally optimal objective value.

$$\sum_{i \in B_c} y_i - \sum_{i \in NB_c} y_i \le |B_c| - 1 \quad \forall c \qquad (15)$$

In Equation (15), B_c is the set containing the index of binary variables y_i with a value equal to 1 in c-th iteration, NB_c is the set containing the index of binary variables y_i taking a value of 0, and $|B_c|$ is the cardinality of the set B_c.

Results

Cross validation study

Using the 55 samples (with label D1-D12, H1-H12, B1-B26, B28-B32) with known periodontal status and comprehensive proteomic analyses results as a basis, the capability of the proposed prediction model and the method of determining model parameters was initially analyzed using a cross-validation study. A training set of size *N* was selected from the 55 samples, the biomarker selection was optimized using the training set, and then the biomarkers were tested on the remaining (55 – *N*) samples. *N* was selected to be 20, 30, or 40 and, for each size, 100 different training sets were randomly selected. The optimization model was run for each of the 100 training sets to determine the prediction model parameters and then make the prediction on the corresponding test set. Finally, the average prediction accuracy for each size is reported.

The proposed optimization method of determining the model parameters is dependent on the candidate protein sets (I_{DB}, I_{DH}, I_{HB}, I_{HH}), so a thorough parametric study was conducted on different cut-off values for HT_i, Hf_i, DT_i, and Df_i (see Supplementary Material). When testing on different cut-off values from (7, 0.5, 7, 0.5) to (10, 0.7, 10, 0.7), the cross-validation accuracy is consistently around 99%, as shown in Table 2. The proposed model is capable of consistently finding an optimal combination of biomarkers that minimizes the error on the training set and performs accurately on the test sets.

The total number of proteins identified from each patient along with the total (i) unique peptide identifications, (ii) peptide-spectrum matches, and (iii) amino acid coverage for each protein is provided as Supplementary Material. The total protein count ranges from 42-190 for each sample, which is consistent with the results obtained from the previous study (Baliban et al. 2012a). The average number of proteins that have been identified in different samples is 62 for two healthy patients, 66 for two diseased patients, and 55 for one healthy and one diseased patient. Though the overlap for one diseased and one healthy patient is less than that for two healthy or two diseased patients, it is important to note that these average values are all within one standard deviation of one another. Thus, there are several proteins within GCF that can be identified regardless of the periodontal status of the patient. The development of a mathematical model that can identify candidate biomarkers is therefore imperative when attempting to perform a large-scale proteomics experiment.

Biomarker selection and model parameters

To generate an optimal combination of biomarkers for a blind test, the optimization model was trained on all aforementioned 55 samples (see Cross validation study). The cut-off values for HT_i, Hf_i, DT_i, and Df_i described above that were selected for the blind test are (7, 0.5, 7, 0.5), and were chosen after parametric analysis to provide the most conservative superset of candidate biomarkers. A comprehensive parametric analysis was performed for 144 distinct combinations of the cut-off values and the results of the blind test are shown in the Supplementary Material. Using integer cuts, all combinations of biomarkers that have the same optimal objective value were identified. The four optimal solutions, denoted as Models 1 – 4, are shown in Table 3. Each of the four solutions contains 10 proteins, 8 of which are common between all the solutions. The results of the biomarker combinations on the training samples are shown in Table S1 of the Supplementary Material which highlights that all the training samples are diagnosed accurately using the optimal solutions.

Blind tests

Using the optimal selection of biomarkers and weights generated from all the 55 samples (see Table 3), the proposed four predictive models were tested on two blind test sets, which have 20 samples (labeled as B33-B52) and 21 samples (labeled as BB1-BB21), respectively. The biomarker proteins that are present in each sample are shown in Table S2 and S3 of the Supplementary Material for each test set. Note that the values in Tables S2 and S3 correspond to the scores of protein identification reported by PILOT_PROTEIN and score of zero indicates that the particular protein is not found in the sample. The scores and the corresponding predicted periodontal status are listed in Table S4 and Table S5 of the Supplementary Material, respectively. For the 20 samples in blind test set 1, 19 were predicted correctly and only one (B33) is predicted incorrectly, corresponding to a prediction accuracy of 95% (Table S5). For blind test set 2, all 21 samples are correctly predicted. Summing over the 41 samples from both blind test sets, the total prediction accuracy is 97.56% (Table 4).

The diagnostic potential for the biomarker combinations is outlined in Table 4 for each model. For detection of chronic periodontitis, a true-positive (TP) is defined as a diseased signal in a CP sample and a false-positive (FP) is a diseased signal in a periodontally healthy sample. A true-negative (TN) is defined as a healthy signal in a periodontally healthy sample and a false-negative (FN) is a healthy signal in a CP sample (Listgarten 1986). The sensitivity of the model is defined as the number of TP divided by the total number of diseased sampled (TP + FN) and specificity of the model is defined as the number of TN divided by the total number of healthy samples (TN + FP). The positive predictive value is defined as the number of TP divided by the total number of positives (TP + FP) while the negative predictive value is defined as the number of TN divided by the number of negatives (TN+ FN). These last two metrics are important diagnostic indicators because they provide an indication of how accurate the model signal will be when the periodontal status of the patient is unknown. For the 20 healthy samples and the 21 diseased samples in the two blind test sets, each model reported a sensitivity of 95.2%, a specificity of 100%, a positive predictive value of 100%, and a negative predictive value of 95.2%.

Discussion

The medical and dental communities have long expressed a considerable interest in developing reliable biomarkers which can evaluate many different types of biological characteristics or parameters including genetics, imaging-based evaluations, and blood or GCF composition assessments. The introduction of novel technologies and technological advancements has allowed for the development of a number of biomarkers which have been integrated in clinical practice such as genomic biomarkers for specific forms of cancers or pharmacogenomic biomarkers for appropriate drug selection (Amur et al. 2008). At least three host-derived substances in GCF (alkaline phosphatase, beta-glucoronidase, and cathepsin B) have been shown to exhibit >77% of diagnostic accuracy in predicting future disease activity (Chapple 2009), while matrix metalloproteinases (MMPs) −8 and −9, neutrophil elastase, and dipeptidyl peptidase II and IV have been shown to correlate with disease presence and/or activity (Loos & Tjoa 2005, Chapple 2009). Although there have been promising results from the above mentioned studies, it is clear that specific and sensitive biomarkers are still required for consistent diagnosis, prognosis, and clinical monitoring of periodontal tissue destruction (Buduneli & Kinane 2011). Until recently, technological limitations hindered the possibility of analyzing individual GCF samples for multiple candidate biomarkers. Advances in technology allowed for the simultaneous identification and/or quantification of multiple such substances in individual GCF samples and therefore exhibited the significance of investigating combinations or panels of potential biomarkers (Offenbacher et al. 2010, Teles et al. 2010).

Human protein biomarkers

In the four different models generated using the proposed method (see Table 3), 8 proteins consistently appear in the biomarker protein selections. G3P_HUMAN, TYPH_HUMAN and KV101_HUMAN were selected as human protein biomarkers for periodontally healthy status. G3P_HUMAN (Glyceraldehyde 3-phosphate dehydrogenase) is an enzyme that participates in glycolysis and serves to break down glucose for energy and carbon molecules. This protein was observed in 24 of the 26 healthy samples and 8 of the 29 diseased samples from the training set. TYPH_HUMAN (Thymidine phosphorylase) is an enzyme that participates in purine metabolism pathway and pyrimidine metabolism pathway and is only found in periodontally healthy GCF samples from the training set. A previous metabolomics study on purine degradation found that the purine metabolic pathway was upregulated during gingivitis and periodontitis (Barnes et al. 2009). The authors note that the purine metabolic pathway is facilitated by multiple proteins (e.g., nuclease, nucleotidase, purine nucleoside phosphorylase) and the absence of TYPH_HUMAN in the diseased samples may be due to the enhanced abundance of these other proteins in the GCF samples. The protein KV101_HUMAN (Ig kappa chain V-I region AG) appeared in 17 of the periodontally healthy samples and 4 of the CP samples from the training data set. The functional role of the above mentioned proteins for periodontal health may be worth further investigation.

LYSC_HUMAN and HBD_HUMAN are two human proteins constantly selected biomarkers for chronic periodontitis. LYSC_HUMAN (Lysozyme C, also known as muramidase or N-acetylmuramide glycanhydrolase) is an antimicrobial protein that is part of the innate immune system and kills bacteria by attacking the bacterial cell walls. Lysozyme C is abundant in a number of secretions (e.g., tears, saliva, human milk, and mucus), and it has been reported that Lysozyme activity in crevicular fluid and in unstimulated saliva correlated with periodontal pocket depth in donors and in patients with gingivitis or periodontitis (Surna et al. 2009). This protein was observed in only 3 of the 26 healthy samples and all 29 of the diseased samples from the training data set. The increased expression of Lysozyme C in GCF samples from CP patients as reported in the present and a previous study (Baliban et al. 2012a) is possibly required to attenuate the antimicrobial activity of this protein in periodontal pockets and protect the host from further infection (Gorr & Abdolhosseini 2011). The HBD_HUMAN protein (Hemoglobin subunit delta) is encoded by the HBD gene and is normally expressed in adults. Two alpha chains plus two beta chains constitute HbA, which in normal adult life comprises about 97% of the total hemoglobin. This protein was found in 23 of the 29 CP samples and only 5 of the 26 periodontally healthy samples from the training data and could be indicative of the intense presence of red blood cells in the periodontal pocket (CP samples) although care was taken not to include GCF samples visibly contaminated with blood.

In addition to the above human proteins, PPIA_HUMAN appears in three of the four models. PPIA_HUMAN (Peptidyl-prolyl cis-trans isomerase A) is a cyclosporin binding-protein and is responsible for the folding of outer membrane proteins and may play a role in cyclosporin A-mediated immunosuppression. PPIA_HUMAN is also reported to have anti-microbial activity (Svensson et al. 2005). This protein appeared in 21 of the 29 diseased samples and only 5 of the 26 healthy samples from the training data set. The increased expression of PPIA_HUMAN is possibly caused by the onset of chronic periodontitis and may play an important role in killing unwanted bacteria.

Bacterial protein biomarkers

Table 3 shows that the bacterial protein biomarkers are a 33 kDa chaperonin (HSLO_OCEIH), a probable succinyl-CoA transferase (SCOB_BACSU), and a ribulose bisphosphate carboxylase (RBL2_RHORT). It is important to note that all three bacterial proteins participate in all four models. Though these proteins were readily identified in the GCF samples when using a database derived from all bacterial taxonomies, the species associated with the above three proteins are not part of a bacterial taxonomy that is generally associated with periodontal status. However, the proteins selected for these bacterial species have high sequence similarity to proteins from bacterial species associated with the periodontal environment that were not part of the Swissprot database.

All peptides identified using the PILOT_SEQUEL algorithm (DiMaggio et al. 2008) that were assigned to the above three proteins were searched against a subset of the NCBI non-redundant database containing only proteins from bacteria which are associated with periodontal status (Socransky et al. 1998). For each of the bacterial proteins, all peptides that were found during the proteomic analysis had very high sequence similarity with a protein in the non-redundant database. The 33 kDa chaperonin protein (HSLO_OCEIH) was identified as a biomarker for periodontal health and generally plays an important role in the bacterial defense system toward oxidative stress. This protein was identified using either or both of the peptide sequences DYLIK and TITITAMMGAMLK. These peptides have high sequence similarity to the peptides DYIVK and TMTATVMMGAMLK that are found in the 33 kDa chaperonin protein for the Prevotella dentalis DSM 3688 species (gi|330686056|). The mass difference of 14.016 Da between DYLIK and DYIVK may be due to a lysine methylation on the latter peptide and the mass difference of 3.941 Da between TITITAMMGAMLK and TMTATVMMGAMLK may be due to the presence of oxidations on the methionine residues rather than a dimethylation of the C-terminal lysine, which was often found on the TITITAMMGAMLK peptide.

The probable succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B protein (SCOB_BACSU) was identified as a biomarker for periodontal disease and catalyzes the reaction of succinyl-CoA and a 3-ketoacid. The GMGGAMDLVNGAK peptide used to consistently identify this protein is very similar to the GMGGAMDLVSGAK peptide that is part of the 3-oxoadipate CoA-succinyl transferase protein from the Campylobacter upsaliensis JV21 species (gi|315638767|). Note that the singular amino acid discrepancy between the two peptides may be caused by an amino acid mutation from a different strain of Campylobacter upsaliensis. The ribulose biphosphate carboxylase protein (RBL2_RHORT) contains the peptide VPEAYR which was found in multiple spectra. This peptide has similar sequence to VAEGLR, which is part of the triphosphate isomerase protein of Corynebacterium diphtheria (TPIS_CORDI) and Corynebacterium glutamicum (TPIS_CORGB/TPIS_CORGL).

Taken collectively, data from the present study provide new insight into the identification of GCF-derived sets of biomarkers which can accurately discriminate between periodontal health or disease. The above proposed combination of biomarkers will be further validated in subsequent experiments regarding various periodontal conditions, as well as dynamic changes during and after periodontal treatment. These findings demonstrate that advances in proteomics technology and optimization-based models for analysis and prediction of biomarkers can greatly contribute in developing reliable tools for Clinical Periodontology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Conflict of interest and Sources of Funding Statement

C.A.F. and B.A.G. acknowledge financial support from the National Science Foundation (CBET-0941143). C.A.F. acknowledges financial support from the National Institute of Health (R01LM009338). B.A.G. acknowledges support from Princeton University and the American Society for Mass Spectrometry Research award. The authors declare that there are no conflicts of interest in this study.

References

Amur S, Frueh FW, Lesko LJ, Huang S. Integration and use of biomarkers in drug development ,regulation and clinical practice: a US regulatory perspective. Biomarkers in Medicine. 2008; 2:305–311. [PubMed: 20477416]

Armitage GC. Development of a classification system for periodontal diseases and conditions. Annals of Periodontology. 1999; 4:1–6. [PubMed: 10863370]

Baliban RC, DiMaggio PA, Plazas-Mayorca MD, Young NL, Garcia BJ, Floudas CA. A novel approach for untargeted post-translational modification identification using integer linear optimization and tandem mass spectrometry. Molecular & Cellular Proteomics. 2010; 9:764–779. [PubMed: 20103568]

Baliban RC, Sakellari D, Li Z, DiMaggio PD, Floudas CA. Novel protein identification methods for biomarker discovery via a proteomic analysis of periodontally healthy and diseased gingival crevicular fluid samples. Journal of Clinical Periodontology. 2012a; 39:203–212. [PubMed: 22092770]

Baliban RC, DiMaggio PA, Plazas-Mayorca MD, Garcia BJ, Floudas CA. PILOT_PROTEIN: Identification of Unmodified and Modified Proteins via High-Resolution Mass Spectrometry and Integer Linear Optimization. Journal of Proteome Research. 2012b in press, doi:10.1021/pr300418j.

Barnes VM, Teles R, Trivedi HM, Devizio W, Xu T, Mitchell MW, Milburn MV, Guo L. Acceleration of purine degradation by periodontal diseases. Journal of Dental Research. 2009; 88:851–855. [PubMed: 19767584]

Biomarkers Definition Group. Biomarkers and surrogate end-points: preferred definitions and conceptual framework. Clinical Pharmacology and Therapeutics. 2001; 69:89–95. [PubMed: 11240971]

Bostanci N, Heywood W, Mills K, Parkar M, Nibali L, Donos N. Application of label-free absolute quantitative proteomics in human gingival crevicular fluid by LC/MS E (gingival exudatome). Journal of Proteome Research. 2010; 9:2191–2199. [PubMed: 20205380]

Buduneli N, Kinane DF. Host-derived diagnostic markers related to soft tissue destruction and bone degradation in periodontitis. Journal of Clinical Periodontology. 2011; 38:85–105. [PubMed: 21323706]

Carneiro LG, Venuleo C, Oppenheim FG, Salih E. Proteome data set of human gingival crevicular fluid from healthy periodontium sites by multidimensional protein separation and mass spectrometry. Journal of Periodontal Research. 2012; 47:248–262. [PubMed: 22029670]

Champagne CM, Buchanan W, Reddy MS, Preisser JS, Beck JD, Offenbacher S. Potential for gingival crevice fluid measures as predictors of risk for periodontal diseases. Periodontology. 2003; 2000(31):167–180.

Chapple ILC. Periodontal diagnosis and treatment-where does the future lie? Periodontology. 2009; 2000(51):9–24.

DiMaggio PA Jr. Floudas CA. De Novo Peptide Identification via Tandem Mass Spectrometry and Integer Linear Optimization. Analytical Chemistry. 2007; 79:1433–1446. [PubMed: 17297942]

DiMaggio PA Jr. Floudas CA. A Mixed-Integer Optimization Framework for De Novo Peptide Identification. AIChE Journal. 2007; 53:160–173. [PubMed: 19412358]

DiMaggio PA Jr. Floudas CA, Lu B, Yates JR III. A Hybrid Method for Peptide Identification Using Integer Linear Optimization, Local Database Search, and Quadrupole Time-of-Flight or Orbitrap Tandem Mass Spectrometry. Journal of Proteome Research. 2008; 7:1584–1593. [PubMed: 18324765]

Gorr SU, Abdolhosseini M. Antimicrobial peptides and periodontal disease. Journal of Clinical Periodontology. 2011; 38(Suppl 11):126–141. [PubMed: 21323710]

Grant MM, Creese AJ, Barr G, Ling MR, Scott AE, Matthews JB, Griffiths HR, Cooper HJ, Chapple ILC. Proteomic Analysis of a Noninvasive Human Model of Acute Inflammation and Its

Resolution: The Twenty-one Day Gingivitis Model. Journal of Proteome Research. 2010; 9:4732–4744. [PubMed: 20662485]

Heitz-Mayfield LZA. Disease progression: identification of high-risk groups and individuals for periodontitis. Journal of Clinical Periodontology. 2005; 32(suppl6):196–209. [PubMed: 16128838]

Jonnson D, Ramberg P, Demmer RT, Kebschull DT, Dahlen G, Papapanou PN. Gingival tissue transcriptome in experimental gingivitis. Journal of Clinical Periodontology. 2011; 38:599–611. [PubMed: 21501207]

Listgarten MA. A perspective on periodontal diagnosis. Journal of Clinical Periodontology. 1986; 13:175–181. [PubMed: 3457804]

Loo JA, Yan W, Ramachandran P, Wong DT. Comparative Human Salivary and Plasma Proteomes. Journal of Dental Research. 2010; 89:1016–1023. [PubMed: 20739693]

Loos BG, Tjoa S. Host-derived diagnosis markers for periodontitis: do they exist in gingival crevice fluid? Periodontology. 2005; 2000(39):53–72.

Ngo LH, Veith PD, Chen Y-Y, Chen D, Cardy IB, Reynolds EC. Mass Spectrometric Analyses of Peptides and Proteins in Human Gingival Crevicular Fluid. Journal of Proteome Research. 2010; 9:1683–1693. [PubMed: 20020772]

Offenbacher S, Barros S, Mendoza L, Mauriello S, Preisser J, Moss K, De Jager M, Aspiras M. Changes in gingival crevicular fluid inflammatory mediator levels during the induction and resolution of experimental gingivitis in humans. Journal of Clinical Periodontology. 2010; 37:324–333. [PubMed: 20447255]

Sakellari D, Menti S, Konstantinidis A. Free soluble receptor activator of nuclear factor-k b ligand in gingival crevicular fluid correlates with distinct pathogens in periodontitis patients. Journal of Clinical Periodontology. 2008; 35:938–943. [PubMed: 18988315]

Schaeffer AS, Richter GM, Nothnagel M, Manke T, Dommisch H, Jacobs G, Arlt A, Rosenstiel P, Noack B, Groessner-Schreiber B, Jepsen S, Loos BG, Schreiber S. A genome-wide association study identifies GLT6D1 as a susceptibility locus for periodontitis. Human Molecular Genetics. 2010; 19:553–562. [PubMed: 19897590]

Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL. Microbial complexes in subgingival plaque. Journal of Clinical Periodontology. 1998; 25:134–144. [PubMed: 9495612]

Surna A, Kubilius R, Sakalauskiene J, Vitkauskiene A, Jonaitis J, Saferis V, Gleiznys A. Lysozyme and microbiota in relation to gingivitis and periodontitis. Medical Science Monitor. 2009; 15:CR66–73. [PubMed: 19179970]

Svensson I, Calles K, Lindskog E, Henriksson H, Eriksson U, Haggstrom L. Antimicrobial activity of conditioned medium fractions from Spodoptera frugiperda Sf9 and Trichoplusia ni Hi5 insect cells. Applied Microbiology and Technology. 2005; 69:92–98.

Teles RP, Gursky LC, Faveri M, Rosa EA, Teles FR, Feres M, Socransky SS, Haffajee AD. Relationships between subgingival microbiota and GCF biomarkers in generalized aggressive periodontitis. Journal of Clinical Periodontology. 2010; 37:313–323. [PubMed: 20447254]

Zhang L, Henson BS, Camargo PM, Wong DT. The clinical value of salivary biomarkers for periodontal disease. Periodontology. 2009; 2000(51):25–37.

Clinical Relevance

Scientific rationale for study: No methodology has been developed to determine the optimal combination of biomarkers to predict the periodontal status of individuals based on analysis of gingival crevicular fluid (GCF) samples.

Principal findings: An optimal combination of 7 human and 3 bacterial protein biomarkers was used to identify the periodontal status of 96 GCF samples with an accuracy of over 95%.

Practical implications: The selection of an optimal combination of biomarkers to diagnose the periodontal health of disease of GCF samples could reach clinical praxis after proper validation.

TableI

Demographic and clinical characteristics of the subject sample.

Diagnosis	***Total***	***Male***	***Female***	***Age range***	***Mean age ± sd***	**Clinical parameters**			**Sampled stes**		
						Probing depth (mean ± sd)	Recession (mean ± sd)	Bleeding on probing (mean ± sd)	Probing depth (mean ± sd)	Recession (mean ± sd)	Bleeding on probing (mean ± sd)
Periodontally healthy	41	17	24	34-46	45.53 ± 6.97	**1.8 ± 0.24**	**0.08 ± 0.12**	**0.09 ± 0.07**	**1.67 ± 0.81**	**0.05 ± 0.21**	**0.05 ± 0.22**
Chronic periodontitis	55	28	27	30-61	48.02 ± 8.03	**3.82 ± 0.85**	**0.95 ± 0.88**	**0.60 ± 0.34**	**6.28 ± 0.55**	**0.7 ± 0.80**	**1**

No differences were observed between groups concerning mean age (Mann-Whitney test p>0.05).

Differences in clinical parameters among groups are indicated by bold lettering (Mann-Whitney test p<0.05).

Table 2

Cross validation accuracy using different training/test set

Training Set Size	Test Set Size	Average Correct Predictions*
40	15	14.98 (99.86%)
30	25	24.92 (99.68%)
20	35	34.68 (99.08%)

*Based on the following cut-off values: $HT_i \geq 7$, $Hf_i \geq 0.5$, $DT_i \geq 7$, $Df_i \geq 0.5$

Table 3

Biomarker Protein Selection and Weights

Protein	Model 1	Model 2	Model 3	Model 4
G3P_HUMAN	2	2	2	2
TYPH_HUMAN	1	0.5	0.5	0.5
KV101_HUMAN	0.5	0.5	0.5	0.5
LYSC_HUMAN	−2	−2	−2	−2
HBD_HUMAN	−0.5	−0.5	−0.5	−0.5
HSLO_OCEIH	1	1.5	1.5	1.5
RBL2_RHORT	−1.5	−1.5	−1.5	−1.5
SCOB_BACSU	−0.5	−0.5	−0.5	−0.5
PPIA_HUMAN	−0.5	−0.5		−0.5
ANGT_HUMAN	0.5		0.5	
LDH6B_HUMAN		0.5		
CAH1_HUMAN			−0.5	
THRB_HUMAN				0.5

Table 4

Blind Test Accuracy Summary

	Model 1	Model 2	Model 3	Model 4	No. Samples
		Blind Set 1			
Healthy	10 (1.000)	10 (1.000)	10 (1.000)	10 (1.000)	10
Diseased	9 (0.900)	9 (0.900)	9 (0.900)	9 (0.900)	10
Total	19 (0.950)	19 (0.950)	19 (0.950)	19 (0.950)	20
		Blind Set 2			
Healthy	10 (1.000)	10 (1.000)	10 (1.000)	10 (1.000)	10
Diseased	11 (1.000)	11 (1.000)	11 (1.000)	11 (1.000)	11
Total	21 (1.000)	21 (1.000)	21 (1.000)	21 (1.000)	21
	Predictive Value for Both Blind Sets				
Sensitivity	20 (0.952)	20 (0.952)	20 (0.952)	20 (0.952)	21
Specificity	20 (1.000)	20 (1.000)	20 (1.000)	20 (1.000)	20
Positive Predictive Value	20 (1.000)	20 (1.000)	20 (1.000)	20 (1.000)	20
Negative Predictive Value	20 (0.952)	20 (0.952)	20 (0.952)	20 (0.952)	21

TP: Diseased prediction for diseased sample; FP: Diseased prediction for healthy sample;

TN: Healthy prediction for healthy sample; FN: Healthy prediction for diseased sample; Sensitivity: TP/(TP + FN);

Specificity: TN/(TN + FP); Positive Predictive Value: TP/(TP + FP); Negative Predictive Value: TN/(TN + FN)