

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Veith, PD;Nor Muhammad, NA;Dashper, SG;Likić, VA;Gorasia, DG;Chen, D;Byrne, SJ;Catmull, DV;Reynolds, EC

Title:

Protein substrates of a novel secretion system are numerous in the Bacteroidetes phylum and have in common a cleavable C-terminal secretion signal, extensive post-translational modification, and cell-surface attachment

Date:

2013-10-04

Citation:

Veith, P. D., Nor Muhammad, N. A., Dashper, S. G., Likić, V. A., Gorasia, D. G., Chen, D., Byrne, S. J., Catmull, D. V. & Reynolds, E. C. (2013). Protein substrates of a novel secretion system are numerous in the Bacteroidetes phylum and have in common a cleavable C-terminal secretion signal, extensive post-translational modification, and cell-surface attachment. *Journal of Proteome Research*, 12 (10), pp.4449-4461. <https://doi.org/10.1021/pr400487b>.

Persistent Link:

<https://hdl.handle.net/11343/158070>

Revised Manuscript PR-2013-00487b

Protein substrates of a novel secretion system are numerous in the Bacteroidetes phylum and have in common a cleavable C-terminal secretion signal, extensive post-translational modification and cell surface attachment

Paul D. Veith¹, Nor A. Nor Muhammad¹, Stuart G. Dashper¹, Vladimir A. Likić², Dhana G. Gorasia¹, Dina Chen¹, Samantha J. Byrne¹, Deanne V. Catmull¹ and Eric C. Reynolds^{1*}.

¹Oral Health CRC, Melbourne Dental School, Bio21 Institute, The University of Melbourne, Australia.

²Bio21 Institute, The University of Melbourne, Australia.

Running title: Protein secretion in Bacteroidetes

*To whom correspondence should be addressed: Professor Eric C. Reynolds, Melbourne Dental School, University of Melbourne, 720 Swanston Street, Victoria, 3010, Australia.
Email: e.reynolds@unimelb.edu.au. Phone +61 3 9341 1547; Fax +61 3 9341 1597

ABSTRACT

The secretion of certain proteins in *Porphyromonas gingivalis* is dependent on a C-terminal domain (CTD). After secretion the CTD is cleaved prior to extensive modification of the mature protein, probably with lipopolysaccharide, therefore enabling attachment to the cell surface. In this study bioinformatic analyses of the CTD demonstrated the presence of three conserved sequence motifs. These motifs were used to construct Hidden Markov Models (HMM) that predicted 663 CTD-containing proteins in 21 fully sequenced species of the *Bacteroidetes* phylum while no CTD-containing proteins were predicted in species outside this phylum. Further HMM searching of *Cytophaga hutchinsonii* led to a total of 171 predicted CTD proteins in that organism alone. Proteomic analyses of membrane fractions and culture fluid derived from *P. gingivalis* and four other species containing predicted CTDs (*Parabacteroides distasonis*, *Prevotella intermedia*, *Tannerella forsythia* and *C. hutchinsonii*) demonstrated that membrane localisation, extensive post-translational modification and CTD-cleavage were conserved features of the secretion system. The CTD cleavage site of ten different proteins from three different species was determined and found to be similar to the cleavage site previously determined in *P. gingivalis* suggesting that homologs of the C-terminal signal peptidase (PG0026) are responsible for the cleavage in these species.

Key words: novel protein secretion system; Bacteroidetes; C-terminal signal peptidase.

INTRODUCTION

The *Bacteroidetes* is a major bacterial phylum whose members are prevalent in a diverse range of habitats and includes species that are of considerable medical, veterinary and environmental importance ¹. In soils they are estimated to represent about 5% of obtained rDNA sequences, ² while in the human gut ³ and the world's oceans ⁴ they account for approximately one quarter of the sequences obtained. In the gut, *Bacteroidetes* are generally considered to be associated with health ^{3,5}, however in the oral cavity, *Bacteroidetes* species are associated with chronic periodontitis in both humans and animals ^{6,7} and are also prevalent in endodontic infections ⁸. Several fish pathogens are *Bacteroidetes spp.* for example *Flavobacterium psychrophilum* ⁹, while other marine *Bacteroidetes* are important contributors to carbon cycling due to their ability to degrade high molecular weight components of dissolved organic carbon such as protein and chitin ^{10,11}.

Arguably, the most studied *Bacteroidetes* species is *Porphyromonas gingivalis*, recognised as the most important pathogen associated with chronic periodontitis ¹². The major virulence factors of this organism are surface-associated cysteine proteinases called gingipains ¹³. The three gingipains (RgpA, RgpB and Kgp) have in common a conserved C-terminal domain of approximately 70 amino acid residues which is referred to as the CTD. The CTD was first postulated to be involved in anchoring these virulence factors to the cell surface over a decade ago ¹⁴. The membrane form of RgpB was found to be extensively modified, containing 30% carbohydrate, while the soluble form lacking the CTD was not surface attached and was not modified ¹⁵. In a proteomic study of the *P. gingivalis* outer membrane (OM), this link was strengthened by the identification of multiple CTD containing proteins that migrated on two-dimensional gel electrophoresis as diffuse spots of considerably elevated molecular weight that reacted with a monoclonal antibody (MAb-1B5) which recognised a Man α 1-2 Man α 1-phosphate epitope found on the anionic polysaccharide

of A-LPS¹⁵⁻¹⁹. Alignment of the CTD sequences of these proteins revealed that the similarity was confined to three motifs separated by poorly conserved regions of variable length¹⁹. Extensive BLAST searching of the CTD sequences led to the identification of 34 proteins containing the CTD within the *P. gingivalis* W83 genome as well many CTD proteins in other *Bacteroidetes* spp.²⁰. Protein motif searches further established the widespread presence of the CTD in this phylum²¹.

Subsequent to the early *P. gingivalis* studies, other CTD proteins from other *Bacteroidetes* spp. have now been identified; including virulence factors of a fish pathogen⁹, cellulose degrading enzymes of *Cytophaga hutchinsonii*²², a xylanase from *Rhodothermus marinus*²³ and a family of proteins from *Parabacteroides distasonis*²⁴. Some of these proteins were also found to be cell-associated or OM-attached and glycosylated. Note that the CTD is often referred to as “D5” in these publications. More recently, a proteomic study of the *Tannerella forsythia* OM also demonstrated the presence of multiple CTD proteins that were glycosylated and attached to the OM²⁵.

While the idea of the CTD being linked to the cell surface-attachment of proteins is evident in the various species mentioned above, the implication that the CTD may also represent a secretion signal enabling the CTD protein to be translocated across the OM has been mainly addressed by *P. gingivalis* researchers; as is the notion that cell-surface attachment may be attained by conjugation of the CTD protein to a particular form of lipopolysaccharide (A-LPS)¹⁵⁻¹⁹. Removal of the CTD from RgpB in *P. gingivalis* leads to the accumulation of RgpB in the periplasm²⁰ and fusing of a CTD to the C-terminus of GFP resulted in the secretion, modification and attachment of GFP to the surface of *P. gingivalis*²⁶ demonstrating a role for the CTD in secretion.

A number of *P. gingivalis* genes have been reported to be required for the secretion of the gingipains²⁷⁻³⁰ or CTD proteins in general^{16, 17, 31}. Many of these genes were chosen for

examination on the basis of their distribution amongst species within the *Bacteroidetes* phylum, as while many sequenced species within the phylum appear to contain CTD proteins and known components of the secretion system, most species within the *Bacteroides* genus do not. Hence a list of proteins common to *P. gingivalis* and *C. hutchinsonii* but absent in *Bacteroides thetaiotaomicron* proved to include many proteins essential for gingipain secretion²⁹.

Electron microscopy of *P. gingivalis* reveals an electron dense layer external to the OM. This electron dense surface layer (EDSL) is dependent on the secretion system since mutants lacking components of the secretion system also lack the EDSL^{16, 17}. Moreover, the gingipains were localised to the surface layer by immunogold labelling and mutants lacking RgpA or Kgp have a reduced EDSL suggesting that the EDSL may be comprised of secreted CTD proteins, particularly the gingipains¹⁶. Similarly, the surface layer of *T. forsythia* consists of the glycosylated surface layer proteins A and B³² which are also CTD proteins²⁵.

Recently, we discovered that the *P. gingivalis* CTD is cleaved during secretion prior to post-translational modification of the matured CTD protein and that the CTD itself is not the site of modification and cell-surface attachment, but rather functions as a C-terminal secretion and attachment signal¹⁷. The PG0026 (PorU) protein was implicated as the C-terminal signal peptidase as PG0026 deletion mutants and mutants lacking the predicted catalytic Cys residue accumulated RgpB on the cell surface with its CTD intact.

In this study we used bioinformatic approaches to define the C-terminal signal of CTD-family proteins and used Hidden Markov Models (HMMs) to predict 663 CTD proteins across 21 *Bacteroidetes* species. We demonstrate in four representative species of the *Bacteroidetes* phylum, namely *P. gingivalis*, *P. distasonis*, *Prevotella intermedia* and *Cytophaga hutchinsonii* that each exhibit highly abundant, modified CTD proteins in the outer membrane fraction. As this is a unique and defining feature of secretion in the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Bacteroidetes phylum, we propose to name it “CoAtSS” (Combined Attachment and Secretion System) which describes its distinguishing feature of attaching its substrates to the cell-surface.

METHODS

Computer Hardware and Software

Bioinformatic analyses were performed on a Core i7 Desktop running Ubuntu 10.04 (<http://www.ubuntu.com>) as a virtual machine and Windows XP (<http://www.microsoft.com>) as the host operating system. Both operating systems had been updated with the latest patches. The virtual machine software used was VMware Player 4.0.2. (<http://www.vmware.com>). The motif analyses tools installed and used were the Gapped Local Alignment of Motifs (GLAM2)³³ and Multiple Em for Motif Elicitation (MEME)³⁴ from The MEME Suite, a package of Motif-based sequence analysis tools (<http://meme.sdsc.edu/>)³⁵.

GLAM2 Analyses

Twenty three CTD proteins from *P. gingivalis* W83 and eleven from *T. forsythia* ATCC 43037 that were previously identified by proteomic studies were selected for these analyses²⁰. The protein sequences of *P. gingivalis* CTD proteins were obtained from NCBI Genbank (<http://www.ncbi.nlm.nih.gov/genbank/>)³⁶ and the protein sequences of *T. forsythia* CTD proteins were from ORALGEN (<http://www.oralgen.lanl.gov>). These thirty four protein sequences [PG1844, TF2339, TF2592, TF2998 (and for the remainder see Fig. 2)] were analysed using GLAM2. Parameters used were 100 for -b: maximum number of aligned columns and default for the remaining parameters. The thirty four sequences were then truncated to their C-terminal 80 amino acid residues and analysed using the global sequence alignment Needleman-Wunsch algorithm (EMBOSS Needle)^{37,38} to check for sequence similarity. Four sequences (from PG1844, TF2339, TF2592 and TF2998) were found to have extensive sequence identity to other sequences and were discarded from further analyses to

reduce redundancy. The remaining thirty CTD protein sequences were analysed using GLAM2 again with the same parameters as before.

MEME Analyses

MEME was used to quantitate individual motifs within the conserved CTD region found by GLAM2. MEME was run in one per sequence mode with the E-value threshold set at 1e-3 and the remaining parameters at their default values. MEME identified two motifs, Motif B and Motif D²⁰. The dominant Motif B sequence was then removed from the CTD protein sequences and these sequences analysed with MEME which resulted in the identification of Motif E. This was achieved by truncating the amino acid sequences from eighty to thirty amino acids in length followed by MEME analysis using the same parameters as above.

Hidden Markov Modelling

HMMER software³⁹ implements methods using probabilistic models named profile Hidden Markov Models (HMM). The multiple sequence alignment of the thirty CTD proteins needed by HMMER as input was produced from the GLAM2 output. The multiple sequence alignment was truncated into two multiple sequence alignment files; the first containing Motif B and the second containing Motif D+E (combined Motif D and Motif E). The variable region between Motif B and Motif D was discarded. The two multiple sequence alignment files were then inputted into HMMER for profiling of Motif B and Motif D+E. The two HMM profiles produced were called Motif B profile and Motif D+E profile.

The protein sequence databases, Swiss-Prot and TrEMBL were obtained from Uniprot release 2011_08 (<http://www.uniprot.org>)^{40, 41}. The Swiss-Prot database was processed into two smaller databases. The first protein sequence database contained the last 80 amino acids of each protein sequence (Swissprot_80aa) and the second database contained the last 30

amino acids of each protein sequence (Swissprot_30aa). The TrEMBL database was processed the same way (Trembl_80aa & Trembl_30aa). The truncation of these databases served to provide the important constraint of C-terminal location. Using hmmpfam, a part of the HMMER2 software package, Motif B profile was searched against Swissprot_80aa and Trembl_80aa database and Motif D+E profile was searched against Swissprot_30aa and Trembl_30aa database. The E-value chosen was $1e-3$ for each HMM profile when searching against the truncated databases for a combined e-value of $1e-6$. This E-value was chosen because it produced less than 1.5% false positives based on evaluation using decoy strategies (see below). Hmmpfam hits towards Motif B and Motif D+E profiles were processed using Microsoft Access to generate the list of proteins that matched Motif B and Motif D+E profiles.

HMM searching was also performed on the proteome sets of fully sequenced *Bacteroidetes* organisms. A total of twenty one *Bacteroidetes* proteomes were downloaded, nineteen from NCBI Genbank and two from ORALGEN (Fig. 3). They were combined to generate one *Bacteroidetes* protein sequence database. Again, two smaller databases were created, *Bacteroidetes*_80aa and *Bacteroidetes*_30aa. Motif B profile was searched against *Bacteroidetes*_80aa database while Motif D+E was searched against *Bacteroidetes*_30aa database. However, the E-value chosen this time was 0.1 for each HMM profile for a combined E-value of $1e-2$. Hmmpfam hits towards Motif B and Motif D+E profiles were processed using Microsoft Access to generate the list of proteins that matched Motif B and Motif D+E profiles. A phylogenetic tree of the 21 *Bacteroidetes* species was constructed using taxonomic data from NCBI Taxonomy Common Tree (<http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi>)^{42, 43} and PHYLIP (the PHYLogeny Inference Package)⁴⁴.

The number of random hits by the HMM methodology towards the *Bacteroidetes* database was estimated using two different decoy strategies. The first strategy was to randomize each protein sequence in the input *Bacteroidetes*_80aa and *Bacteroidetes*_30aa databases. This produced two randomized databases, *Bacteroidetes*_80aa_R1 and *Bacteroidetes*_30aa_R1, which maintained the same amino acid composition of each protein sequence in the input databases. The second strategy was to extract random regions of each protein sequence from the full *Bacteroidetes* database. The length of the extracted regions was 80 amino acids for the generation of *Bacteroidetes*_80aa_R2 and 30 amino acids for *Bacteroidetes*_30aa_R2. Once the four databases had been generated, Motif B was searched against *Bacteroidetes*_80aa_R1 and *Bacteroidetes*_80aa_R2 while Motif D+E was searched against *Bacteroidetes*_30aa_R1 and *Bacteroidetes*_30aa_R2. Results of all four hmmpfam search were processed using Microsoft Access to produce a list of proteins that hit Motif B and Motif D+E.

Bacterial Strains and Culture

P. gingivalis W50 was grown as previously described¹⁶. *Parabacteroides distasonis*⁴⁵ and *Prevotella intermedia* (ATCC 25611) were grown on horse blood agar containing 10% v/v defibrinated horse blood and 5 mg L⁻¹ menadione. The bacteria were grown in tryptic soy-enriched brain heart infusion broth (TSBHI), containing per litre, 25 g of tryptic soy, 30 g of brain heart infusion, 0.5 g of cysteine, 5 mg of haemin and 1 mg of menadione. Bacteria were incubated at 37°C in an anaerobic workstation (MACS, Don Whitely Scientific, UK). *P. distasonis* and *P. intermedia* were grown in planktonic culture until they reached a cell density equivalent to ~0.6 AU_{650nm} and then harvested.

Cytophaga hutchinsonii ATCC 33406²² was grown in Dubos Broth containing per litre, 0.5 g NaNO₃, 1 g K₂HPO₄, 0.5 g MgSO₄.7H₂O, 0.5 g KCl, 10 mg FeSO₄.7H₂O. The

pH of the medium was adjusted to 7.2 and then filter sterilized. To Dubos Broth 1.5% (w/v) of agar was added for solid medium. Sterile strips of Whatman No.1 filter paper were laid on the surface of the agar and inserted into the broth. Bacteria were maintained on solid agar by adding a drop of bacterial suspension or streak from a previous plate and placing filter paper on top. Plates were incubated for 7-10 days at 26°C in the anaerobic workstation. Broth cultures were inoculated with bacteria grown on plates and incubated for 10 days, these planktonic cultures were then used to inoculate larger volumes (1/50) which were incubated for 10 days. Bacteria were grown at 26°C in a shaking incubator at 40 rpm.

Bacteroides thetaiotaomicron ATCC 29148⁴⁶ was grown in either Chopped Meat/Tryptic Soy Broth Medium with added Yeast Extract (containing per litre, 500 g ground beef, 25 mM NaOH, 30 g Trypticase Soy Broth, 5 g Yeast Extract, 5 g K₂PO₄ and added supplements after sterilization 0.5g cysteine-HCl, 5 mg hemin and 1 mg Vitamin K) or Brain Heart Infusion Medium with supplements as previously described⁴⁷. All media were pre-reduced in an anaerobic workstation for 24 h prior to use and all bacterial manipulations were performed in the anaerobic workstation. Stocks were grown for anywhere between 24-48 h in the initial passage and 24 h for subsequent passages. Purity of the cultures was determined by microscopic observation and Gram stain.

Bacterial Fractionation

Bacterial cells were harvested by centrifugation (8,000 g, 4°C, 25 min) and the cell pellet was washed once with PBS + 5 mM MgCl₂ (PBSM) and resuspended in 2 mL of 20% PBSM in a 15 mL plastic tube. Proteinase inhibitor cocktail, 5 uL (PIC, Sigma-Aldrich) was added to the cells of all species and in addition, 5 mM Na-tosyl-L-lysine chloromethyl ketone (TLCK, Sigma-Aldrich) was added to *P. gingivalis* cells before, during and after lysis. Cells were lysed using an ultrasonic processor (model CPX 750, Cole Parmer) fitted with a 3 mm

stepped microtip. The amplitude was set to 40%, pulser to 1 s on, 2 s off for a total of 30 min. The membranes were collected by centrifugation (48,400 g, 20 min) washed twice with PBSM, and stored at -80°C until use. *P. gingivalis* outer membrane vesicles were prepared by collecting the culture fluid which was then passed through a 0.22 µm filter and ultracentrifuged (170,000 g, 4°C, 45 min). The vesicle pellet was washed once in PBS.

SDS-PAGE and Western Blot

SDS-PAGE was conducted using NuPAGE® Novex 10% Bis-Tris gels using MOPS running buffer (Invitrogen). MES running buffer was used to resolve cleaved CTDs. The optimum load of membrane sample to separate by SDS-PAGE was determined empirically. For Western blots, gel-separated proteins were transferred to a nitrocellulose membrane and probed with MAb 1B5 as previously described¹⁶. MAb 1B5 was a gift from Professor Mike Curtis¹⁵⁻¹⁷.

In-gel Digestion and Mass Spectrometry

In-gel digestion using trypsin was performed after reduction with DTT and alkylation with iodoacetamide as previously published¹⁶. Tryptic digests were acidified with TFA and analysed by LC-MS/MS using a Dionex Ultimate 3000 system (Thermo Scientific) coupled to an HCTultra ion trap mass spectrometer as previously described¹⁶. Proteins were identified by MS/MS Ion Search using Mascot v 2.2 (Matrix Science) against the *P. gingivalis* W83 sequence database (2,227 sequences) or other database as detailed below. Search parameters were as follows: enzyme = trypsin, missed cleavages = 1, fixed modifications = carbamidomethyl (Cys), optional modifications = oxidation (Met), MS tolerance = 1.5 Da, MS/MS tolerance = 0.5 Da. To enable the identification of mature N-termini resulting from signal peptide cleavage, additional searches using semi-trypsin and

pyro-Glu (N-term Q or E) as an extra variable modification were conducted. The *P. intermedia* sequence database was sourced from the Oralgen databases in May 2011 and contains 2807 sequences (www.oralgen.lanl.gov), and sequence databases for *P. distasonis* ATCC8503 (3861 sequences) and *C. hutchinsonii* ATCC 33406 (3789 sequences) were obtained from JCVI's Comprehensive Microbial Resource (<http://cmr.jcvi.org>) in November 2011 and August 2011 respectively. Proteins were considered identified when at least one peptide match was greater than the Mascot identity threshold ($p < 0.05$). All peptides were only considered identified if they scored higher than the relevant threshold score ($p < 0.05$) which was 23 for *P. gingivalis* and *P. intermedia*, 25 for *C. hutchinsonii* and 26 for *P. distasonis*. For CTD proteins identified on the basis of a single peptide, the score was greater than 30. The average false discovery rate (FDR) as determined by the Mascot decoy searches was 1.0% (*P. gingivalis*), 0.5% (*P. intermedia*), 1.1% (*P. distasonis*) and 1.3% (*C. hutchinsonii*).

RESULTS

The CTD consists of three conserved motifs

Twenty-two non-redundant CTD proteins from *P. gingivalis* and eight non-redundant CTD proteins from *T. forsythia* were analysed by GLAM2 and MEME to detect, align and define the motifs of the CTD. GLAM2 analyses revealed that these 30 CTD proteins have a conserved C-terminal domain that is approximately 60 amino acids in length. After truncation of the CTD protein sequences to their C-terminal 80 amino acids a second round of GLAM2 analyses were conducted. The analyses revealed a conserved motif that is 42 amino acids long with a score of 946.6 (Fig. 1). Three shorter motifs within the CTD motif labeled Motif B, Motif D and Motif E ²⁰were observed in the GLAM2 data (Fig. 1) and were verified by MEME (Table 1, Supplementary Figure 1). Absolutely conserved Gly residues were observed at positions 8 and 25, while absolutely conserved hydrophobic residues were observed at positions 1, 3, 29 and 31 (Fig. 1). Motif E, although small, exhibited an absolutely conserved basic amino acid residue at position 39 (almost exclusively Lys). Hydrophobic amino acid residues were strongly preferred in the next three positions in Motif E (Fig. 1).

Using the combined results of the GLAM2 and MEME analyses a multiple sequence alignment was constructed with the three motifs marked, allowing species-specific differences to be noted (Fig. 2). In Motif B of *P. gingivalis* CTD proteins, position 5 preferred Asp which is acidic while in *T. forsythia*, this position preferred Asn, Ser or Thr which are polar, non-charged residues. At position 6, *P. gingivalis* preferred Met or Leu while *T. forsythia* preferred Val or Ile. At position 9, *P. gingivalis* preferred the positively charged residues, Arg and Lys while *T. forsythia* preferred the polar non charged residues, Ser and Gln. In Motif D and Motif E, there was little difference in preferred residues between *P. gingivalis* and *T. forsythia*.

The CTD is present only in the *Bacteroidetes/Chlorobi* phylum

The two generated HMM profiles, Motif B and Motif D+E profiles (Fig. 2) were used to search the Uniprot's Swiss-Prot and TrEMBL protein sequence databases. Sequences that matched both Motif B and Motif D+E profiles were defined as putative CTD proteins.

Putative CTD proteins were identified in 87 organisms belonging to the *Bacteroidetes* and in 8 unclassified organisms. 682 putative CTD proteins were identified using a combined E-value of $1e-6$. CTD proteins had a wide species distribution but all were contained in organisms from the *Bacteroidetes* phylum (Supplementary Table 1).

For a more sensitive analysis, the search for putative CTD proteins was repeated against 21 fully sequenced *Bacteroidetes* species using a less strict combined E-value of $1e-2$. The 21 species covered a broad distribution of *Bacteroidetes* organisms accounting for 8 of 15 known families. The total number of proteins that matched both HMM profiles was 663 proteins (Fig. 3). In 12 of the 21 *Bacteroidetes* species, at least 20 putative CTD proteins were predicted by the HMM searching and five other *Bacteroidetes* species had at least nine putative CTD proteins. Organisms that had at least nine putative CTD proteins were found in seven different families in the *Bacteroidetes* phylum while hits to the *Bacteroidaceae* family were very few (Fig. 3). Upon further analysis of the hits to the *Bacteroidaceae* family, four of the five hits either lacked an N-terminal signal peptide or else exhibited a Motif D+E hit that was not at the C terminus. Only one hit, BVU_0089 was considered an acceptable hit, however it was missing the conserved basic residue in Motif E (data not shown), and hence these species were considered to be CTD-negative (coloured red in Fig. 3). The CTD proteins predicted for *P. gingivalis*, *T. forsythia*, *P. distasonis*, *P. intermedia* and *C. hutchinsonii* are listed in Supplementary Table 2.

HMM searching was also conducted against decoy databases in order to estimate the number of false positives. Results for the first decoy strategy identified eight proteins that hit Motif B and Motif D+E profiles equating to 0.11% of the entire *Bacteroidetes* database. The second strategy found six proteins equating to 0.008% of the database. The false positive rate was therefore estimated to be 1.2% (8/663) and 0.9% (6/663), respectively.

CTD proteins are abundant and extensively modified in representative *Bacteroidetes*

To discern whether LPS-modification is common to all *P. gingivalis* CTD proteins, *P. gingivalis* cells were fractionated and subjected to proteome analysis. The outer membrane vesicle fraction was found to be enriched with CTD proteins and was therefore comprehensively analysed. A total of 29 CTD proteins were identified from LC-MS/MS analyses of 17 gel segments (Supplementary Table 3). Of these, 19 were identified from segments of considerably higher MW than predicted from the protein sequence, consistent with previously reported LPS-modification (Fig 4A). The modification increased the apparent MW of most proteins by approximately 20-50 kDa. The modified form of RgpA (PG2024) observed was the C-terminal A4 adhesin domain, as previously reported^{13, 19}, and for the first time, the modified form of the C-terminal A5 adhesin domain of Kgp (PG1844) was found (Fig 4A) as previously predicted¹³. Of the ten CTD proteins that did not exhibit a significantly increased MW, only PG0026 and PG1604 appeared to be intact, and therefore not extensively modified (Supplementary Table 3). The other eight CTD proteins may be modified but were too big to assess (e.g PG0183), or were proteolytically cleaved into smaller domains or fragments of unknown size (e.g PG1798, PG1427). Most also exhibited low sequence coverage making it difficult to assess the form of the protein identified (Supplementary Table 3). Apart from PG0026 and PG1604, it appeared that all other intact CTD proteins were extensively modified.

To determine whether the presence of abundant, cell-associated, and extensively post-translationally modified CTD proteins are general features of other *Bacteroidetes* species that contain predicted CTD proteins, the membrane fractions of three predicted CTD-positive species, namely *P. intermedia*, *P. distasonis* and *C. hutchinsonii* (Fig 3) were analysed by SDS-PAGE and LC-MS/MS. In total, 298 proteins were identified from *P. intermedia*, 639 proteins from *P. distasonis* and 660 proteins from *C. hutchinsonii*. The complete identification data for all of the proteins identified are presented in Supplementary Table 3.

The membrane fraction of *P. distasonis* exhibited an SDS-PAGE band at approximately 160 kDa that was considerably stronger than any other band (Fig 4B). The major protein present in this band was BDI_2562, one of the predicted CTD family proteins (Supplementary Table 2) with a Mascot score of 4,624 (Fig 4B). The next five strongest hits (ranked 2-6) for this band were also predicted CTD proteins with Mascot scores ranging between 408 – 606 (Fig 4B). These and other predicted CTD proteins also dominated bands 24, 29 and 30, and were strongly represented in band 23 (Fig 4B). Of the ten identified CTD proteins, eight were observed at elevated MW relative to the estimated MW of the predicted mature sequence (Fig 4B). The PG0026 homolog, BDI_2576 was identified in band 27 in accordance with its predicted full length MW (Fig 4B) suggesting that like PG0026 (see above), it may lack extensive modification. For BDI_3364, BDI_2055 and BDI_2580 that exhibited forms at an equal or lower MW than expected (Fig 4B), in each case, the sequence coverage obtained by the MS analyses indicated that the form of the protein identified corresponded to an unmodified N-terminal fragment.

For *P. intermedia*, of the 298 non-redundant proteins identified from the membrane fraction, 15 were predicted to be CTD proteins by our bioinformatic analyses (Supplementary Table 2). In addition to the 15 predicted CTD proteins, two abundant proteins of elevated MW, namely PI_0610 and PI_1993 were identified (Fig. 4C). These proteins matched to

motif B but not motif D+E in the HMM search, however upon multiple alignment of their CTD sequences with the predicted CTD sequences, it was clear that they did in fact contain the major features of motif D+E (Supplementary Fig. 2), and were therefore considered to be CTD proteins. In total, twelve of the proteins exhibited a substantial increase in MW consistent with their conjugation to a large polymer such as LPS (Fig 4C). In terms of abundance, the identified CTD proteins dominated the high MW region of the gel from 120 kDa and higher, obtaining the top protein rank for each of the four bands in this region (Fig 4C). Although there were no single proteins as abundant as some of CTD proteins in *P. distasonis*, there were many more CTD proteins identified. The PG1604 homolog, PI_2159 was identified in band 15 in accordance with its predicted full length MW indicating that like PG1604 (see above) it was not extensively modified (Fig 4C). The proteomic data for PI_2050 and PI_1698 indicated extensive N-terminal truncation.

For *C. hutchinsonii*, 18 of the predicted CTD proteins were identified, and an increase in MW was observed for eight proteins (Fig 4D). Again, the identified CTD proteins were very abundant in the high MW region of the gel, with the putative glycosidases CHU1075 and CHU1107 achieving the highest Mascot scores (Fig 4D). Homologs to both PG0026 (CHU3237) and PG1604 (CHU3302) were identified in bands corresponding to their expected full length MW (Fig 4D). Seven of the predicted CTD proteins identified were annotated as “CHU large proteins”, however eleven additional proteins with this annotation were identified (Supplementary Table 3). Most of these proteins appear to have a C-terminal sequence that is closely related to the CTD. Of particular note was a group of CHU large proteins expressed from mostly adjacent genes (CHU_3435, CHU_3437, CHU_3439-3441 and CHU_3654) that shared an almost identical 87-residue C-terminal sequence. The MW of these proteins ranged from 242-297 kDa, and therefore it could not be easily determined by SDS-PAGE and MS whether they were C-terminally truncated and modified. To determine

whether these C-terminal sequences could be predicted by HMMs, a new HMM was generated based on the 95 predicted CTDs for *C. hutchinsonii* (Supplementary Figure 3). Using the new model an additional 76 putative CTD proteins were predicted, including the CHU large proteins listed above. Comparison of the alignment of these newly predicted CTD proteins with the original model demonstrates the similarity of the sequences, and reveals interesting differences in all three motifs (Supplementary Fig. 4). Twelve of the newly predicted CTD proteins were identified by MS, three of which were elevated in MW (Supplementary Table 3).

A proteomic analysis was not conducted on *B. thetaiotaomicron* since the only predicted CTD protein (BT_4596) was only 82 amino acids long and did not have an N-terminal signal peptide. In contrast, all of the CTD proteins identified above in *P. gingivalis*, *P. distasonis*, *P. intermedia* and *C. hutchinsonii* exhibited predicted N-terminal signal peptides. The cleavage of five N-terminal signal peptides from *P. distasonis*, *P. intermedia* and *C. hutchinsonii* was demonstrated by the identification of mature N-terminal peptides, four of which contained pyroglutamate at the N-terminus (Supplementary Table 3).

The CTD is cleaved in representative *Bacteroidetes*: Of all the modified CTD proteins identified from the four organisms examined, no peptides from the CTD itself were identified, consistent with the CTD being cleaved in each organism. Strong examples of this include PG2102 (P59), PG0553 and BDI_2562 where the sequence coverage obtained for the mature region was 69%, 70%, and 62% respectively, yet no peptides for the CTD region were identified (Fig 5A-C). Interestingly, in *P. intermedia* the sequence coverage obtained for five CTD proteins extended to within 50 - 60 residues from their C-terminus, indicating that the CTD of *P. intermedia* proteins can be shorter than that found in the other species examined (Fig 5D). To verify that the CTD was indeed cleaved in each species, the culture fluid

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

fractions were fractionated by SDS-PAGE and the region less than ~12 kDa analysed for the presence of cleaved CTDs. The culture fluid of *T. forsythia* was also examined. A total of twelve cleaved CTDs were identified from *T. forsythia*, five from *P. distasonis*, six from *C. hutchinsonii* and seven from *P. intermedia* (Supp Table 3). Included in this data set were five semi-tryptic peptides from *T. forsythia*, four from *P. distasonis* and one from *C. hutchinsonii* that reveal the likely CTD cleavage sites (Fig. 6). One of the cleaved CTDs identified from *P. distasonis* was BDI_0323, which was predicted by HMM to contain motif D+E only. Together the data show that CTD cleavage is a conserved feature of the secretion system.

DISCUSSION

Characterisation of the CTD

Based on our bioinformatic analyses, the C-terminal domain of 30 *P. gingivalis* and *T. forsythia* protein sequences contained a conserved region of about 42 amino acids in length (excluding gaps) consisting of three motifs that were statistically significant according to MEME. Seers et al (2006)²⁰ divided the CTD into five motifs named Motif A to E, however our findings with GLAM2 indicated that only Motif B, D and E were significantly conserved. The HMM based on these three motifs was successful in predicting large numbers of CTD proteins both from species within the same taxonomic class (*Bacteroidia*) as *P. gingivalis* and *T. forsythia* and also from more distantly related organisms within the *Flavobacteriia*, *Cytophagia* and *Sphingobacteriia* classes of the *Bacteroidetes* phylum (Fig. 3). In representative organisms, motifs B, D and E were always found together in CTD proteins that were detected to be extensively modified by SDS-PAGE. While the CTD can be predicted using the above motifs that encompass approximately 42 amino acid residues, the CTD was shown to extend 27-36 residues N-terminal of motif B. The total length of the CTD in those CTDs with known or strongly predicted cleavage sites was between 67 and 88 amino acid residues (Fig. 6). However, as shown in Fig 5D, several CTD proteins of *P. intermedia* appear to have CTDs that are very short, with some being less than 50 residues. CTD cleavage of these short CTDs however remains to be demonstrated. The CTD cleavage sites determined (Fig. 6) are similar in that the P1' residue is always small (Ser, Ala, or Gly, or rarely Asp). A pattern can also be seen for the P1 position, but is species dependent. For example Thr is preferred in *P. distasonis*, Val is preferred in *T. forsythia*, while *P. gingivalis* likes a polar or negatively charged residue. Species specific preferences were also detected for positions P2' to P5'. Thus consensus cleavage sites were T/SNExI (*P. distasonis*), V/ANETI (*T. forsythia*), and T/SLADV (*P. gingivalis*).

Prediction of CTD proteins

A Hidden Markov Model was used to profile the chosen CTD proteins. Due to the large variability of the CTD, conventional bioinformatic methods (e.g BLAST) for finding CTD sequences in other species were not suitable. Using HMM, we were able to capture the information of 30 CTD proteins in a HMM profile which acts like a super sequence having the features of all the 30 CTD proteins. Using this method, many more candidate CTD proteins were found compared to conventional protein homology searching.

In addition to *P. gingivalis*²⁰, and *T. forsythia*²⁵, the CTD motif or D5 domain has been reported in *Rhodothermus marinus*²³ *Cytophaga hutchinsonii*²² and *Flavobacterium psychrophilum*⁹.

P. gingivalis contained 23 experimentally identified CTD proteins that were used to build the HMM profile, and an additional 11 CTD proteins predicted by BLAST only²⁰. The HMM method employed predicted two extra CTD proteins, PG0183 and PG1604, but did not support the prediction of PG0111, PG0290 or PG1035. The 11 *T. forsythia* CTD protein sequences that were used in building the HMM profile were experimentally identified CTD proteins from Veith et al (2009)²⁵. Veith et al (2009)²⁵ further identified 15 proteins that might be CTD proteins using multiple BLAST searches. All 15 proteins were also identified using the HMM methods. In addition, another 11 *T. forsythia* CTD proteins were predicted for a total of 37 CTD proteins. In comparison, 25 *T. forsythia* CTD proteins were predicted using a simpler protein motif search²¹.

The HMM methods found 95 candidate CTD proteins in *C. hutchinsonii* compared to only 57 found by methods using BlastP²², clearly demonstrating the power of the HMM to predict proteins from distantly related organisms. However, 19 out of the 57 proteins found by Xie et al (2007)²² were not found by the HMM. Further HMM modelling using the

initial 95 candidates predicted a further 76 CTD candidates including all but one of the proteins predicted by Xie et al (2007)²². In the alignment of Xie et al (2007)²² it is apparent that the YPNP motif was important. The YPNP motif is however present in only a few *P. gingivalis* and *T. forsythia* CTD sequences, and was therefore not used in our model.

Previously, a total of nine proteins were found in *P. distasonis* that had C-terminal similarity to each other and to the surface layer proteins of *T. forsythia*²⁴ in contrast to the 22 predicted here by HMM. Karlsson et al (2005)²³ described a conserved C-terminal domain in *Rhodothermus marinus* proteins, RmX10A (Rm1069) and RmM26A (Rm0016) which they called the D5 domain. Using BLAST, they found another 12 D5-containing proteins. In contrast, the HMM profiles predicted 54 CTD proteins in *R. marinus*. Similarly, in *Flavobacterium psychrophilum*, 34 proteins containing the CTD have been predicted previously⁹. Our HMM methodology predicted 33 of these proteins and predicted an extra eight CTD proteins. In each case the HMM profile despite being built from proteins of another bacterial class was able to predict more CTD proteins than that found by conventional searching within a specific organism.

CTD proteins were found to be widespread in the *Bacteroidetes* phylum being predicted in 87 different species, a number that will continue to grow as the genomes of more *Bacteroidetes* species are sequenced. A total of 663 putative CTD proteins were predicted from only 17 fully sequenced CTD-positive species, at an average of 39 different CTD proteins per organism (Fig. 3). A second round of HMM profiling of *C. hutchinsonii* predicted a total of 171 CTD proteins suggesting that the numbers presented in Fig. 3 may be an underestimate. The ease of predicting the CTD signal contrasts sharply with the prediction of secretion signals in the substrates of other major secretion systems. In Type II systems, the largest number of substrates reported in a single species is 25 as determined by proteomic analyses of specific mutants⁴⁸. In type III and IV systems, due to the complex nature of the secretion

signals, sophisticated algorithms are needed to predict substrates which ultimately need to be confirmed experimentally⁴⁹.

Characterisation of CTD proteins

The putative functions of the predicted proteins based on their annotation included proteases, glycosidases, adhesins, hemagglutinins and internalins (Supplementary Table 2 & 3). In this study, a comprehensive analysis of *P. gingivalis* OM vesicles demonstrated that extensive modification of CTD proteins, probably with A-LPS¹⁵⁻¹⁹ was a standard feature with few exceptions. This rule of extensive modification was also true for the CTD proteins identified in *P. distasonis*, *P. intermedia*, and *C. hutchinsonii*, and has been previously demonstrated in *T. forsythia*²⁵. Although extensive modification has not yet been definitively demonstrated by elucidating the nature or structure of the linkage between protein and polysaccharide, the cumulative evidence for modification is strong. Some proteins, especially when in high concentration can produce high MW SDS-PAGE bands corresponding to oligomeric forms of the protein. However in such cases the monomer is the most abundant form. The results of this study showed that in general, CTD proteins do not produce bands at their expected monomer MW and therefore it is extremely unlikely that the appearance of abundant higher MW forms is due to oligomerisation. The molecule responsible for the extensive modification in each species may be different, as LPS and other surface polysaccharides are generally species-specific. Immunoblot analysis of the cell envelope fractions of *P. intermedia*, *C. hutchinsonii*, and *P. distasonis* using MAb-1B5 which recognises a Man α 1-2Man α 1-phosphate epitope of the modified CTD proteins of *P. gingivalis*¹⁵⁻¹⁹ seemed to weakly cross-react with proteins from *C. hutchinsonii* but not to the other two species (data not shown).

Other conserved features of secretion were the high abundance of the identified CTD proteins, and their high MW. The dominant CTD proteins of each of the species examined, as well as of *T. forsythia* had a predicted MW of at least 110 kDa (Fig 4). In *P. gingivalis*, the most abundant CTD proteins are RgpA and Kgp which each have a precursor MW of above 180 kDa. After proteolytic processing of their individual proteolytic and adhesin domains¹⁹,⁵⁰ they associate non-covalently to form multimeric complexes of up to 660 kDa^{51, 52}. In *P. distasonis*, the nine CTD proteins identified had a predicted MW between 90 and 122 kDa and in *C. hutchinsonii*, the two most abundant CTD proteins had a MW of 274 and 135 kDa respectively. These two proteins are annotated as CHU large proteins along with 16 others with a combined average MW of 224 kDa. In *T. forsythia*, 12 of 14 identified CTD proteins had a MW between 114 and 242 kDa²⁵. The reason for the predominance of large CTD proteins is uncertain. Potentially, the proteins are large to position their active sites as far as possible from the cell surface to maximise unhindered access to substrate. This would be particularly useful for proteins like adhesins, hemagglutinins and internalins that interact directly with large substrates such as host tissues/cells or in the case of *C. hutchinsonii*, expansive polysaccharides such as cellulose. The abundance, number and size of CTD proteins identified indicates a remarkable secretion system that must be capable of secreting large proteins very efficiently. It is not known yet whether these proteins are secreted in a folded or unfolded state.

The Combined Attachment and Secretion System (CoAtSS) of the *Bacteroidetes*

At least six major protein secretion pathways have been identified in Gram negative bacteria. Type I, III, IV & VI transport proteins directly from the cytoplasm to the outside of the cell, whereas Type II & V systems secrete proteins from the periplasm⁵³. The secretion system of *P. gingivalis* that is used to secrete the gingipains and other CTD proteins has been

1
2
3 recognised as a new type, and given the number IX²⁹⁻³¹. We propose that this novel Type IX
4
5 secretion system be named CoAtSS, recognising that secretion is combined with surface
6
7 attachment to “coat” the bacterial cell. In each case CoAtSS is characterised by the secretion
8
9 of substrates containing a universal N-terminal signal peptide directing translocation across
10
11 the inner membrane and a conserved C-terminal signal (the CTD) that directs translocation
12
13 across the OM. Recently we demonstrated that CTD cleavage was performed by PG0026 in
14
15 *P. gingivalis* on the cell surface¹⁷. We also concluded that extensive modification of the CTD
16
17 protein was concomitant with CTD cleavage¹⁷. PG0026 is well conserved amongst the
18
19 *Bacteroidetes* and like the *P. gingivalis* form, its homologs in *T. forsythia*²⁵, *P. distasonis*
20
21 and *C. hutchinsonii* (Fig 4) were found to be full length and unmodified consistent with its
22
23 function as the C-terminal signal peptidase on the cell surface. Similar results were found for
24
25 PG1604 and its homologs suggesting that PG1604 may also be a secreted component of the
26
27 secretion system. Furthermore, since CTD cleavage and extensive modification were
28
29 observed in all species analysed (Fig 4, Supplementary Table 3) it would appear that the
30
31 mechanism of cleavage and modification is conserved and is integral to CoAtSS.
32
33
34
35

36 Many CoAtSS components have been demonstrated or predicted to be localised to the
37
38 OM including LptO¹⁶, PorT^{16, 54}, Sov⁵⁵, PorN, PorP, PorK²⁹, PG0026¹⁷, PG0534⁵⁶ and
39
40 PorQ since it belongs to the FadL family of outer membrane proteins (data not shown).
41
42 However, in most secretion systems, only one integral OM protein is required to form the
43
44 secretion channel, while most of the other proteins required for secretion are located in the
45
46 inner membrane, periplasm and cytoplasm. Therefore it is likely that in addition to PG0026
47
48 and LptO, some of the other CoAtSS components listed above are involved in the attachment
49
50 stage of this system. Further to our recommendation that this secretion system is named
51
52 CoAtSS, we propose that CTD proteins that are secreted and attached to the cell surface are
53
54 named CoAtSS substrate proteins (CSPs), and that the genes encoding components of the
55
56
57
58
59
60

1
2
3 system be named *cotA*, *cotB* etc. There is much confusion in the literature of other secretion
4
5 systems where different names have been adopted for every different species, and therefore a
6
7 species-neutral name such as cot (pronounced as “coat”) that could be utilised for all
8
9 *Bacteroidetes* species appears useful. In summary, the major protein secretion system present
10
11 in the *Bacteroidetes* and which we propose to name CoAtSS is characterised by its unique
12
13 cleavable C-terminal signal, the attachment of its substrates to the cell surface, and its
14
15 tendency to secrete copious quantities of large substrates.
16
17
18
19

20 21 **ACKNOWLEDGEMENTS**

22
23 This project was supported by the Australian National Health and Medical Research Council
24
25 Project Grant # 1027812. Ms Rita Paolini is thanked for technical assistance with bacterial
26
27 culture.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ASSOCIATED CONTENT

Supporting Information: This material is available free of charge via the internet at <http://pubs.acs.org>

Supplementary Figure 1. Protein logo representations of the individual motifs present in the last 80 amino acids of the C-terminal region of 30 experimentally identified CTD proteins from *Porphyromonas gingivalis* and *Tannerella forsythia* produced by MEME (<http://meme.sdsc.edu>). (A) Motif B (B) Motif D (C) Motif E.

Supplementary Figure 2. Multiple sequence alignment of the C-terminal region of 54 predicted *P. intermedia* CTD proteins, PI0610 and PI1993. Alignment was extracted from GLAM2 analysis of the 56 *P. intermedia* protein sequences. Parameters used for the GLAM2 analysis were -b (maximum number of aligned columns) = 100, -z (minimum number of sequences in the alignment) = 56 and default for the remaining parameters. *P. intermedia* protein sequences were from the Oralgen Sequence Databases (www.oralgen.lanl.gov).

Supplementary Figure 3. The *C. hutchinsonii* CTD motif derived from 95 HMM predicted CTD proteins in *C. hutchinsonii* (using HMM from *P. gingivalis* and *T. forsythia*). The motif was discovered using GLAM2 with parameters set to: -b (maximum number of aligned columns) = 100, -z (minimum number of sequences in the alignment) = 95 and default for the remaining parameters. *C. hutchinsonii* protein sequences were sourced from NCBI Genbank (www.ncbi.nlm.nih.gov/genbank).

Supplementary Figure 4. Multiple sequence alignment of the C-terminal region of 102 *C. hutchinsonii* proteins. The alignment was constructed from 95 predicted CTD proteins and nine *C. hutchinsonii* large proteins (two of which were in the 95 predicted CTD proteins). Alignment was extracted from GLAM2 analysis of the 102 *C. hutchinsonii* protein sequences. Parameters used for the GLAM2 analysis were -b (maximum number of aligned columns) = 100, -z (minimum number of sequences in the alignment) = 102 and default for the remaining parameters. *C. hutchinsonii* protein sequences were sourced from NCBI Genbank (www.ncbi.nlm.nih.gov/genbank).

Supplementary Table 1 - Predicted CTD proteins in Swiss-Prot and TrEMBL databases.

Supplementary Table 2 - Predicted CTD proteins in selected bacteria.

Supplementary Table 3 - Proteomic data.

REFERENCES

(1) Thomas, F.; Hehemann, J. H.; Rebuffet, E.; Czejzek, M.; Michel, G. Environmental and gut bacteroidetes: the food connection. *Front. Microbiol.* **2011**, *2*, 93.

(2) Janssen, P. H. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl. Environ. Microbiol.* **2006**, *72*, 1719-1728.

(3) Frank, D. N.; St Amand, A. L.; Feldman, R. A.; Boedeker, E. C.; Harpaz, N.; Pace, N. R. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 13780-13785.

(4) Pommier, T.; Canback, B.; Riemann, L.; Bostrom, K. H.; Simu, K.; Lundberg, P.; Tunlid, A.; Hagstrom, A. Global patterns of diversity and community structure in marine bacterioplankton. *Mol. Ecol.* **2007**, *16*, 867-880.

(5) Ley, R. E.; Turnbaugh, P. J.; Klein, S.; Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **2006**, *444*, 1022-1023.

(6) Hardham, J.; Dreier, K.; Wong, J.; Sfintescu, C.; Evans, R. T. Pigmented-anaerobic bacteria associated with canine periodontitis. *Vet. Microbiol.* **2005**, *106*, 119-128.

(7) Socransky, S. S.; Haffajee, A. D.; Cugini, M. A.; Smith, C.; Kent, R. L., Jr. Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* **1998**, *25*, 134-144.

(8) Siqueira, J. F., Jr.; Rocas, I. N. Diversity of endodontic microbiota revisited. *J. Dent. Res.* **2009**, *88*, 969-981.

(9) Duchaud, E.; Boussaha, M.; Loux, V.; Bernardet, J. F.; Michel, C.; Kerouault, B.; Mondot, S.; Nicolas, P.; Bossy, R.; Caron, C.; Bessieres, P.; Gibrat, J. F.; Claverol, S.; Dumetz, F.; Le Henaff, M.; Benmansour, A. Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. *Nat. Biotechnol.* **2007**, *25*, 763-769.

- (10) Bauer, M.; Kube, M.; Teeling, H.; Richter, M.; Lombardot, T.; Allers, E.; Wurdemann, C. A.; Quast, C.; Kuhl, H.; Knaust, F.; Woebken, D.; Bischof, K.; Musmann, M.; Choudhuri, J. V.; Meyer, F.; Reinhardt, R.; Amann, R. I.; Glockner, F. O. Whole genome analysis of the marine Bacteroidetes 'Gramella forsetii' reveals adaptations to degradation of polymeric organic matter. *Environ. Microbiol.* **2006**, *8*, 2201-2213.
- (11) Kirchman, D. L. The ecology of Cytophaga-Flavobacteria in aquatic environments. *FEMS Microbiol. Ecol.* **2002**, *39*, 91-100.
- (12) Haffajee, A. D.; Socransky, S. S. Microbial etiological agents of destructive periodontal diseases. *Periodontol.* **2000** *1994*, *5*, 78-111.
- (13) O'Brien-Simpson, N. M.; Veith, P. D.; Dashper, S. G.; Reynolds, E. C. *Porphyromonas gingivalis* gingipains: the molecular teeth of a microbial vampire. *Curr. Protein Peptide Sci.* **2003**, *4*, 409-426.
- (14) O'Brien-Simpson, N. M.; Paolini, R. A.; Hoffmann, B.; Slakeski, N.; Dashper, S. G.; Reynolds, E. C. Role of RgpA, RgpB, and Kgp proteinases in virulence of *Porphyromonas gingivalis* W50 in a murine lesion model. *Infect. Immun.* **2001**, *69*, 7527-7534.
- (15) Curtis, M. A.; Thickett, A.; Slaney, J. M.; Rangarajan, M.; Aduse-Opoku, J.; Shepherd, P.; Paramonov, N.; Hounsell, E. Variable carbohydrate modifications to the catalytic chains of the RgpA and RgpB proteases of *Porphyromonas gingivalis* W50. *Infect. Immun.* **1999**, *67*, 3816-3823.
- (16) Chen, Y. Y.; Peng, B.; Yang, Q.; Glew, M. D.; Veith, P. D.; Cross, K. J.; Goldie, K. N.; Chen, D.; O'Brien-Simpson, N.; Dashper, S. G.; Reynolds, E. C. The outer membrane protein LptO is essential for the O-deacylation of LPS and the co-ordinated secretion and attachment of A-LPS and CTD proteins in *Porphyromonas gingivalis*. *Mol. Microbiol.* **2011**, *79*, 1380-1401.

(17) Glew, M. D.; Veith, P. D.; Peng, B.; Chen, Y. Y.; Gorasia, D. G.; Yang, Q.; Slakeski, N.; Chen, D.; Moore, C.; Crawford, S.; Reynolds, E. C. PG0026 is the C-terminal signal peptidase of a novel secretion system of *Porphyromonas gingivalis*. *J. Biol. Chem.* **2012**, 287, 24605-24617.

(18) Rangarajan, M.; Aduse-Opoku, J.; Paramonov, N.; Hashim, A.; Bostanci, N.; Fraser, O. P.; Tarelli, E.; Curtis, M. A. Identification of a second lipopolysaccharide in *Porphyromonas gingivalis* W50. *J. Bacteriol.* **2008**, 190, 2920-2932.

(19) Veith, P. D.; Talbo, G. H.; Slakeski, N.; Dashper, S. G.; Moore, C.; Paolini, R. A.; Reynolds, E. C. Major outer membrane proteins and proteolytic processing of RgpA and Kgp of *Porphyromonas gingivalis* W50. *Biochem. J.* **2002**, 363, 105-115.

(20) Seers, C. A.; Slakeski, N.; Veith, P. D.; Nikolof, T.; Chen, Y. Y.; Dashper, S. G.; Reynolds, E. C. The RgpB C-terminal domain has a role in attachment of RgpB to the outer membrane and belongs to a novel C-terminal-domain family found in *Porphyromonas gingivalis*. *J. Bacteriol.* **2006**, 188, 6376-6386.

(21) Nguyen, K. A.; Travis, J.; Potempa, J. Does the importance of the C-terminal residues in the maturation of RgpB from *Porphyromonas gingivalis* reveal a novel mechanism for protein export in a subgroup of Gram-Negative bacteria? *J. Bacteriol.* **2007**, 189, 833-843.

(22) Xie, G.; Bruce, D. C.; Challacombe, J. F.; Chertkov, O.; Detter, J. C.; Gilna, P.; Han, C. S.; Lucas, S.; Misra, M.; Myers, G. L.; Richardson, P.; Tapia, R.; Thayer, N.; Thompson, L. S.; Brettin, T. S.; Henrissat, B.; Wilson, D. B.; McBride, M. J. Genome sequence of the cellulolytic gliding bacterium *Cytophaga hutchinsonii*. *Appl. Environ. Microbiol.* **2007**, 73, 3536-3546.

(23) Karlsson, E. N.; Hachem, M. A.; Ramchuran, S.; Costa, H.; Holst, O.; Svenningsen, S. F.; Hreggvidsson, G. O. The modular xylanase Xyn10A from *Rhodothermus marinus* is

- cell-attached, and its C-terminal domain has several putative homologues among cell-attached proteins within the phylum Bacteroidetes. *FEMS Microbiol. Lett.* **2004**, *241*, 233-242.
- (24) Fletcher, C. M.; Coyne, M. J.; Bentley, D. L.; Villa, O. F.; Comstock, L. E. Phase-variable expression of a family of glycoproteins imparts a dynamic surface to a symbiont in its human intestinal ecosystem. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 2413-2418.
- (25) Veith, P. D.; O'Brien-Simpson, N. M.; Tan, Y.; Djatmiko, D. C.; Dashper, S. G.; Reynolds, E. C. Outer membrane proteome and antigens of *Tannerella forsythia*. *J. Proteome Res.* **2009**, *8*, 4279-4292.
- (26) Shoji, M.; Sato, K.; Yukitake, H.; Kondo, Y.; Narita, Y.; Kadowaki, T.; Naito, M.; Nakayama, K. Por secretion system-dependent secretion and glycosylation of *Porphyromonas gingivalis* hemin-binding protein 35. *PLoS One* **2011**, *6*, e21372.
- (27) Ishiguro, I.; Saiki, K.; Konishi, K. PG27 is a novel membrane protein essential for a *Porphyromonas gingivalis* protease secretion system. *FEMS Microbiol. Lett.* **2009**, *292*, 261-267.
- (28) Saiki, K.; Konishi, K. Identification of a *Porphyromonas gingivalis* novel protein required for the secretion of gingipains. *Mol. Immunol.* **2007**, *51*, 483-491.
- (29) Sato, K.; Naito, M.; Yukitake, H.; Hirakawa, H.; Shoji, M.; McBride, M. J.; Rhodes, R. G.; Nakayama, K. A protein secretion system linked to bacteroidete gliding motility and pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 276-281.
- (30) Sato, K.; Sakai, E.; Veith, P. D.; Shoji, M.; Kikuchi, Y.; Yukitake, H.; Ohara, N.; Naito, M.; Okamoto, K.; Reynolds, E. C.; Nakayama, K. Identification of a new membrane-associated protein that influences transport/maturation of gingipains and adhesins of *Porphyromonas gingivalis*. *J. Biol. Chem.* **2005**, *280*, 8668-8677.

- (31) Sato, K.; Yukitake, H.; Narita, Y.; Shoji, M.; Naito, M.; Nakayama, K. Identification of *Porphyromonas gingivalis* proteins secreted by the Por secretion system. *FEMS Microbiol. Lett.* **2013**, *338*, 68-76.
- (32) Lee, S. W.; Sabet, M.; Um, H. S.; Yang, J.; Kim, H. C.; Zhu, W. Identification and characterization of the genes encoding a unique surface (S-) layer of *Tannerella forsythia*. *Gene* **2006**, *371*, 102-111.
- (33) Frith, M. C.; Saunders, N. F.; Kobe, B.; Bailey, T. L. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comp. Biol.* **2008**, *4*, e1000071.
- (34) Bailey, T. L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 28-36.
- (35) Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202-208.
- (36) Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res* **2013**, *41*, D36-42.
- (37) Olson, S. A. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform* **2002**, *31*, 87-91.
- (38) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276-277.
- (39) Finn, R. D.; Clements, J.; Eddy, S. R. MMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **2011**, *39*, W29-37.
- (40) Magrane, M.; Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, *2011*.
- (41) Consortium, T. U. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **2013**, *41*, D36-42.

- (42) Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res* **2012**, *40*, D136-143.
- (43) Jenuth, J. P. The NCBI. Publicly available tools and resources on the Web. *Methods Mol. Biol.* **2000**, *132*, 301-312.
- (44) Retief, J. D. Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* **2000**, *132*, 243-258.
- (45) Sakamoto, M.; Benno, Y. Reclassification of *Bacteroides distasonis*, *Bacteroides goldsteinii* and *Bacteroides merdae* as *Parabacteroides distasonis* gen. nov., comb. nov., *Parabacteroides goldsteinii* comb. nov. and *Parabacteroides merdae* comb. nov. *Int. J. Syst. Evol. Microbiol.* **2006**, *56*, 1599-1605.
- (46) Xu, J.; Bjursell, M. K.; Himrod, J.; Deng, S.; Carmichael, L. K.; Chiang, H. C.; Hooper, L. V.; Gordon, J. I. A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science* **2003**, *299*, 2074-2076.
- (47) Bacic, M. K.; Smith, C. J. Laboratory maintenance and cultivation of bacteroides species. *Curr. Protoc. Microbiol.* **2008**, *Chapter 13*, Unit 13C 11.
- (48) Cianciotto, N. P. Many substrates and functions of type II secretion: lessons learned from *Legionella pneumophila*. *Future Microbiol.* **2009**, *4*, 797-805.
- (49) McDermott, J. E.; Corrigan, A.; Peterson, E.; Oehmen, C.; Niemann, G.; Cambronne, E. D.; Sharp, D.; Adkins, J. N.; Samudrala, R.; Heffron, F. Computational prediction of type III and IV secreted effectors in gram-negative bacteria. *Infect. Immun.* **2011**, *79*, 23-32.
- (50) Veith, P. D.; Chen, Y. Y.; Reynolds, E. C. *Porphyromonas gingivalis* RgpA and Kgp proteinases and adhesins are C terminally processed by the carboxypeptidase CPG70. *Infect. Immun.* **2004**, *72*, 3655-3657.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

(51) Bhogal, P. S.; Slakeski, N.; Reynolds, E. C. A cell-associated protein complex of *Porphyromonas gingivalis* W50 composed of Arg- and Lys-specific cysteine proteinases and adhesins. *Microbiology* **1997**, *143*, 2485-2495.

(52) Takii, R.; Kadowaki, T.; Baba, A.; Tsukuba, T.; Yamamoto, K. A functional virulence complex composed of gingipains, adhesins, and lipopolysaccharide shows high affinity to host cells and matrix proteins and escapes recognition by host immune systems. *Infect. Immun.* **2005**, *73*, 883-893.

(53) Holland, I. B. The extraordinary diversity of bacterial protein secretion mechanisms. *Methods Mol. Biol.* **2010**, *619*, 1-20.

(54) Nguyen, K. A.; Zylicz, J.; Szczesny, P.; Sroka, A.; Hunter, N.; Potempa, J. Verification of a topology model of PorT as an integral outer-membrane protein in *Porphyromonas gingivalis*. *Microbiology* **2009**, *155*, 328-337.

(55) Saiki, K.; Konishi, K. The role of Sov protein in the secretion of gingipain protease virulence factors of *Porphyromonas gingivalis*. *FEMS Microbiol. Lett.* **2010**, *302*, 166-174.

(56) Saiki, K.; Konishi, K. Identification of a novel *Porphyromonas gingivalis* outer membrane protein, PG534, required for the production of active gingipains. *FEMS Microbiol. Lett.* **2010**, *310*, 168-174.

Figure Legends

Figure 1. Protein logo representations of the motifs present in the C-terminal 80 amino acids of 30 experimentally identified CTD proteins from *Porphyromonas gingivalis* and *Tannerella forsythia* produced by GLAM2 (meme.sdsc.edu). The GLAM2 score for the full alignment was 946.6. The boundaries of Motif B, Motif D and Motif E were defined by MEME analysis (Supplementary Figure 1). The colours of each amino acid were based on the biochemical properties of the amino acids. A, C, F, I, L, V, W & M are most hydrophobic and coloured blue. N, Q, S & T are polar, non-charged, non-aliphatic residues and are coloured green. D & E are acidic and coloured magenta. K & R are positively charged and are coloured red. H is coloured pink, G is coloured orange, P is coloured yellow and Y is coloured turquoise.

Figure 2. Multiple sequence alignment of the CTD motifs from 30 experimentally identified CTD proteins from *Porphyromonas gingivalis* and *Tannerella forsythia*. Labelled are regions of Motif B, Motif D, Motif E and the variable gaps between Motif B and Motif D. Motif logo representation of the multiple sequence alignment is shown in Fig 1. The amount of amino acid conservation in each column can be seen in the bar graph below the alignment.

Figure 3. Distribution of 663 predicted CTD proteins that were detected by HMM methodology in 21 *Bacteroidetes* organisms organized in a phylogenetic tree. The phylogenetic tree was obtained from NCBI taxonomy common tree (www.ncbi.nlm.nih.gov/Taxonomy/CommonTree). Species that were predicted to have true CTD proteins were labelled blue while red represents species that do not have true CTD

proteins. The predicted CTD proteins in the red coloured species were lacking one of the three motifs or the motifs positions were wrong in the sequences.

Figure 4. Identification of CTD proteins from SDS-PAGE gels. Each gel was excised into segments as shown. The identified CTD proteins are shown as shaded rectangles at the correct MW relative to the gel and with their accession number shown above. The shading reflects the Mascot score with darker shaded proteins being identified with higher scores and therefore likely to be of higher abundance. The calculated MW of each protein excluding signal peptide, pro domain (if any) and CTD is shown by a bar. The MW of bars outside the scale of the gels are provided. The accession numbers of PG0026 and PG1604 homologs are shaded and are found close to their calculated MW (see text). A) *P. gingivalis* vesicles. Only CTD proteins elevated in MW are shown. *PG1837 (HagA) and PG1844 (Kgp) have similar A5 adhesin domains. B) *P. distasonis* membrane fraction. In addition to the gray scale, Mascot scores are also shown in each rectangle. C) *P. intermedia* membrane fraction. D) *C. hutchinsonii* membrane fraction

Figure 5. Sequence coverage obtained for selected CTD proteins. Peptides identified from membrane or vesicle samples are shown in red. The CTD is shown in blue based on the cleavage site shown in Table 1 or previously reported¹⁷. The signal peptides are shown in green. A) PG2102 (P59) B) PG0553. C) BDI_2562. D) The C-terminal region of five CTD proteins identified from *P. intermedia* showing the most C-terminal peptide identified (in red) demonstrating very short CTDs. The numbers show the number of residues from the C-terminus.

Figure 6. Clustal W alignments of full length CTDs (from cleavage site to C-terminus).

Identified semi-tryptic peptides are highlighted yellow and mark experimentally determined cleavage sites. The MS data for these peptides is shown in Supplementary Table 3, or in the case of *P. gingivalis* from Glew et al., 2012¹⁷ & data not shown. A) *P. gingivalis* CTD proteins from Fig 4. B) *T. forsythia* CTD proteins from Veith et al., 2009²⁵. C) *P. distasonis* CTD proteins with identified cleaved CTDs and CTD proteins listed in Figure 4. D) Single CTD sequence from *C. hutchinsonii* with identified semitryptic peptide. Due to only one cleavage site being determined in this species, reliable prediction of cleavage sites in other CTD proteins was not possible.

Table 1. Motif elucidation of *Porphyromonas gingivalis* and *Tannerella forsythia* CTD proteins by MEME (meme.sdsc.edu). All three motifs had significant E-Values validating their existence.

Motif	Length	E-Value	Regular Expression
B	13	1.9e-56	[VA]x[ILV][YF]D[ML]NG[RK]LV[ARY]A
D	11	7.7e-57	x[LA][PQ]KG[VT]Y[VI][VL]R[VI]
E	7	2.4e-04	Y[VT]EK[VIL]I[VI]

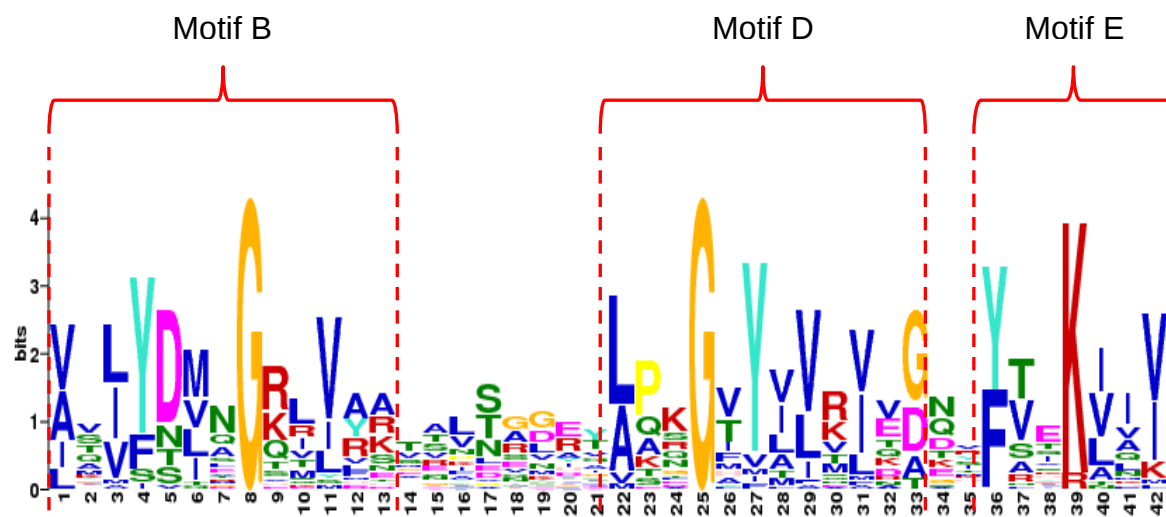


Figure 1.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

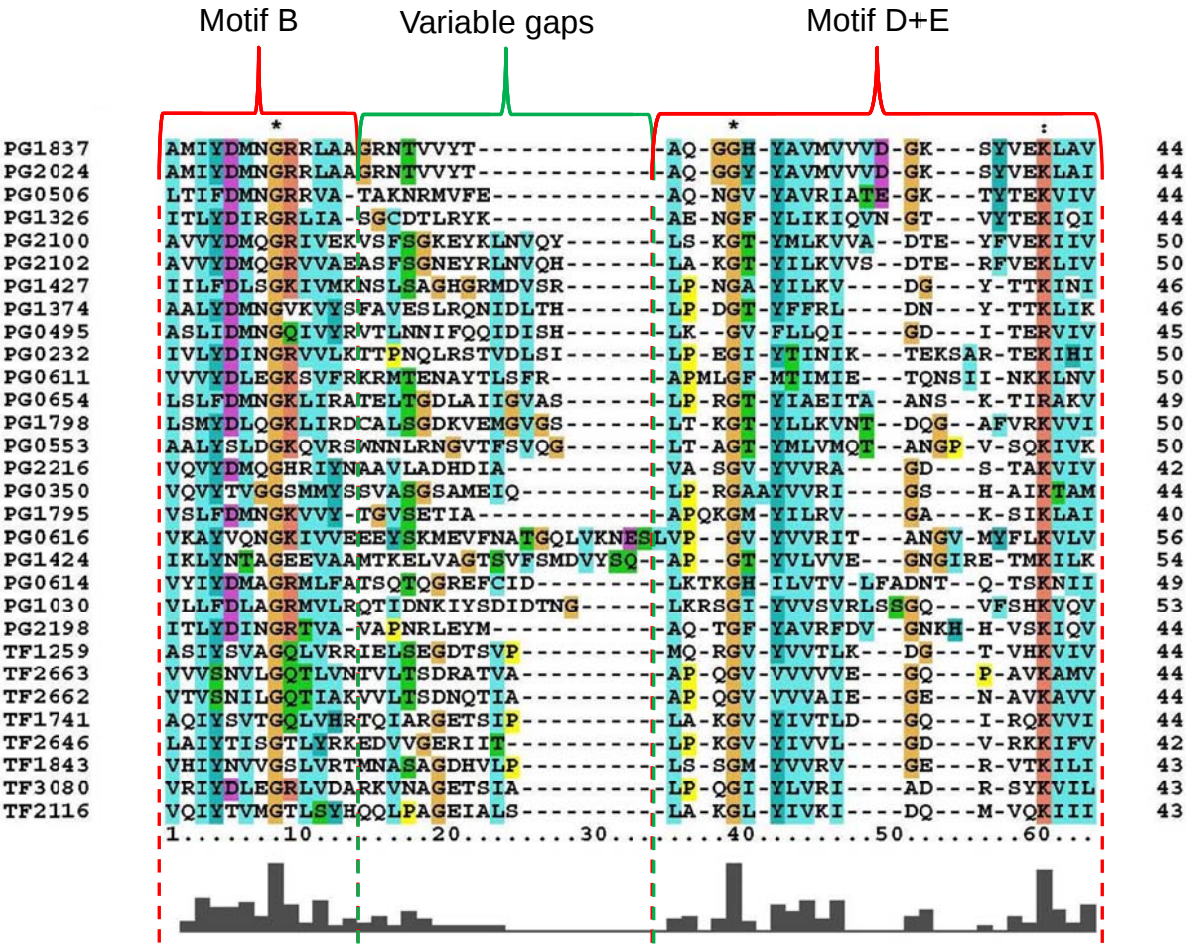


Figure 2.

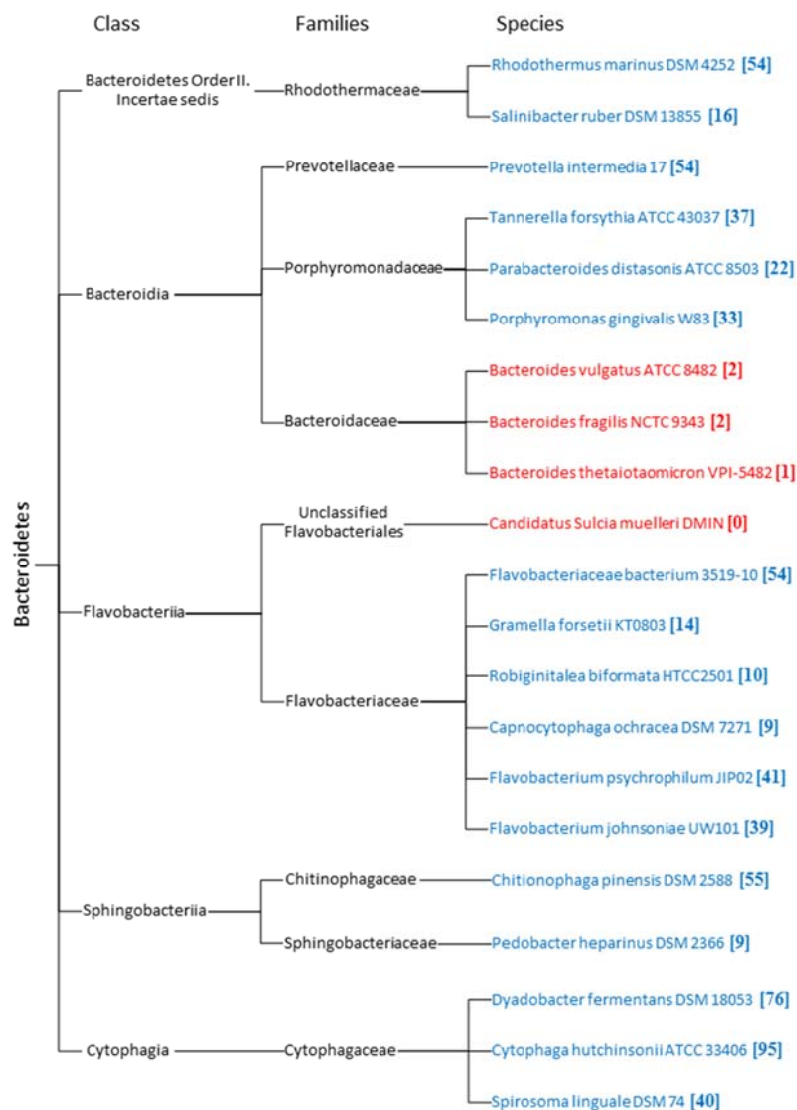


Figure 3.

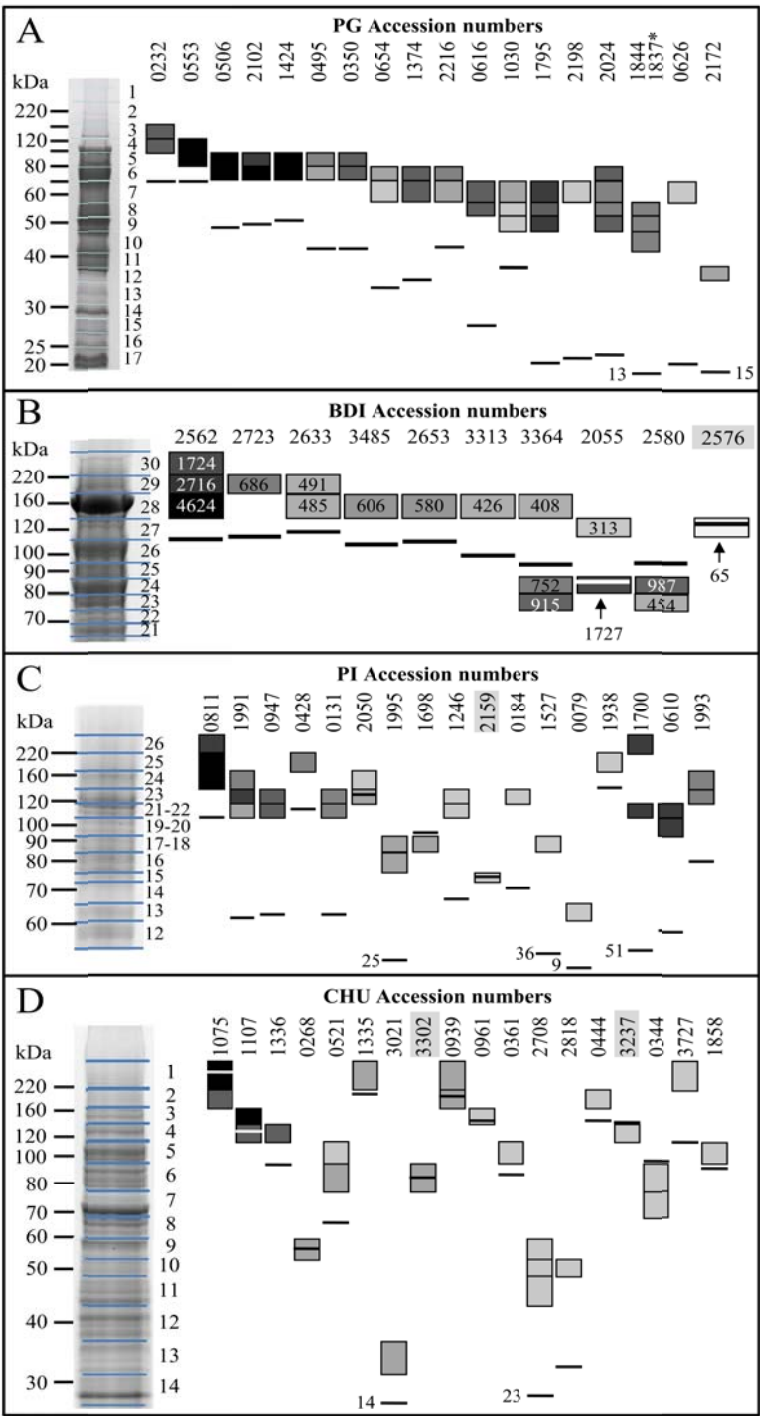


Figure 4.

A	1	MKTKVLRKFFV	VAAFAVATLC	PLAQAQTMGG	DDVKVVQYNQ	EKLVTQTRMSV	ADNGWIYVMT	HSGYDTGNSN
	71	VKIFRSKDQG	ATYQKLRDWD	PSDDYQFQDF	DIVVTGKNES	DIKIWSVELM	NKPGGYKSRV	AVFSRDANAQ
	141	NAKLVIKEDF	SNVQLYDVDI	ASNYSRPSL	NNGGNPFALA	FAYTGFNTH	KISFVDYVFS	LNGGQNFNKN
	211	LLFSQDGEKK	IDKVDLSLGS	TSESMGHNAW	PLMGVVFEMN	KQGGKSDIGF	LSNFVDNDPE	FQWSGPIKVS
	281	ESDMSFSPKI	QMLLEDNNT	INGESCHNEM	ITYSDYDSEY	SDWDIRYVYP	KKSFKYEKGK	TPTMDDLVEA
	351	FLTASYQSET	NSGLGYDKNA	NHYLITYAKK	EENGNTNLKY	RWANYDKIHN	KDLWSDTFTY	TSSANALYTP
	421	QVDINPTKGL	VCWSWVEYLP	GKRIVWSDTQ	WTHANGVEDI	VMQEGSMKLY	PNPAQEYAVI	SLPTAANCKA
B	491	VVYDMQGRVV	AEASFSGNEY	RLNVQHLAGK	TYILKVVSDT	ERFVEKLIVE		
	1	MNKFYKSLQ	SGLAAFVSM	TALTASAQIS	FGGEPLSFSS	RSAGTHSFDD	AMTIRLTPDF	NPEDLIAQSR
	71	WQSQRDGRPV	RIGQVIPVDV	DFASKASHIS	SIGDQDVYRL	QFKLEGAKAI	TLYYDAFNIP	EGGRLYIYTP
	141	DHEIVLGAYT	NATHRRNGAF	ATEPVPGESE	IMDYEVSRGG	TLDPDIKISGA	GYIFDKVGGP	PVTDNHYGIG
	211	EDDSDSDCI	NINCEPADW	QAEKNGVVM	IMVKGQYISM	CSGNLLNNTK	GDFTPLIISA	GHASITTNF
	281	GVTQSELDKW	IFTFHYEKRG	CSNGTLAIFR	GNSIIGASHK	AFLPIKGKSD	GLLLQLNDEV	PLRYRVYNG
	351	WDSTPDIPSS	GAGIHHFAGD	AMKISILKKT	PALNTWISSS	GSGGTDDHFY	FKYDQGGTEG	GSSGSSLFNQ
C	421	NKHVVGTLTG	GAGNCGGTEF	YGRNLNHWNE	YASDGNTSRM	DIYLDQNNNG	QTTILNGTYR	DGYKPLPSVP
	491	RLLLQSTGDQ	VELNWTAVPA	DQYPSYQVE	YHIFRNGKEI	ATTKELSYSD	AIDESIIGSG	IIRYEVSARF
	561	IYPSPLDGE	SYKTDKTS	DLAIGDIQTK	LKPDVTPPLG	GGVSLSWKVP	FLSQLVSRFG	ESPNPVFKTF
	631	EVPIVSAAAA	QTPNPPVGVV	IADKFMAGTY	PEKAAIAAVY	VMPSADPSTF	HLFLKSNTRN	RLQKVTTSPD
	701	WQAGTWLRIN	LDKPPFVNND	HMLFAGIRMP	NKYKLNRAIR	YVRNPDNLFS	ITGKKISYNN	GVSFEGYGIP
	771	SLGYMAIKY	LVVNTDAPKI	DMSLVQEPYA	KGTNVAPFPE	LVGIYVYKNG	TFIGTQDPVS	TTYSVSDGTE
	841	SDEYEIKLVY	KSGGISNGVA	QIENNNAVVA	YPSVVTDRFS	IKNAHMHVAA	ALYSLDGKQV	RSWNNLRNGV
D	911	TFSVQGLTAG	TYMLVMQTAN	GPVSQKIVKQ				
	1	MNKKFSTLLC	ASLLFSSAFT	AQAEDLSGAL	KGGDATEMSE	LGSTTPKGVY	QLRTSDGKIL	TVNDKGEYKL
	71	VTSANLSVDK	SSLWCQVITE	AQDRGQQPIY	DFVNGQGSAM	LAINVPDGMK	VGETTAALSV	GESYGGWAFS
	141	SVYKESLERN	QTLTYTQEAD	NVLILLAEGL	GIAKKIKAN	EAENYQDALK	FTLFNAGTYV	LSAAEINAYL
	211	KENKNVLKFN	PDANADVNP	STTAFVAKGI	NSDTHNFVY	VTDAKENS	LKVDTAANGV	GTQFLKFGWT
	281	DVKAKGSEDA	LTSGEIANQH	KFLFTYRPSG	DSLFIQVQEA	RYLPEKRTEN	DYWKDMTPIK	YGTVYHAVTC
	351	EGTDSADSLF	VKLQNFVAD	RIATIGKKPI	NTHISFGVDG	CFTSTDKTSV	NDGVVVIKNA	KGEVLAAPIH
E	421	LNDNVAGEAN	NDVQVVKLDA	QDPKHMPAYQ	WVVLKTMHSD	KAIETSPVVI	TNREFFQFTV	TVQFTKNEKG
	491	ELVGNTSIDN	KVLSAVTFDA	LEESLIKDAK	LGKYISDAE	LSVNKYKFN	LNPFTPDYWI	ANGDGKDSLL
	561	YVKKEASRYE	LVKGSTAAAYG	YKVTDEAKKW	IKGLAQLERT	NYVVKLANTD	KYLREEFVYG	QKYAMGAIAT
	631	PDTTTTFKENN	HDGAKHFIYAI	LKSHYVANE	GFDYKYAANI	ENVATTHKVG	IADDGKEAGL	KVELLNEKRT
	701	SSFSVEPDNT	PLYRRFNNEL	LGENVNDASD	SLYFVESIRK	EYLMDENNTN	LQHSTVDYAG	IWNKDKANGK
	771	LAFRVDTAWL	KRGAGEIKPQ	YLISAFQDVI	AGEKIIPCTE	DGKHIPDGT	PTDNPYECVH	AVHIKTPAFV
	841	YGKYLNVFSD	SAATCEADVN	PFIFNTKAVS	NSSYTRVGTV	PAVQYGDTL	VLTKGYASMT	KEQLKNEGIE
F	911	SMIATYKKEY	PSFINYLNGD	NHKNVTSFR	YVNEPAAAVA	YEDGKEGENN	AFLIESNVYA	AGEKYQTVKG
	981	NKDEAIAPEK	AAAWLKMHN	CLVLTDVNAT	FDNAKTGGDG	ALVFNAFQKT	EKEDMVTSD	BIATEGVSVV
	1051	AGNGTIVTVQ	AAGKSVVITN	ILGKVVAETV	LTSDNATIAV	PAGIVAVAVE	GEEAVKTIK	

Figure 5.

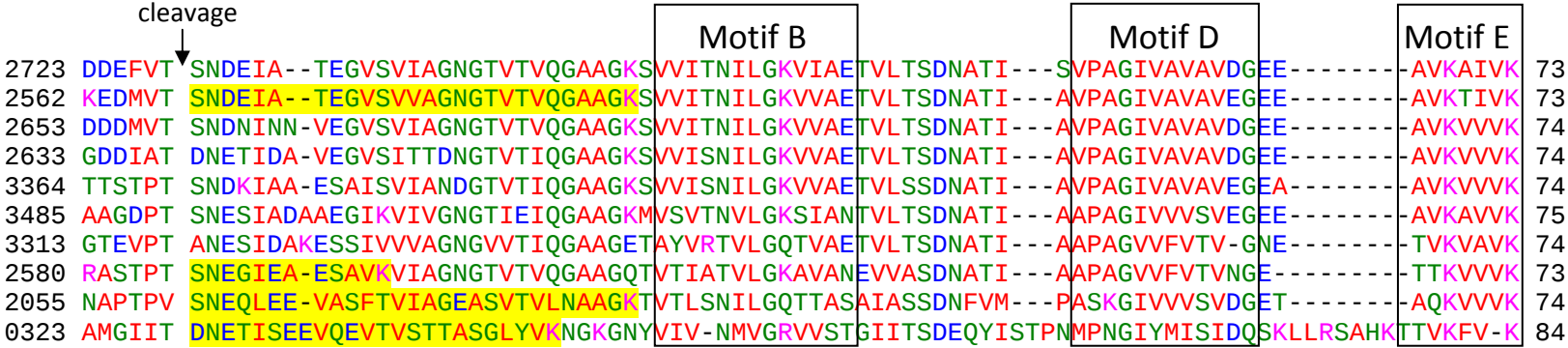
A) *P. gingivalis* CTDs

		cleavage		Motif B		Motif D		Motif E	
1		↓							
2	1844		ATLNIT	SLADVTA--QKPYTLTVVG---K	TTITVTCQ--GEAMIYDMNGRRLAAG-----RNTVVYTAQGG-HYAVMVVVDGK-SYVEKLAVK	72			
3	2024		VDYIPD	GVADVTA--QKPYTLTVVG---K	TTITVTCQ--GEAMIYDMNGRRLAAG-----RNTVVYTAQGG-YYAVMVVVDGK-SYVEKLAIK	72			
4	0506		VKVEGT	SIADVAN--DKPYTVAVSG---K	TTITVESP--AAGLTIFDMNGRRVATA-----KNRMVFEAQNG-VYAVRIATEGK-TYTEKVIVK	73			
5	2198		VTVTNS	SLSNVDG--QAPYTLRVEG---K	KIIIAEAH--GMITLYDINGRTVAVA-----PNRLEYMAQTG-FYAVRFDVGNK-HHVSQIQR	72			
6	1795		HLKKGE	GVEAVLTN-DKANAYVQNG---V	IYVAGAN--GRQVSLFDMNGKVVTG-----VSETIAAPQKG-MYILRVGAKS-----IKLAI	69			
7	2216		NKPVIT	SLAAPIS--HEIRIWATAG---R	IFIAGAPA-GTSVQVYDMQGHRIYNAV-----LADHDIASG-VYVVRAGDST-----AKVIVP	72			
8	0350		PFPGIT	ALISIEGE-SEYSVYAQDG---I	LYLSGMEQ-GLPVQVYTVGGSMYSSVA-----SGSAMEIQLPRGAAYVVRIGSHA-----IKTAMP	75			
9	1030		KESFIT	SFISPTVV-QGVDVYTLAG---K	IRIESETP-VSEVLLFDLAGMRVLRQTI--DNKIYSDIDTNGLKRS-G-IYVSVRLSSGQVFSHKVQV	83			
10	0616		DKAAPT	ATEQIVAT-PSVKAYVQNG---K	IVVEEY--SKMEVFNATGQLVKNE-----SLVPG-VYVVRITANGV-MYFLKVLVP	67			
11	2172		YKGGGT	GLTNIGL---GRIALIQSGN-TCT	LQYNSNGK-RLALEVYNLLGVKVFSQLPAGS---GSYTLPVRLQRG-VHIFRITEGGKPAFVQKYLK	84			
12	0232		EVILDD	SVEDIVAQ-TGIVIRPQNG---T	KQILIEANAIAKAIIVLYDINGRVVLKTPN-QL---RSTVDLSILPEG-IYTINIKTEKS-ARTEKIHIG	83			
13	0654		TDMQGN	ALTDVAVN-ESIKIYPRPAT---D	FLRIEGSQ-LLRLSLFDMNGKLIRATELTG-D---LAIIGVASLPRG-TYIAEITAANSKTIRAKVSLR	83			
14	0626		RILNYQ	SLQEVEQ--EGIRIYPNPT--TDR	LMLDNVGD-VSAIRIIDMYGRPLVVGNAGSS---KLSIDVSDLGRG-NYVVDMQCKGH-RKISRLITKE	84			
15	2102		QWTHAN	GVEDIVMQEGSMKLYPNPAQ-EY	AVISLPTAA-NCKAVVYDMQGRVVAEASFSG-N---EYRLNVQHLAKG-TYILKVVSDE-RFVEKLIVE	85			
16	0495		LMKKS	AVGEVDT--KGFHVYPIPTSKDLT	IEIPAEMV-GKVASLIDMNGQIVYRVTLN--N---IFQQIDISHLKG-VFLLQIGDIT-----ERVIVQ	79			
17	1374		YEGSPT	SNLAVDA--PTVRIYPNPVGRYAL	VEIPESLL-GQEAAALYDMNGVKVYSFAVES-L---RQNIDLTHLPDG-TYFFRLDNYT-----TKLIKQ	80			
18	0553		GSGISN	GVAQIENN-NAVVAYPSVVT--DR	FSIKNAHM-VHAAALYSLDGKQVRSWNNLR-N---GVTFSVQGLTAG-TYMLVMQTANG-PVSQKIVKQ	83			
19	1424		NTCTVT	GAAKAL-----RAWFNAGRSELAV	SVSLNIAGTYRIKLYNTAGEEVAAMTKELVAGTSVFSDVYSQAPG-TYVLVVEGNGI-RETMKILK	84			

B) *T. forsythia* CTDs

24		cleavage							
25		↓			Motif B		Motif D		Motif E
26	2645		RKPGDD	ATVE-VNDQSIYAS--EGYVHIRSGKPVSLRIY	ILGALYRHEERLVGDHVIA-LPRGLYIVVLNHTA-KKIMVR	70			
27	2646		TVRKTN	ASED-IVEN-VDIWSFDNSLRIRTDQLRPLAIY	TISGTLRKED-VVGERIIT-LPKGVIIVVLGDVR-KKIFVQ	70			
28	1741		TGVSPV	SNVS-IDRG-YALRTEEGALHIRTAKATDAQIYS	VTGQLVHRTQIARGETSIP-LAKGVYIVTLDGQIRQKVIR	72			
29	0193		NGVSPV	SNAS-MP-Q-HKVWAHRQHLYIEAQYATIMQLY	TLDGKLHTQQMIPQGNTSFP-LAAGFYTVVMDGRV-YKVMVE	70			
30	1259		FGPDHV	ANMS-IE-G-TAVWSHGTEIYIRVERKDIASIYS	VAGQLVRRIELSEGDTSP-MQRGVYVVTLKDGTVHKVIVK	71			
31	2320		SGVSPV	GNQE-IIRKVSRAWAYEGKIYVETDRPTTVYV	FTTMGHLYRREKVQAGLTTFDAPQPGVYFVKFDNGYSGKILIE	74			
32	3163		SGITPL	ANTN-ITEQPQKAWAHKGLYVETERNETVYIY	TVTGQLYKREDAQAGLSVFD-LESGYYIIRLASGYSGKVFIE	73			
33	2116		SGIKPT	ANQL-IEEYAAVWADRNG-IHIRTDPAMVQIY	TVMGTLSTYHQQLPAGEIALS-LAKGLYIVKIDQ-MVQKIIK	71			
34	1589		DVIRTV	ANET-ID-G-LRIYAADGALRLALPAPETVHIY	NVEGSLVKTLILPAGDHVQP-LPSGVYVVRVGERV-TKVMVK	70			
35	2922		EVLSAV	ANET-ID-G-LRIFAADGTLRLTLPSPATAHIY	NVNGAMVKTLALPSGDHVQPLAP-GVYVVRVGEHV-TKVMVK	70			
36	0535		EVISTV	GNET-ID-G-LRVYAVDGALHLTLAPQIVHLY	DVSGALVRTMDAPAGDHVLP-LAP-GVYLVRIGEQI-AKVLIK	70			
37	1843		TFTHTV	ANQT-ID-G-LRVYAADGALHLTLPKAETVHIY	NVVGSLVRTMNASAGDHVLP-LSSGMVVRVGERV-TKILIK	70			
38	1985		AYIAEV	ANQERID-G-LRIYVAHGILYLTLPAPQTVQIY	NVNGMAVKILTLPGEHIEP-LAQGIYLVRIGDTV-EKILVK	71			
39	2662		TKKSPL	ANAEA-SVTDVKVAVAGAEVMIRNASGKKVTV	SNILGQTIKVVLTSDNQTIA-APQGVVVVAIEGENAVKAVK	73			
40	2663		TKEDPV	ANEITPAITEVKVVAENGAVTILNAAGKVVV	SNVLGQTLVNTVLTSDRATVA-APQGVVVVVVEGQPAVKAMVK	74			
41	1458		LTVDPH	ANSSAVTGE-KVWSYKDHVLLSVKASTDVRIY	DLEGRVLARKVNAGETSIA-LPQGIYLVRIADRS-YKVILR	72			
42	1032		SLTEPV	AN-IVVE-SEVSFRTQREQLCTVTRPVTVHIY	TIIGALLRKQNLNEGECFT-LPQGMIVSLSNGTVRKIVIP	72			
43	1273		KAGNPH	SNLVIEKPG--KVYAAEGSVVIENEQPALVAVY	TFGGQLVEQRRMAAGINRIP-LPKGVIIVVGRQG-HKIVVR	71			
44	0537		KGGEEN	ANELIVDDSNSRVWAADGILYLSTEQPIDVEV	FSFSGLLKKFKASTGTTPEHLPTGLYVVRVGQQT-TKVIVR	74			

C) *P. distasonis* CTDs



D) *C. hutchinsonii* CTDs

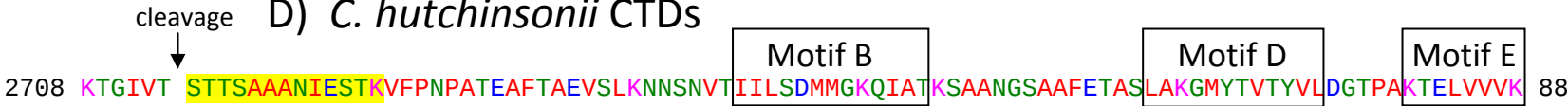


Figure 6.