Sequence analysis

Informatic modeling of enamel pellicle interactions

Jeremy A. Horst[1,2], E. Emre Oren[3], Orapin V. Horst[4], L. Hong Hung[1], Ram Samudrala[1,*]

[1]Computational Genomics Group, Department of Microbiology, University of Washington, Seattle, USA; [2]Division of Pediatric Dentistry, Department of Orofacial Sciences, University of California, San Francisco, USA; [3]Department of Biomedical Engineering, TOBB University of Economics and Technology, Ankara, Turkey; [4] Division of Endodontics, Department of Preventive and Restorative Dental Sciences, University of California, San Francisco USA.

Received on April xx, 2012; revised on xx xx, xxxx; accepted on xx xx, xxxx

Advance Access publication . . .

ABSTRACT

Motivation: Protein-hydroxyapatite interactions govern development and homeostasis of mineralized tissues including tooth and bone. Little is known about these interactions because no available bench techniques produce robust data for assessing phase interfaces. Characterization will enable design of peptides for repair and regeneration of mineralized tissues.

Results: We show that tooth enamel pellicle peptides have subtle sequence similarities that encode hydroxyapatite-binding mechanisms, by segregating them from control peptides using a substitution matrix-based peptide comparison protocol. We improved discrimination in leave-one-out experiments from 0.81 by our previously developed protocol (Oren et al., 2007) to 0.99 AUC, by considering many matrices, adding biological control sequences, and optimizing the matrix refinement algorithm and statistical formalism. Applying the selected refined matrix (pellitrix) to cluster, align, and analyze the pellicle peptides identified residues differentially conserved for the common function of enamel binding.

Availability: Software to apply this protocol is freely available at http://software.compbio.washington.edu/pellitrix/.

Contact: ram@compbio.washington.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A wealth of data has recently been created to describe the human contribution to the proteomic environment of the mouth **[salivary proteome ref]**, particularly the saliva-derived tissues that combat dental caries, including the tooth enamel-coating pellicle (pellicle peptides). 78 pellicle peptide sequences have been characterized, and described as consistent across patients (Siqueira et al., 2007; Vitorino et al., 2007; Vitorino et al., 2008; Siqueira and Oppenheim, 2009). However, current techniques fail to find similarities to explain the shared function of enamel binding (Siqueira and Oppenheim, 2009).

Pellicle peptides bind tooth enamel on one side, and some bind plaque bacteria on the other. As one handle by which oral flora adhere to the tooth, adhesion has been explored for several peptides (Mei et al., 2009), to the extent of redesigning selectivity (Li et al., 2009). Yet mechanisms for enamel adhesion, and protein-hydroxyapatite interactions in general, are still poorly understood.

*To whom correspondence should be addressed.

These 78 enamel binding-specific peptide sequences present a Rosetta stone for understanding protein-hydroxyapatite interactions. We hypothesize that a methodology to identify weak signals in a set of short sequences with shared rare function will be applicable to this set, and conversely that the evolutionary information in this set is sufficient to drive the training of a sequence comparison algorithm to successfully discriminate the enamel-binding pellicle peptides from control sequences.

1.1 Protein-hydroxyapatite interaction mechanisms

A few mechanistic bases of protein-hydroxyapatite interactions are obvious from clues in nature. For example, the many aspartate - serine - serine (DSS) repeats in dentin phosphoprotein (DPP) hint that three residue spacing of carboxylates supports calcium interactions, and the intervening hydroxyls form favorable interactions to phosphates within the surface of forming or mature hydroxyapatite. These data are supported by similar or even enhanced affinities upon mutation to residues bearing the same functional groups but different patterns of side chain length (Yarbrough et al., 2010).

Patterns within the sequences of the 78 enamel pellicle peptides may be used to drive analysis for less obvious aspects of sequence-controlled hydroxyapatite binding. While the greatest mechanistic insights into protein interactions arise from experimentally determined structures, no atomic resolution structures of proteins that physiologically interact with hydroxyapatite are known, except osteocalcin (PDB identifiers 1q8h and 1vzm). No analogous regions are observed between the pellicle peptides or enamel matrix-proteins and the DSS repeats of DPP or γ-carboxy glutamic acids of osteocalcin, so homology-based-inferences from well understood systems are not accessible either. Meanwhile, position-specific information in profile or hidden Markov model comparisons fails to identify sufficient nonobvious relationships to enable meaningful analysis (Supplementary Figure 1).

1.2 Substitution matrix-based peptide similarity

Previously we exploited sequence similarities among phage display peptides that bind to inorganic surfaces to program an amino acid substitution matrix that designed peptides with enhanced binding (Oren et al., 2007). A similar approach may learn the patterns of naturally occurring enamel binding peptides of varying lengths.

The Needleman-Wunsch dynamic programming algorithm compares two protein sequences to find the optimal global alignment with respect to the scoring system being used (Needleman and Wunsch, 1970), which includes a substitution matrix. The Smith-Waterman algorithm is essentially a variant of the Needleman-

Wunsch algorithm with zeroed negative matrix values, such that local alignments are optimized over the global alignments of the Needleman-Wunsch algorithm (Smith and Waterman, 1981).

Application of the Needleman-Wunsch algorithm to evaluate the similarity of a given pair of sequences requires a substitution matrix and two gap penalties for opening or extending gaps in the alignment. These parameters need to be calculated or trained for each particular application. Optimal gap penalties are easy to find using a simple grid search. However, finding the optimal values to score the potential alignment of two sequences is a challenge (Kawashima et al., 2008). The combination of 39 integer values (from -19 to 19) for each of the 210 possible amino acid substitutions in a symmetric matrix, 39^{210}, is too many to calculate (39^{400} if asymmetric). Substitution matrices can be calculated directly by comparative analysis between sets, but some alignments must already be known, and unless the set is large enough to represent the relevant evolutionary relationships, this approach has the propensity to become too specific to the data set to connote biological information (Moult et al., 1997).

One technique to avoid overtraining that performed well for the phage display-derived inorganic surface-binding problem was to exploit a matrix calculated with a widely diverse set of proteins (e.g. BLOSUM62, PAM250) and refine the values to the specific application (Oren et al., 2007). However, it is unlikely for refinement to create an optimal matrix, as coarse scoring functions like matrices have many local maxima and weak trajectories to guide improvement. Therefore here we add sampling of many starting matrices from the diverse substitution matrices in the AAindex (Kawashima et al., 2008; accessed May 2010). Assessing the ability of many matrices to discriminate pellicle peptides from controls allows analysis of matrices that are particularly successful, as the commonality and uniqueness of the proteins used to construct that matrix may inform our understanding of pellicle peptides.

Specific biological sequences that do not bind tooth enamel have not been observed by bench experiment, so we fabricate decoy nonbinder sets as the negative control instances to feed the machine learning algorithm. It is possible that the sequences least likely to bind enamel are the areas of the source protein from which the pellicle peptides are not derived - they are exposed to the same environment that enables enamel interactions and therefore it is likely that they would be observed if they did bind enamel. We derive the decoy control set from these protein regions. Omission by lack of observation is not sufficient evidence to identify absent function (enamel binding), but the ability to help discriminate pellicle peptides would give evidence of differential evolution for not binding enamel, and demonstrate the flexibility of our approach.

A substitution matrix has previously been trained from an existing matrix to discriminate strong, moderate, and weak inorganic surface binding peptides, yet the source sequences were taken from artificial phage display constructs absent of evolutionary information (Oren et al., 2007). The uniform length of the phage display sequences presented a simpler task for machine learning than what is found in nature. Therefore in this work we ask whether a sequence analytic algorithm can select and refine a substitution matrix to discriminate functional peptides of dissimilar lengths from controls, find these peptides from within their source proteins, and identify mechanistic patterns in these natural sequences.

2 MATERIALS AND METHODS

2.1 Data sets

2.1.1 Acquired enamel pellicle peptides. The principal data set used in this work is comprised by the 78 acquired enamel pellicle peptides of 29 salivary proteins (Siqueira and Oppenheim, 2009). The researchers swabbed electron wick paper across the buccal surfaces of the central incisors and first molars in ten patients with good oral health, two hours after dental prophylaxis. Fractionates <10kDa were used because this molecular weight range is known to specify pellicle peptides (Siqueira and Oppenheim, 2009). These were applied through a single in-line high performance liquid chromatography to electrospray ionization tandem mass spectrometer (LC-EPI-MS/MS; Siqueira and Oppenheim, 2009). For use in our bioinformatic experiments, we aligned the peptide sequences, removed 100% redundant sequences, and combined overlapping portions of the same protein. The resulting new pellicle peptide fragment set includes 49 peptides, eight to 36 residues in length (Supplemental Table 1). This set was constructed before any selection of the matrices or training of the algorithm.

2.1.2 Control sequences from the same proteins. To model controls for training and back-testing we used fragments of the 29 proteins not observed within the 78 acquired enamel pellicle peptides. We retrieved random fragments matching the number and length of the peptides, in regions of the 29 proteins not overlapping the pellicle peptide sequences. In cases for which intervening stretches were not abundant or long enough to derive a matching set, we retrieved additional fragments from random other proteins in the set. The resulting decoy nonbinder control set includes 49 peptides, eight to 36 residues in length (Supplemental Table 2).

2.1.3 Additional negative sequences from other proteins. To increase information content for matrix training and to enhance relevance to nonpellicle proteins, we produced additional presumed nonfunctional sets matching the pellicle peptide set in length and quantity. Extracting random parts of any human protein found to be secreted in the saliva produced one set. The resulting salivary proteome-derived control set includes 49 peptides, eight to 36 residues in length (Supplemental Table 3). Additional sets were constructed from random sequences by combination of amino acids selected to mimic the distribution in UniProt accessed May 25th, 2010 (The UniProt Consortium, 2007): A 0.08559150, C 0.01316784, D 0.05286585, E 0.06143265, F 0.04036596, G 0.07075697, H 0.02212168, I 0.05972078, K 0.05282025, L 0.09822349, M 0.02436435, N 0.04181098, P 0.04786202, Q 0.03893130, R 0.05510762, S 0.06789354, T 0.05612079, V 0.06705310, W 0.01323359, Y 0.03055564. The resulting additional UniProt amino acid type distribution-derived fragment set includes 49 peptides, eight to 36 residues in length (Supplemental Table 4).

2.1.4 Training data set combinations. We attempted training with and without each of the additional background sequence sets. The additional negative sequences were not included as controls during assessment; they were only used for training. As long as inclusion of these sequences did not disrupt training we used the matrix trained with them, as this may enhance relevance to proteins outside the 29 enamel pellicle proteins.

2.2 Training protocol

2.2.1 Similarity calculations. The total similarity score (TSS) was used to apply substitution matrices for the specific purpose of differentiating between inorganic surface binding from moderate or weak binding peptides (Oren et al., 2007). The basis is the raw score of a particular matrix applied with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), normalized for the length of each peptide involved and the number of sequences in each set, summed over the input set. The calculation used previously was the mathematical difference between the pellicle to pellicle TSS (TSS.pp) and the pellicle to control TSS (TSS.pc; Oren et al., 2007). Here we explore the utility of also considering the control to control TSS (TSS.cc) and the control to pellicle TSS (TSS.cp), attempted using the difference (TSS.pp + TSS.cc - TSS.pc - TSS.cp) or the quotient ((TSS.pp * TSS.cc) / (TSS.pc * TSS.cp)). We also attempted training with

the difference between the lowest scoring pellicle peptide and the highest scoring control sequence (or the third of each to allow for outliers).

2.2.2 Gap penalties. Gap penalties were trained by selecting the maximal score in an integer grid based search [-16, -1] for the gap open penalty and [-8, -1] for the gap extend penalty. Gap penalties were only trained before altering substitution matrices, and not iteratively, due to their potential volatility during a training process.

2.2.3 Amino acid substitution matrices. We selected starting matrices from 75 amino acid substitution matrices in the AAindex (Kawashima et al., 2008), as described in the Introduction.

2.2.4 Refinement paths. We evaluated three substitution matrix re finement paths. We perturb the starting matrix values by either greedy or modified Monte Carlo trajectories. The greedy algorithm considers all possibilities and then chooses the path that makes the largest magnitude of improvement. We also attempted either local maximization by using the minimum unit of the matrix, or a modified Monte Carlo search for the global maximum by using a random value less than the maximum difference in the matrix, with the decision of keeping each sequential step made after local maximization. We also attempted a set of paths wherein the importance of query versus database amino acid and overall trends in amino acid type were simultaneously examined rather than amino acid type combinations (e.g. the target position being an alanine versus both query and target being alanine), as all sequential combinations of mutating columns, rows, and cells of the matrix.

2.3 Assessment

2.3.1 Leave out one protein experiments. We attempted to discriminate pellicle peptides from control sequences by total similarity score (Figure 1). To assess performance, we performed modified leave-one-out experiments: while scoring a peptide we remove all sequences (pellicle peptides and controls) from the same protein. A normal leave one out experiment involves removing one constituent from the set, training on the rest, scoring the constituent, and repeating for each instance. Here peptides are separated by protein such that in the benchmark the algorithm never learns from and applies information to peptides from the same protein, because sequences in the same protein are likely to contain mutual information.

2.3.2 Statistical metrics. The receiver operating characteristic compares sensitivity (true positives) across all ranges of specificity (true negatives; Figure 2a). The precision recall curve compares the precision at all ranges of recalled selections (Figure 2b). The Matthews correlation coefficient (MCC; Matthews, 1975) measures the correlation of true positives, false positives, false negatives, and true negatives. We demonstrate here that the complexity of a MCC curve informs the capacity for improvement by further training, and identifies the threshold cutoff score that results in the most informative predictions (Figure 2c). Receiver operating characteristic area under the curve (AUC) and one tailed unpaired unequal variance Student's T-test (p values) analyses were used to evaluate significance for each experiment.

2.3.3 Amino acid content calculation. For each amino acid in a peptide, a score was calculated as the difference in abundance for that amino acid type between pellicle peptides and control sequences. This served as a control for the predominance of amino acid type on enamel binding, to ask the question of whether sequential orientation (sequence) matters.

2.4 Application to full protein sequences

We recapture functional regions from full protein sequences by generating a score for each residue in the protein, considering the surrounding region. In the sliding window technique, all possible continuous segments of a particular length are taken from the sequence. We applied the sliding window approach for each unique length of pellicle peptides. For this problem, it is uncertain whether it would be better to choose segments of one particular length, or to exhaustively create segments of all pellicle peptide lengths. Even then, it is not known how to consider the similarity scores for the various segments to which each particular residue contributes. For both a single window length (the median of all peptide lengths) and exhaustive enumeration of the lengths, we evaluated the application of the mean of the similarity scores for overlying segments and the maximum score for each. Maintaining consistent fragment lengths between query and comparison sets avoids a difficult normalization problem. We compared the predictive ability of residue scores to recapture the pellicle peptides from the entire protein sequences, again leaving out all pellicle peptides and control sequences derived from each protein as it was evaluated (Figure 3).

2.5 Cluster analysis

To study the sequence patterns in pellicle peptides, we derived sequence clusters by analyzing a matrix comparison of each enamel pellicle peptides against all others using the best selected and refined matrix (pellitrix). We filtered the resulting similarity scores by the threshold cutoff that gave maximum information in the benchmark according to the MCC plot (Figure 2c). We then input the suprathreshold similarity values as clustering force vectors. Subcluster networks were identified from 2D depictions of the contiguous network, and constructed into a multiple sequence alignment (MSA) using pellitrix (Figure 4, bottom). The importance of each cluster alignment column relative to the entire set was estimated as the sum of the pellitrix values for all possible residue pairs within the column divided by the number of columns (Figure 4, blue bars). The importance of each residue among its cluster was similarly estimated as the sum of pellitrix values for comparison to other residues in the same column (Figure 4, white to green coloring of MSA, bold letters mark the top 25th percentile of scores).

2.6 Software

All code was written in the Python programming language version 2.5. The Needleman-Wunsch algorithm used was ggsearch35 within fasta-35.4.11. Multiple sequence alignments were generated by CLUSTALW (Larkin et al., 2007). Statistical tools employed in the assessment were written locally and extensively checked against both SPSS and STATA. Depiction of assessment in Figures 1-3,5 was performed with gnuplot. Clustering and depiction of the network in Figure 4 were performed with Cytoscape.

3 RESULTS

3.1 Selected and refined peptide discrimination

We demonstrate the ability of the matrix sampling and refinement protocol to optimize performance in discriminating pellicle from control sequences (Figures 1 and 2). We obtained marked improvement for two highly different substitution matrices by three rigorous statistical metrics (Figure 2). The β-3D-Ali matrix (MEHP950102) was selected for optimal peptide discrimination, and refined from 0.92 ($p=3.44*10^{-15}$) to 0.99 AUC ($p=3.43*10^{-26}$). We present the optimized substitution matrix and values changed during training in Supplemental Table 5. The PAM250 matrix (DAYM780301) was refined from 0.76 ($p=5.00*10^{-7}$) to 0.84 AUC ($p=4.53*10^{-10}$). We extended the refined β-3D-Ali matrix to estimate the likelihood of any single residue binding tooth enamel and calculate recovery of the pellicle peptides (0.75 AUC; Figure 3). Finally, we used the refined selected matrix to analyze similarities between the pellicle peptides for possible mechanistic bases of enamel interactions (Figure 4).

3.2 Matrix sampling

Sampling within AAindex (Kawashima et al., 2008) identified matrices that discriminate pellicle peptides from control sequences. Performance ranged from discriminating the majority of pellicle peptides, to not discriminating any significantly. Figure 1 shows the distribution of scores for pellicle peptides and control sequences for the top twenty matrices and the amino acid content score (parameters and evaluation for all matrices in Supplemental Table 6). The matrix that most accurately separated pellicle peptides from controls, the β-3D-Ali matrix, was used for further analysis. The PAM250 matrix was run through the same analyses for comparison.

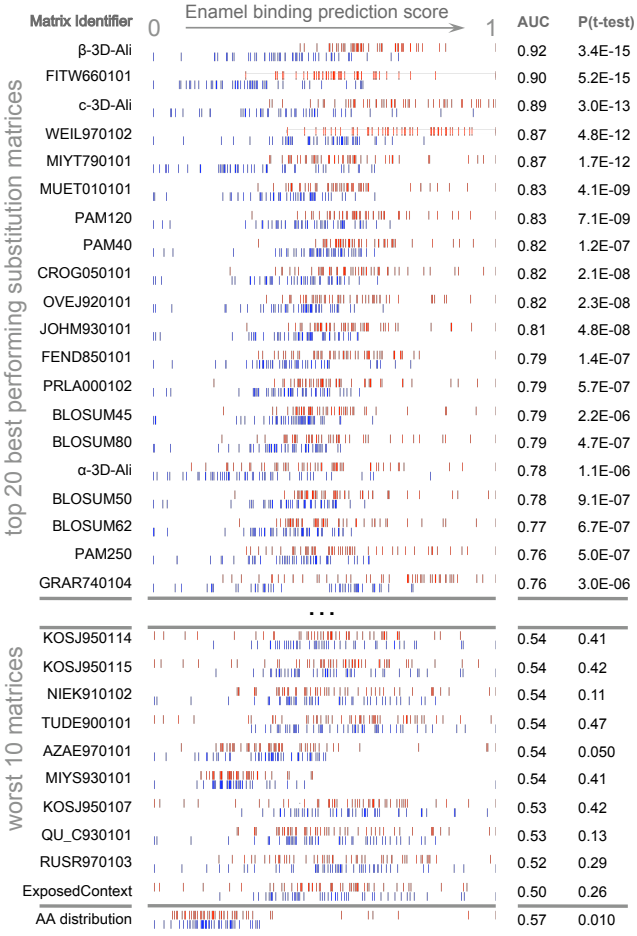

Fig. 1. Discrimination of enamel pellicle peptides. The selection of 49 pellicle peptides (red) from 49 control sequences (blue) by the top 20 and worst 10 performing substitution matrices is shown. Each row corresponds to one matrix, for which normalized scores are plotted for each pellicle and control sequence. Better discrimination is seen at top, with pellicle peptides assigned higher scores (red to the right) and controls assigned lower scores (blue to the left). Nonoverlap for the profiles of pellicle and control markers would indicate perfect discrimination. Most matrices discriminate more accurately than amino acid content (at bottom), which demonstrates the importance of the sequential and spatial arrangement of residues. The ability of the algorithm to separate pellicle peptides from decoy control sequences that are not observed to bind enamel verifies that the controls do not bind and are actually evolved to not bind enamel.

3.3 Matrix refinement

The refinement protocols improved performance on the task of sorting pellicle peptides from control sequences for both the PAM250 and β-3D-Ali matrices (Figure 2, Supplemental Table 7).

3.3.1 Similarity calculations. All three subtraction-based similarity calculations resulted in improvement for the PAM250 and β-3D-Ali matrices, whereas the quotient based similarity calculation did not. The most significant improvements in the matrices arose consistently from including the relation of control sequences to themselves and to the pellicle peptides in the total similarity score (TSS.pp + TSS.cc - TSS.pc - TSS.cp).

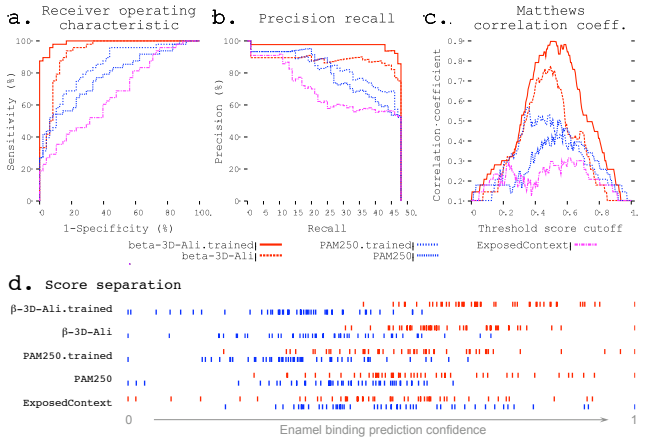

Fig. 2. Refinement improves enamel pellicle peptide discrimination. The β-3D-Ali.trained (solid red line, see key below panels a-c) and PAM250.trained (blue coarsely dashed line) matrices demonstrate increased predictive ability across three rigorous metrics from the β-3D-Ali (red dashed line) and PAM250 (blue thinly dashed line) matrices, respectively. Comparison is given to the worst performing matrix (ExposedContext = KOSJ950113). a. receiver operating characteristic curve. b. precision recall curve. c. Matthews correlation coefficient (MCC) curve. The complexity of each MCC curve informs the capacity for improvement: the untrained matrices both show a large local minimum, lost with improvement in the correlate trained curve. d. Score distributions (as in Figure 1) show greater separation of pellicle peptide (red) and control sequence (blue) scores after training.

3.3.2 Refinement paths. The best and most consistent matrix refinement protocol was achieved by a greedy path, exhausting improvements from changing all values in each column together, exhausting improvements similarly in the rows, then optimizing whole columns and rows with the modified Monte Carlo search. The greedy algorithm uses more processor time than a random or Monte Carlo path, as both positive and negative trajectories for each position must be considered before progressing to the next step. Nonetheless, each training combination was able to reach completion in less than one day on a single 4.8 GHz processor.

The order of starting permutations with matrix row (query amino acid type) or column (pellicle / control amino acid type) affected the performance of the matrix. Only a few random paths starting with rows increased performance, while many training conditions improved accuracy when starting with columns. Adding Monte Carlo perturbations of columns and then rows as a last set of steps after the described greedy path improved performance in nearly all cases, whereas Monte Carlo perturbations of the cells never did.

3.3.3 Training data set combinations. Inclusion of the additional background sequences into the controls improved the discriminatory performance of both PAM250 and β-3D-Ali matrices slightly (AUC ~1%) with statistical significance ($p<0.01$).

3.3.4 Preferences of the matrix. Pairwise amino acid substitution scores for the identical residue and for the mean of all possible residue substitutes indicate the importance of matching each particular amino acid type (Supplemental Table 8). For example, it is preferred that glutamic acid is aligned with another glutamic acid (score = 2.00), but self match is penalized for leucine (-1.40) and arginine (-2.00).

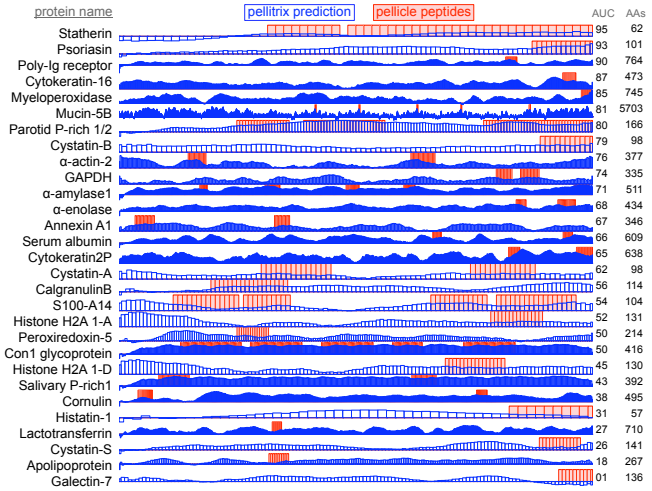

Fig. 3. Enamel pellicle peptide recapture from complete proteins. Predictions of enamel affinity by the refined β-3D-Ali matrix for each residue are plotted in blue for each enamel pellicle protein. Scores represent the mean of the similarity scores between all peptides derived from other proteins (leave one out experiment) and all possible overlapping sequence fragments of lengths matching the pellicle peptides (sliding window fragmentation). Experimentally derived pellicle peptides are shown as red blocks. Overlap of high blue bars with the red blocks denotes recapture of pellicle peptides from the parent protein. Protein length (AAs) and recapture accuracy (AUC) are listed at far right.

3.4 Protein binding region recapture

Accuracy of pellicle peptide recapture from full protein sequence depended largely on the formalism. Comparing protein segments of length corresponding to the median pellicle peptide length (fourteen residues) achieved 0.75 AUC for the mean score, and 0.54 AUC for the maximum. A similar difference was found for enumerating all lengths: 0.69 AUC for the mean and 0.54 AUC for the maximum. A caveat to this experiment should be noted: while the leave-one-out design avoids comparing peptides directly to any part of their source protein sequence, the information trained into the matrix in the selection and refinement steps cannot be removed and so biases this experiment. Without training, the β-3D-Ali matrix achieves 0.73 AUC using the mean of the multiple sliding window, again the highest of all matrices (Supplemental Table 7).

3.5 Pellicle peptide sequence cluster analysis

Application of the selected and refined β-3D-Ali matrix to compare all 78 pellicle peptides to each other resulted in an informative network of context-specific sequence similarities (Figure 4). Multiple sequence alignments constructed with the matrix illustrate in each column the amino acid types that can function similarly within the specific context of protein-hydroxyapatite interactions.

4 DISCUSSION

4.1 Advancement in biomineralization

The ability of many amino acid substitution matrices to accurately discriminate enamel pellicle peptides from control sequences (Figure 1) demonstrates the presence of mechanistic information encoded in the sequences. The common function of enamel hydroxyapatite binding is the most likely explanation for the hidden sequence patterns that enable discrimination. Cluster analysis (Figure 4) suggests peptide groups likely to share similar hydroxyapatite binding mechanisms, and sequence patterns to facilitate those mechanisms. The refined selected matrix can be used to analyze sequences for likelihood of contributing protein-hydroxyapatite interactions in peptides (Figure 2), whole protein sequences (Figure 3), and to design novel peptides with tunable affinity.

Novel biocompatible peptides may be designed with controllable affinities, used as a supplementary pellicle coat to control the attachment of oral flora, or as an adjuvant vehicle for controllable delivery of saliva replacements such as anticariogenic antibiotics (He et al., 2007), remineralizing agents (Yarbrough et al., 2010), buffers, or lysozyme (Hannig et al., 2009).

4.2 Advancement in bioinformatics

The improvements we introduce to our protocol to develop peptide similarity detection tools increased the final trained matrix discriminatory ability from 0.81 AUC with the old protocol to an unprecedented 0.99 AUC with the new protocol. MCC plot analyses indicate the training of this matrix has approached saturation (Figure 2c). The most significant improvements arose from sampling many starting substitution matrices, including the total similarity scores of peptides to controls and vice versa, and optimizing after greedy refinement. This approach may be generalizable to learning patterns in any group of functional peptides, and is available as software for use and development.

4.3 Matrix sampling

Outstanding discriminatory performance by a matrix may indicate relevance to the context for which the matrix was calculated. A matrix that gives specific structural context performed best, matrices built for general protein sequence comparison exhibited intermediate performance, and matrices built for intuitively irrelevant contexts performed no better than random.

The best performing matrix, AAindex name MEHP950102, was calculated from the alignment of β-strands in 38 3D-Ali protein structure families (Mehta et al., 1995; AUC=0.925, $p=3.44*10^{-15}$; Figure 1). The relevance of the β-3D-Ali matrix to these peptides may be the similarity between extended conformations of beta strand and polyproline type II, which seems to be the secondary structure for many hydroxyapatite-interacting proteins (Le et al., 2006; Jin et al., 2009; Lyngstadaas et al., 2009). The matrix derived from coil portions of the 3D-Ali protein set (non α-helix or β-strand) was the third best performing matrix (MEHP950103, AUC=0.89, $p=3.0*10^{-13}$), which may suggest that coil secondary structure is nearly as relevant as β-strand or simply that α-helices are not. Structural alignments, or the 38 3D-Ali protein structure

families, may be particularly relevant to this group of peptides, as even the α-helix derivation performs 16th best (MEHP950101, AUC=0.78, $p=1.06*10^{-6}$).

The ability of substitution matrices to discriminate pellicle peptides from decoy sequences presumed to not bind enamel demonstrates unique patterns within these sequences. The ability to sort the nonbinding regions gives evidence for divergent evolution of these sequence regions from the source proteins, suggesting that the regions not observed in the pellicle may have evolved to specifically not bind enamel. Significantly worse discrimination by amino acid content (Figure 1, last row) demonstrates that sequential arrangement of amino acids is important for enamel binding.

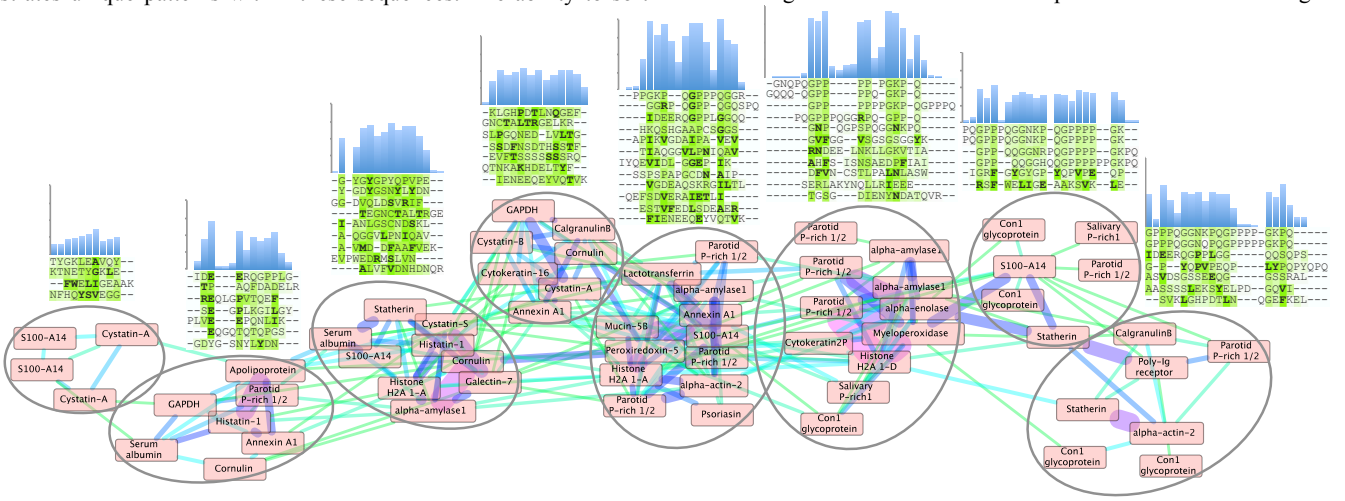

Fig. 4. Cluster analysis of enamel pellicle peptide sequences identifies enamel-binding residues. A network of pairwise pellicle peptide sequence alignments was clustered, with edge weights given by alignment scores (increasing edge width and green to violet color) and threshold cutoff corresponding to that of the maximum Matthews correlation coefficient in Figure 2c. Protein names that appear multiple times indicate alternate peptides derived from the same protein. Multiple sequence alignments for each cluster are analyzed: the estimated importance of each alignment column relative to the entire set is shown as blue bars; and that of each residue to its cluster is shown in white to green; bold font marks the top 25th percentile of scores. This analysis identifies the mutual information in the pellicle peptides that encodes the function they have in common: enamel binding. All comparisons were calculated with pellitrix.

4.4 Matrix refinement

The matrix refinement results demonstrate success, but a limit to the protocol: refining the PAM250 matrix never achieves the accuracy of the existing β-3D-Ali matrix (Figure 2). This observation highlights the importance of selecting from a diverse set of substitution matrices, as it increases the sample space with dramatic efficiency.

4.5 Protein binding region recapture

Comparison of the score profiles to experimentally derived pellicle peptide sequences in Figure 3 shows that we have successfully modeled a significant subset of enamel binding mechanisms, and predicted other regions not yet consistently observed within the enamel pellicle. High scoring regions at locations where pellicle peptides have not been measured are predictions of areas that may bind enamel, for example the amino terminal regions of α-actin 2, cystatin-A, S100-A14, histone H2As 1-A and 1-D (Figure 3).

Recapture of pellicle peptides from whole protein sequences is better than average for 21 of 29 proteins, with a by-residue AUC of 0.75 across all proteins. Poor performance of the PAM250 matrix (AUC=0.31) highlights the uniqueness of sequence traits within these peptides of such rare function, and therefore the importance of using similarity matrices with maximal relevance to any particular group of proteins under study. This analysis demonstrates novel ability to understand, predict, and potentially design protein to hydroxyapatite interactions.

4.6 Pellicle peptide sequence cluster analysis

Each cluster displays trends in the multiple sequence alignments (Figure 4). Generally we observe tolerance for swapping residue identity but maintenance of chemical moieties: adjacent carboxyl or amide residues may facilitate calcium interactions (Horst and Samudrala, 2010), and alternating hydroxyl moieties may mediate phosphate interactions. Stretches of prolines may stabilize extended conformations, facilitating surface interactions. Proline almost never aligns with glutamine, suggesting non-interchangeable roles for the two most abundant residues in these peptides. Residue types most commonly involved in protein-mediated catalysis (in order: EKDHRSTCYNQAFGMLWIVP; Wang et al., 2008) are seldom aligned with identical amino acid types in these clusters. Generally these patterns suggest greater structural conservation with variance allowed for chemical interactions; this description fits the manifold presentation of calcium and phosphate on the enamel hydroxyapatite surface.

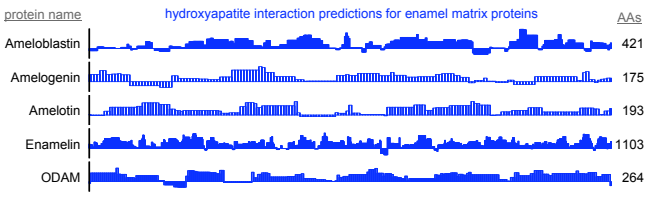

Fig. 5. By-residue likelihood of hydroxyapatite interactions for enamel matrix proteins. The refined selected matrix was applied to find the similarity of the region surrounding each residue to the enamel pellicle peptides. Scores are normalized to the highest (1.0) and lowest (0.5) scores observed for all peptides and control sequences. Length of proteins shown at right.

High scoring regions likely correspond to functional areas that interact with mature or maturing enamel. Low scoring areas may have carry out functions not consistent with mature enamel, such as hydroxyapatite nucleation and endoprotease cleavage.

4.7 Application to enamel matrix proteins

To demonstrate the utility of our approach as a scientific tool, we applied the refined selected matrix and data set to the five known enamel matrix proteins (Figure 5). High scoring regions likely correspond to functional areas that interact with mature or maturing enamel. Low scoring areas may carry out functions not consistent with mature enamel, such as hydroxyapatite nucleation and endoprotease cleavage. These data support design of peptides for nanotechnology, and mutation experiments to develop a mechanistic understanding of enamel development.

4.8 Matthews correlation coefficient plot

The complexity of each MCC curve informs the capacity for improvement: the untrained matrices each show one large local minimum, which is lost with improvement in the correlate trained curve (Figure 2c). The trained matrix MCC curves are more broad and show decreased complexity, suggesting that these are near the end of the respective training paths. While not communicated directly by standard PR nor ROC plots, the MCC plot directly shows the cutoff value with the most discriminative ability and informational content.

4.9 Comparison to previous work

We extended the methodology for sequence-based prediction of inorganic binding peptides to naturally occurring peptides observed in the human enamel pellicle. We invoked a powerful step to efficiently sample the amino acid substitution matrix space by selecting from an existing diverse database of these matrices. The profile of accuracies for these matrices informs relevance of the pellicle peptides to the matrix derivation. Analysis might be strengthened by comparison to other function prediction methods (e.g. Wang et al., 2008), differential aspects of calcium phosphate nucleation, hydroxyapatite maturation, and hydroxyapatite binding.

As seen previously for artificial phage display derived inorganic surface binding peptides (Oren et al., 2007), amino acid substitution matrix methods can learn contextual patterns, now including natural salivary enamel pellicle peptides. Further understanding and specificity will be gained by considering endoprotease cleavage sites, post-translation processing, and evolutionary conservation among the by-residue pellicle similarity scores. While no other tool known to us can dissect sequences with such unique function, the analysis presented here demonstrates the ability to understand, predict, and therefore design protein-hydroxyapatite interactions.

5 CONCLUSIONS

We demonstrated that enamel pellicle peptides contain subtle sequence similarities that encode hydroxyapatite binding mechanisms. With various experimental and algorithmic improvements, our substitution matrix-based peptide comparison protocol was able to represent the pellicle peptide similarities in an amino acid substitution matrix (pellitrix) that discriminates the pellicle peptides from sequences with near perfect accuracy (0.99 AUC). We showed that pellitrix can recapture the peptides from the source protein sequences, and that this can be applied as a tool to predict hydroxyapatite interaction regions of relevant proteins. Analysis of relationships between the pellicle peptide sequences indicates that adjacent carboxyl or amide residues facilitate calcium interactions, that alternating hydroxyl moieties mediate phosphate interactions, and that stretches of prolines stabilize extended conformations. This protocol was built as a freely available software suite to learn similarities in any set of peptides, for bioengineering design and analysis of biologic function.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grants F30-DE017522 (J.A.H.), K08-DE022377 (O.V.H.), and DP1-OD006779 to (R.S.).

Conflict of Interest: none declared.

REFERENCES

Hannig,C. et al. (2011) Targeted immobilisation of lysozyme in the enamel pellicle from different solutions. *Clin Oral Investig*, 2011, 15, 65-73.

He,J. et al. (2007) Novel synthetic antimicrobial peptides against Streptococcus mutans. *Antimicrob Agents Chemother*, 51, 1351-1358.

Horst,J.A. and Samudrala,R. (2010) A protein sequence meta-functional signature for calcium binding residue prediction. *Pattern Recognit Lett*, 31, 2103–2112.

Hu,S. et al. (2008) Salivary Proteomics for Oral Cancer Biomarker Discovery. *Clin Cancer Res*, 14, 6246-6252.

Jin,T. et al. (2009) Elongated polyproline motifs facilitate enamel evolution through matrix subunit compaction. *PLoS Biol*, 7, e1000262.

Kawashima,S. et al. (2008) AAindex: amino acid index database, progress report. *Nucleic Acids Res*, 36, D202-D205.

Larkin,M.A. et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.

Le,T.Q. et al. (2006) Comparative calcium binding of leucine-rich amelogenin peptide and full-length amelogenin. *Eur J Oral Sci*, 114 Suppl 1, 320-329.

Li,M.Y. et al. (2009) Effect of a dentifrice containing the peptide of streptococcal antigen I/II on the adherence of mutans streptococcus. *Arch Oral Biol*, 54, 1068-1073.

Lyngstadaas,S.P. et al. (2009) Enamel matrix proteins; old molecules for new applications. *Orthod Craniofac Res*, 12, 243-253.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 405, 442-451.

Mehta,P.K. et al. (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci*, 4, 2517-2525.

Mei,L. et al. (2009) Poisson analysis of streptococcal bond-strengthening on saliva-coated enamel. *J Dent Res*, 88, 841-845.

Moult,J. et al. (1997) Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins*, Suppl 1:2-6.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48, 443-453.

Oren,E.E. et al. (2007) A novel knowledge-based approach for designing inorganic binding peptides. *Bioinformatics*, 23, 2816-2822.

Siqueira,W.L. et al. (2007) Identification of protein components in in vivo human acquired enamel pellicle using LC-ESI-MS/MS. *J Proteome Res*, 6, 2152-2160.

Siqueira,W.L. and Oppenheim,F.G. (2009) Small molecular weight proteins/peptides present in the in vivo formed human acquired enamel pellicle. *Arch Oral Biol*, 54, 437-444.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, 147, 195-197.

The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 35, D193-D197.

Vitorino,R. et al. (2007) Peptidomic analysis of human acquired enamel pellicle. *Biomed Chromatogr*, 21, 1107-1117.

Vitorino,R. et al. (2008) Peptide profile of human acquired enamel pellicle using MALDI tandem MS. *J Sep Sci*, 31, 523-537.

Wang,K. et al. (2008) Protein meta-functional signatures from combining sequence, structure, evolution and amino acid property information. *PLoS Comp Bio*, 4, e1000181.

Yarbrough,D.K. et al. (2010) Specific binding and mineralization of calcified surfaces by small peptides. *Calcif Tissue Int*, 86, 58-66.