Polysaccharide utilization locus and CAZYme genome repertoires reveal diverse ecological adaptation of *Prevotella* species

Tomaž Accetto*, Gorazd Avguštin

University of Ljubljana, Biotechnical Faculty, Animal Science Department, Groblje 3, 1230 Domžale, Slovenia

ABSTRACT

The results of metagenomic studies have clearly established that bacteria of the genus *Prevotella* represent one of the important groups found in the oral cavity and large intestine of man, and they also dominate the rumen. They belong to the *Bacteroidetes*, a phylum well-known for its polysaccharide degrading potential that stems from the outer membrane-localized enzyme/binding protein complexes encoded in polysaccharide utilization loci (PULs). Dozens of *Prevotella* species have been described, primarily from the oral cavity, and many of them occur simultaneously at the same sites, but research on their ecological adaptation has been neglected. Therefore, in this study, the repertoires of PULs and carbohydrate acting enzymes (CAZYmes) found in *Prevotella* genomes were analyzed and it was concluded that the *Prevotella* species were widely heterogeneous in this respect and displayed several distinct adaptations with regard to the number, source and nature of the substrates apparently preferred for growth.

© 2015 Elsevier GmbH. All rights reserved.

Introduction

The anaerobic bacterial species of the genus *Prevotella* are members of the large bacterial phylum *Bacteroidetes* that encompasses polysaccharide and protein degraders found in seawater, fresh water and soils, as well as the animal gut [56]. The *Bacteroidetes* are special in several ways: the base composition and spacing of their promoters are distinct [6], the Shine-Dalgarno sequence is not used in translation initiation [1], and their capability to degrade a wide array of carbohydrates depends on the starch utilization system-like (Sus-like) multiprotein complexes that bind and partially degrade substrates prior to their transport into the periplasm where final breakdown occurs [37]. The central elements of the Sus-like systems are encoded by the paralogues of *Bacteroides thetaiotaomicron* VPI-5482 *susC* and *susD* whose products were shown to account for >60% of starch binding, while SusC also transfers oligosaccharides across the outer membrane [37]. The three SusD-like proteins with hitherto known structures, the *B. thetaiotaomicron* VPI-5482 proteins SusD, BT1043 and BT3984, are structural homologues but their binding strategies differ: for SusD it is thought that the overall helical shape of amylose is recognized, while for BT1043 the specific hydrogen bond interactions are

crucial [5,30,31]. *susC* and *susD* are usually the central elements of larger gene clusters, the polysaccharide utilization loci (PULs), that are also formed of genes coding for carbohydrate active enzymes (CAZYmes) [34], response regulators and transporters, which all contribute to efficient degradation of a single substrate [37]. The genomes of gut-dwelling *B. thetaiotaomicron* VPI-5482 and *B. ovatus* ATCC 8483 contain 88 and 110 PULs, respectively [38]. The PULs are inducible and transcriptional profiling has revealed the substrates of numerous *B. thetaiotaomicron* VPI-5482 and *B. ovatus* ATCC 8483 PULs [36,38]. These bacteria occupy distinct niches in the gut and they both degrade starch, storage polysaccharides and pectin, although *B. thetaiotaomicron* VPI-5482 cannot degrade hemicelluloses, while *B. ovatus* ATCC 8483 cannot use host mucin O-glycans for growth, which is mirrored in their PUL repertoires [38]. These repertoires are plastic due to lateral gene transfer, which is attested to by a recent transfer of PULs specialized for the degradation of porphyran from a marine bacteroidete into the gut-dwelling *B. plebeius* 17135 isolated in Japan, probably due to consumption of non-sterile red algae [27].

According to Bergey's Manual [32], the genus *Prevotella* contains 34 species and at least a further 13 species have been described in recent years [2,15–19,26,33,44–46,53,57]. The majority of prevotellas have been isolated from the oral cavity of mammals, whereas other species inhabit the rumen, large bowel and urogenital tract, and some species have been described from clinical specimens only [2,15–19,26,32,33,44–46,53,57]. While some oral species are considered true pathogens (e.g. *Prevotella intermedia* contributes to periodontitis [10]), other oral species may be

---

opportunistic pathogens as they are frequently isolated from a variety of polymicrobial abscesses [7]. However, the ruminal and large bowel prevotellas are symbionts that contribute to polysaccharide breakdown [25]. It is noteworthy that *Prevotella* are widespread in oral cavity surfaces, account for more than 10% of all bacteria in saliva [48] and, on average, 11 *Prevotella* species may be found at the same time in healthy gingiva alone [40], yet the niches they occupy in the oral ecosystem remain obscure.

In order to gain an insight into their roles in the oral microbiota using genomic data, we cataloged, annotated and compared the PUL and CAZYme genomic repertoires of 39 *Prevotella* species that were already sequenced. A clear grouping of *Prevotella* species emerged that reflected the number, origin and variability of the polysaccharides most likely used for growth by different species. Furthermore, the rumen/bowel and urogenital tract species provided examples of *Prevotella* adaptation to non-oral habitats.

Materials and methods

The genomes

The *Prevotella* genomic data were obtained in February 2014 from the EBI and GenBank. The strains and degrees of their genome sequence completion are listed in Table S1. Genomes that contained only sequence data were annotated using Prokka 1.7 [47]. The *Hallella seregens* ATCC 51272 and *Alloprevotella rava* F0323 genomes were included since *H. seregens* was recently reclassified as *Prevotella dentalis* [59] and *A. rava* belongs to a recently proposed related genus that now also encompasses the former *Prevotella tannerae* [20].

Identification of susCD-*like loci in* Prevotella *genomes and clustering of* Prevotella *SusD-like proteins on the basis of their amino acid similarity*

Since all hitherto described PULs contain *susC* and *susD* in tandem [36], we searched for putative PULs in *Prevotella* genomes using SusD-like proteins as proxies with HMMERv3 hmmsearch [22] and the SusD.hmm, SusD-like_2.hmm and SusD-like_3.hmm models downloaded from PFAM [24]. Structures of *B. thetaiotaomicron* VPI-5482 proteins SusD, SusD-like BT1043 and BT3984 have already been resolved and they display similar folds, although the amino acid identities are low and blastp [8] searches fail to detect O-mucin specialized BT1043 as a match to starch-binding SusD [5,30,31]. Therefore, we reasoned that SusD-like proteins from PULs with the same target substrate in various *Prevotella* genomes should exhibit significant amino acid identity, whereas the hits to SusD-like proteins binding other substrates would be insignificant. SusD-like proteins were thus compared among themselves by a blastp search using blast 2.2.27+ [8], and hits exhibiting >50% coverage and >40% amino acid identity were retained. The proteins exhibiting high amino acid identity were clustered with mcl [58] using an inflation parameter of 1.5. The SusD-like protein profiles were then generated by noting the presence/absence of a specific SusD-like protein group in the individual *Prevotella* genome.

Assigning putative substrates to identified Prevotella *PULs*

The results obtained through transcriptomic profiling of *B. thetaiotaomicron* VPI-5482 [36,38], *B. ovatus* ATCC 8483 [38] and *Prevotella bryantii* B$_1$4 [14] were used to infer most probable substrates for the identified *Prevotella* PULs. SusD-like protein groups that contained five or more members were manually investigated. For each member in a group, the immediate genome neighborhood of the *susCD*-like genes was inspected using Artemis [9] and searched using blastp against the *B. thetaiotaomicron* VPI-5482,

B. ovatus ATCC 8483 and *P. bryantii* B$_1$4 proteins. If a significant match was found, the matched protein was searched for among the proteins known to be coded in PULs whose substrates had been identified in transcriptomic profiling [14,36,38], thereby revealing the putative substrate of the *Prevotella* PUL. Additionally, the protein family data [24,35] for the *susCD*-like neighborhoods were used for evaluation of the conclusions obtained using the blastp search.

SusD-like protein phylogeny reconstruction

The maximum likelihood tree was reconstructed by MEGA6 [54] using a JTT amino acid substitution matrix and 1000 bootstrap replicates.

CAZYme and MEROPS annotation of Prevotella *genomes*

The CAZYme repertoires of *Prevotella* species were obtained using the CAZYme family hmm models available in dbCAN release 3.0 [62]. The *Prevotella* proteins were searched for against the dbCAN hmm models using HMMERv3 hmmscan [22] and the suggested E-value cutoff was applied [62]. The peptidase content of *Prevotella* genomes was estimated using the Merops peptidase database [43] compilation of peptidases available in merops_scan.lib (downloaded in May 2014) used in Merops batch-BLAST (http://merops.sanger.ac.uk/cgi-bin/batch_blast). The searches were carried out locally using blastp 2.2.27+ [8] and an *E*-value cutoff of 1e−10.

Hierarchical clustering of Prevotella *species according to their SusD-like protein and CAZYme repertoires*

Clustering was carried out with R package pvclust [52] using average linkage and 10,000 bootstrap replicates. For CAZY repertoires, the distance measure chosen was Manhattan, since it is known that enzymes from the same CAZY family may act on several substrates or in different ways on the same substrate [50]. Thus, the Manhattan coefficient would capture both the variability of CAZYme families in a genome as well as the number of enzymes in a family. The CAZY glycoside hydrolases, polysaccharide lyases and carbohydrate esterases were included in the analysis. For SusD-like protein profiles, the binary distance metric was chosen, since we wanted to emphasize the degree of shared SusD-like proteins that may reflect the common substrates bound by *Prevotella* species and the index should not be affected by the total number of SusD-like proteins per genome, which in fact varied greatly. The SusD-like protein repertoire comparison was based on groups with more than one member.

Ordination analysis of annotated SusD-like protein and CAZYme Prevotella *repertoires*

Non-metric multidimensional analysis was performed on data contained in Table S3 using R package Vegan [42]. The distance measure chosen was Bray–Curtis.

Results

Diversity of Prevotella *SusD-like protein repertoires*

Almost 1200 full-length SusD-like proteins were detected in 50 *Prevotella* genomes, and there was a wide variation in their number per genome. A distinct reduction in the number of SusD-like proteins to four or less was seen in several *Prevotella* species (i.e. *P. corporis*, *P. disiens*, *P. intermedia*, *P. micans*, *P. nigrescens*, *P. pallens* and *P. tannerae*) (Table 1). The SusD-like proteins were divided by mcl into 274 groups of which 99 contained at least