INFECTION AND IMMUNITY, Mar. 1990, p. 667–673
0019-9567/90/030667-07$02.00/0

Vol. 58, No. 3

Sequence Analysis of the Gene for the Glucan-Binding Protein of *Streptococcus mutans* Ingbritt

JEFFREY A. BANAS,[1] ROY R. B. RUSSELL,[2] AND JOSEPH J. FERRETTI[1]*

Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73190,[1] and Dental Research Unit, Royal College of Surgeons of England, London Hospital Medical College, London E1 2AD, United Kingdom[2]

Received 22 August 1989/Accepted 21 November 1989

The nucleotide sequence of the *gbp* gene, which encodes the glucan-binding protein (GBP) of *Streptococcus mutans*, was determined. The reading frame for *gbp* was 1,689 bases. A ribosome-binding site and putative promoter preceded the start codon, and potential stem-loop structures were identified downstream from the termination codon. The deduced amino acid sequence of the GBP revealed the presence of a signal peptide of 35 amino acids. The molecular weight of the processed protein was calculated to be 59,039. Two series of repeats spanned three-quarters of the carboxy-terminal end of the protein. The repeats were 32 to 34 and 17 to 20 amino acids in length and shared partial identity within each series. The repeats were found to be homologous to sequences hypothesized to be involved in glucan binding in the GTF-I of *S. downei* and to sequences within the protein products encoded by *gtfB* and *gtfC* of *S. mutans*. The repeated sequences may represent peptide segments that are important to glucan binding and may be distributed among GBPs from other bacterial inhabitants of plaque or the oral cavity.

The ability to synthesize extracellular glucans is generally believed to be one of the virulence properties of *Streptococcus mutans* which contributes to plaque formation and to the subsequent development of dental caries (11). The initiatory component of plaque produced by *S. mutans* is an insoluble glucan called mutan, which is synthesized by the glucosyltransferase (GTF) enzymes. The precise mechanisms whereby mutan, receptors on the *S. mutans* cell surface, and various glucan-binding proteins (GBPs) interact to form plaque are unknown.

S. mutans produces an extracellular protein designated GBP which becomes associated with the cell in the presence of sucrose (1a, 15). There is some evidence that the GBP may be involved in the formation of cohesive plaque, since a GBP-deficient mutant forms loosely adherent plaque in vitro (17). The GBP has been shown to copurify with a fructosyltransferase (FTF) (17), although there is no such activity expressed from the *gbp* gene cloned in *Escherichia coli* (16). Furthermore, the genes for GBP and that for FTF from *S. mutans* have distinct restriction maps, and there is no antigenic cross-reactivity between their products (1). The relationship between GBP and FTF therefore remains unclear.

In this report, we present the nucleotide sequence of *gbp* and an analysis of the sequence and the putative protein that it encodes. By analyzing the *gbp* gene, it is anticipated that the role(s) of GBP in caries etiology can be defined more precisely.

MATERIALS AND METHODS

Bacteria and media. The cloned *gbp* gene from *S. mutans* Ingbritt was obtained in *E. coli* JM109 containing the plasmid pMLG43 (1). This is a low-copy-number plasmid and contains a 4-kilobase-pair (kb) insert from the lambda clone of *gbp* (16) in the vector pGD103, a derivative of pLG339 (1, 21). The M13 bacteriophage vectors (26) were used for sequencing. *E. coli* JM109 was used as the host strain for transfection with M13 and was grown in 2× YT broth (13). Detection of recombinant phages was accomplished by using soft agar (0.75%) overlays of 2× YT broth base supplemented with 0.33 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and 0.02% 5-bromo-4-chloro-3-indoyl-3-galactoside (X-Gal). *E. coli* isolates that were transfected with recombinant M13 phages were grown in terrific broth (22) for the isolation of single-stranded phage DNA that was to be used as the template in sequencing. For purification of GBP, *S. mutans* Ingbritt was grown in a Casamino Acid (Difco Laboratories, Detroit, Mich.) minimal medium (17). Verification of the presence or purity of GBP in *S. mutans* or *gbp* clones was accomplished by Western immunoblotting with antisera specific for the GBP (1a).

Enzymes and chemicals. Restriction enzymes, exonuclease III, T4 DNA ligase, Klenow fragment, and M13 17-mer primer were all purchased from either Bethesda Research Laboratories, Inc. (Gaithersburg, Md.), or Fisher Scientific Co. (St. Louis, Mo.) and were used in accordance with the specifications of the manufacturers. The deoxy- and dideoxynucleotide triphosphates and dextran T10 were purchased from Pharmacia LKB Biotechnology, Inc. (Piscataway, N.J.), and 7-deaza-dGTP was purchased from Boehringer Mannheim Biochemicals (Indianapolis, Ind.). Some sequencing was done with the Sequenase DNA sequencing kit developed by U. S. Biochemical Corp. (Cleveland, Ohio). The [α-^{32}P]dATP was purchased from either Dupont, NEN Research Products (Boston, Mass.) or ICN Radiochemicals (Irvine, Calif.). IPTG, X-Gal, sodium azide, and sucrose were purchased from Sigma Chemical Co. (St. Louis, Mo.).

Nucleotide sequencing. Fragments to be sequenced were separated after restriction enzyme digestion by electrophoresis (75 mA) in 0.6% low-melting-point agarose and isolated as described by Kuehn et al. (8) or by using the GENECLEAN kit of Bio 101 (LaJolla, Calif.). The isolated fragments were cloned into M13, and sequential deletion clones were derived by limited digestion with exonuclease III by

* Corresponding author.

```
                         16                  32
                          *                   *
     CCG GCT ATA AGT TGA AAT ATT GTA GGT ATT AAA AAC TAT CTT TAG TTT
                     -35                               -10

                         64                  80
                          *                   *
     AGT ATT TAC ATT AAT TTT AAA AAT GTT ATA GTG GAA GTG TCA TGT TGA

                         112                 128
                          *                   *
     TTA CTA TTT TTT TAA GGA GGT AAA ATG ATG AAA GAA AAG ACA CGT TTT
                           RBS            M   K   E   K   T   R   F

                         160                 176
                          *                   *
     AAA CTG CAC AAG GTT AAA AAG CAG TGG GTG GCG ATT GCC GTG ACT AGT
   8  K   L   H   K   V   K   K   Q   W   V   A   I   A   V   T   S

                         208                 224
                          *                   *
     CTA GCT CTA GCT GCG ATA TTG TCA GGA GCT CAC TTG ACT CAG GCT GAG
  24  L   A   L   A   A   I   L   S   G   A   H   L   T   Q   A   E

                         256                 272
                          *                   *
     GAA CAA TCC GGC GGT ACT GAC AGT AAG CCA AGA CTG ACA GCG ACT GTA
  40  E   Q   S   G   G   T   D   S   K   P   R   L   T   A   T   V

                         304                 320
                          *                   *
     CAG GAA AGC TCA GAA CAA CCA ATT ACA AAA GCT CCA GCA GCT GAT TCA
  56  Q   E   S   S   E   Q   P   I   T   K   A   P   A   A   D   S

                         352                 368
                          *                   *
     TCT GTA GAA AAT AAC AGT GCT AAC GCT GTT AAA AGT TCT GAA ACA GCA
  72  S   V   E   N   N   S   A   N   A   V   K   S   S   E   T   A

                         400                 416
                          *                   *
     GAG GCA GCT GAA GTA TCC GAT GGA GGC AGA GCC AGC CAA ACT GAA GCA
  88  E   A   A   E   V   S   D   G   G   R   A   S   Q   T   E   A

                         448                 464
                          *                   *
     GTA ACA AAC CAA ACA AAC TCT GAA GAG CAC CAT CCA GCA GAA AAA GCC
 104  V   T   N   Q   T   N   S   E   E   H   H   P   A   E   K   A

                         496                 512
                          *                   *
     ACA GCC GTT TCT GGA GAA GCT CAG TCA GTG CAA AAT GCT CCA TCA GAA
 120  T   A   V   S   G   E   A   Q   S   V   Q   N   A   P   S   E

                         544                 560
                          *                   *
     AAT GCT GCC CAG CAG GAA ACG GCT AAA ACC GAG CCA GCG ACT GCT GCA
 136  N   A   A   Q   Q   E   T   A   K   T   E   P   A   T   A   A

                         592                 608
                          *                   *
     GAA AAT AAT GAC GCT GCT CCA ACC AAT AGC TTC TTT AAA AAA GAT GGT
 152  E   N   N   D   A   A   P   T   N   S   F   F   K   K   D   G

                         640                 656
                          *                   *
     AAA TGG TAC TAC AAA AAG GCC GAT GGA CAG CTG GCA ACC GGT TGG CAG
 168  K   W   Y   Y   K   K   A   D   G   Q   L   A   T   G   W   Q

                         688                 704
                          *                   *
     ATA ATT GAT GGA AAG CAG CTC TAT TTC AAC CAA GAT GGT AGT CAG GTC
 184  I   I   D   G   K   Q   L   Y   F   N   Q   D   G   S   Q   V

                         736                 752
                          *                   *
     AAA GGA GAA ATT CAT GTG GAG ACA GGG GAT CAA ATC ATT TAT CAT CCT
 200  K   G   E   I   H   V   E   T   G   D   Q   I   I   Y   H   P

                         784                 800
                          *                   *
     GTT TTC ATA AGT GAT TCA CCT TCA GTT TTG GAA GTC AAT AAG ATT TAT
 216  V   F   I   S   D   S   P   S   V   L   E   V   N   K   I   Y

                         832                 848
                          *                   *
     TAC TTT GAT CCT GAT AGT GGT GAA CTC TGG AAG GAT CGT TTT GTC TAT
 232  Y   F   D   P   D   S   G   E   L   W   K   D   R   F   V   Y

                         880                 896
                          *                   *
     TCT AGT TAT GCA GAT CCC CTC CAT TAT GAA AAT ATT AAA CAT GAA GGC
 248  S   S   Y   A   D   P   L   H   Y   E   N   I   K   H   E   G

                         928                 944
                          *                   *
     TGG TTC TAT CTT GGA GAA GAT GGA AAG GCT GCT ATC GGC TGG AGA ACT
 264  W   F   Y   L   G   E   D   G   K   A   A   I   G   W   R   T

                         976                 992
                          *                   *
     ATT GGC GGT AAA AAA TAC TAT TTT GAC ACT AAT GGT GTT CAA GTC AAA
 280  I   G   G   K   K   Y   Y   F   D   T   N   G   V   Q   V   K
```

```
                         1024                1040
                          *                   *
     GGA AAG CTA ATT AGT ACA GAT GGC AAT TAT AAT CTA ATT AGC CAG AAG
 296  G   K   L   I   S   T   D   G   N   Y   N   L   I   S   Q   K

                         1072                1088
                          *                   *
     TAT GGC AAG AAA TCT TTC CTA GAT CCT GAC ACC GGT GAA GCT TGG ACT
 312  Y   G   K   K   S   F   L   D   P   D   T   G   E   A   W   T

                         1120                1136
                          *                   *
     AAT CGT TTT GTC AAT GCA AAG TAT TAT TTC TAC AAC TTT GCA GGA TAC
 328  N   R   F   V   N   A   K   Y   Y   F   Y   N   F   A   G   Y

                         1168                1184
                          *                   *
     GTC TCT ACG ACA GAC TGG TTC TAT ATG GGA GCC GAT GGT ATC GGC GTG
 344  V   S   T   T   D   W   F   Y   M   G   A   D   G   I   G   V

                         1216                1232
                          *                   *
     ACC GAT TGG CAA AAG ATC GAT GGT ATG GAT TAC TAT TTC GAA CCT TCC
 360  T   D   W   Q   K   I   D   G   M   D   Y   Y   F   E   P   S

                         1264                1280
                          *                   *
     AGT GGT ATT CAG GTT AAA GGC GAC ATT GCT GAG CGT GAT GGC AAG GTC
 376  S   G   I   Q   V   K   G   D   I   A   E   R   D   G   K   V

                         1312                1328
                          *                   *
     TAT TAT TTA GAT GAA GAC AGT GGA CAA GTT GTT AAG AAT CGT TTT GGC
 392  Y   Y   L   D   E   D   S   G   Q   V   V   K   N   R   F   G

                         1360                1376
                          *                   *
     ACA ACA CCT GCC GAG CGT ATC AGT ACA GTT GAG GCT CGT TTC CCT AAA
 408  T   T   P   A   E   R   I   S   T   V   E   A   R   F   P   K

                         1408                1424
                          *                   *
     ACT TAT TAT TTT GGA GCG GAC GGT AGC GCG AAA GAT CTA ACT GGT TGG
 424  T   Y   Y   F   G   A   D   G   S   R   K   D   L   T   G   W

                         1456                1472
                          *                   *
     CAG ATT ATT GAT GGT AAA ACT TAT TAC TTT AAG GAT GAT CAC AGC ATA
 440  Q   I   I   D   G   K   T   Y   Y   F   K   D   D   H   S   I

                         1504                1520
                          *                   *
     AAA GCA AAG TCA GAG TAT AGT CAA ATT GGT GGT TCT GTG CCT GAT GAC
 456  K   A   K   S   E   Y   S   Q   I   G   G   S   V   P   D   D

                         1552                1568
                          *                   *
     GGT TTT GCA GAG ATT GAT GGT GAT GGT TAC TTT TTT GAT ACT CAA GGT
 472  G   F   A   E   I   D   G   D   G   Y   F   F   D   T   Q   G

                         1600                1616
                          *                   *
     CAA TTC GTA ACG AAT AGA TTT GTC AGA AAA TAC GAC TAC AGT AAT ATT
 488  Q   F   V   T   N   R   F   V   R   K   Y   D   Y   S   N   I

                         1648                1664
                          *                   *
     TGG TAT TAT TAT GGA AGC GAT GGC AAA CGT GTA TCA GGC TGG CAA ACT
 504  W   Y   Y   Y   G   S   D   G   K   R   V   S   G   W   Q   T

                         1696                1712
                          *                   *
     ATC GAC GGT AAG CGC TAC TAC TTT AGC CAA GAT GAA AAG ACA AAG GGC
 520  I   D   G   K   R   Y   Y   F   S   Q   D   E   K   T   K   G

                         1744                1760
                          *                   *
     CGT CAA ATT AAA GGA CAA ACC ATC ACT ATC GAT GGT AAA GAA TAT ACT
 536  R   Q   I   K   G   Q   T   I   T   I   D   G   K   E   Y   T

                         1792                1808
                          *                   *
     TTT GAC AAA GAC AGC GGT GAA GTT ATC AAT AGT AAC TAG TTG GTA AAT
 552  F   D   K   D   S   G   E   V   I   N   S   N   -

                         1840                1856
                          *                   *
     CCC ATG GCA CAC AAA AAC GAG CAG AAT TCA TAC TCT GTT CGT TTT TTC

                         1888                1904
                          *                   *
     GCC TTA AAA CTT ATA TAT TTA TAA ATC GTC AAT AAA GTG TTT ACT TGA

                         1936                1952
                          *                   *
     AAA CGG TAA ATA TGC CAA GAG TTT GAC TGT TAT CAA TTA ATG GGA AAG
```

FIG. 1. The nucleotide sequence of *gbp* and the deduced amino acid code are shown. The underlined regions upstream of the start site indicate the putative promoter and ribosome-binding site (RBS). The underlined regions downstream of the termination codon, with arrows pointed toward one another, represent potential stem-loop structures. The numbers above the asterisks represent the nucleotide number; the left-hand margin contains the number of the first amino acid in each line of the reading frame. Base number 75 was ambiguous, as it consistently read as a T in one direction but as a C in the opposite direction.

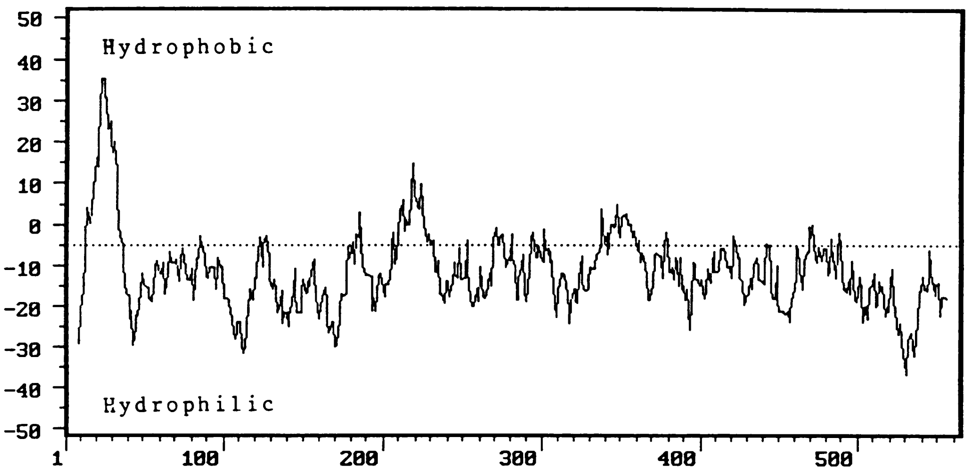


FIG. 2. Hydropathy plot of GBP. The hydrophilic nature of GBP is clearly visible. The hydrophobic core of the signal peptide is evident at the N terminus.

the method of Henikoff (6). Sequencing was performed by using the dideoxy chain-termination method of Sanger et al. (18) and the Klenow fragment of DNA polymerase I in conjunction with the sequencing protocol of Amersham Corp. (Arlington Heights, Ill.) or with a modified T7 DNA polymerase supplied with the Sequenase kit. The sequence was confirmed with overlapping clones, and the entire gene was sequenced in both orientations. Portions of the sequence that were compressed or conflicted between clones were resequenced by using 7-deaza-dGTP in place of dGTP or 25% formamide in the gel. The DNA sequence of *gbp* and its putative amino acid sequence were analyzed by the James M. Pustell DNA and protein sequencing program (International Biotechnologies, Inc., New Haven, Conn.), Staden-Plus (Amersham), and the CLUSTAL program (7).

RESULTS

Cloning of the *gbp* gene in M13. The *gbp* gene was contained on a 4-kb *Eco*RI fragment in pMLG43. This *Eco*RI fragment was subcloned intact into the *Eco*RI site of the bacteriophage M13mp18, and this clone was subsequently used to make sequential deletion clones in one orientation. For sequencing in the opposite direction, the 4-kb fragment containing *gbp* was cleaved and several segments were cloned separately. A Western blot (immunoblot) of a lysate of *E. coli* infected with a recombinant M13 containing the 4-kb *Eco*RI fragment confirmed that GBP was being synthesized and showed a pair of bands as previously reported (16).

Nucleotide sequence. *gbp* was identified as an open reading frame of 1,689 bases preceded by a ribosome-binding site (GGAGG) 8 bases upstream from the ATG start codon (Fig. 1). A putative promoter region preceded *gbp* with −35 (TTGAAA) and −10 (TATCTT) consensus sequences beginning 113 bases upstream from the start site. Potential stem-loop structures were identified near the termination codon of *gbp*.

Amino acid sequence. Analysis of the amino acid sequence of GBP determined from the nucleotide sequence revealed that GBP is a highly hydrophilic protein of 563 amino acids (Fig. 2). The N-terminal portion of the protein corresponds to a signal peptide consisting of a typical basic N terminus

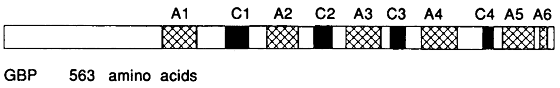
A Repeats:

1.	W	Y	Y	K K A	D G	Q - - L A T G	W	Q I	I	D	G	K Q L Y	F	N - Q D	G	S Q V	K G
2.	W	F	Y	L G E	D G	K - - A A I G	W	R T	I	G	G	K K Y Y	F	D - T N	G	V Q V	K G
3.	W	F	Y	M G A	D G	I - - G V T D	W	Q K	I	D	G	M D Y Y	F	E P S S	G	I Q V	K G
4.	-	Y	Y	F G A	D G	S R K D L T G	W	Q I	I	D	G	K T Y Y	F	K - D D	-	H S I	K A
5.	W	Y	Y	Y G S	D G	K R - - V S G	W	Q T	I	D	G	K R Y Y	F	S - Q D	-	E K T	K G
6.								T	I	D	G	K E Y T	F	D - K D			

C Repeats:

1.	V N	K	I Y Y F	D	P	D	S	G	E L W K D	R F	V
2.	Y G	K	K S F L	D	P	D	T	G	E A W T N	R F	V
3.	D G	K	V Y Y L	D	E	D	S	G	Q V V K N	R F	G
4.	D G	-	- Y F F	D	T	-	Q	G	Q F V T N	R F	V

FIG. 3. Amino acid sequences of the GBP repeat regions. The amino acid sequences of each A repeat and each C repeat are compared among themselves. The boxed residues represent amino acids that were conserved throughout each repeat.



A REPEATS ▨	AMINO ACID LOCATION	TOTAL LENGTH
A 1	169 - 201	33
A 2	264 - 296	33
A 3	349 - 382	34
A 4	425 - 457	33
A 5	504 - 535	32
A 6	544 - 555	12

C REPEATS ■	AMINO ACID LOCATION	TOTAL LENGTH
C 1	227 - 246	20
C 2	312 - 331	20
C 3	388 - 407	20
C 4	479 - 495	17

FIG. 4. Relative positions of the GBP repeat regions. The sizes (amino acid residues) and positions of the A and C repeats within the GBP molecule are illustrated. The precise amino acid location of each repeat is also tabulated.

followed by a hydrophobic central region and a polar C terminus. Cleavage of the signal peptide is believed to be after amino acid 35, in accordance with the −3, −1 rule of von Heijne (24). The molecular weight of the unprocessed protein was 62,909; that of the processed protein was 59,039.

Repeat regions. Two series of repeats were identified within the GBP and were designated A and C repeats. The A repeats were represented by five regions of 32 to 34 amino acids, with a partial sixth repeat of 12 amino acids at the C terminus of the protein. The C repeats ranged from 17 to 20 amino acids and occurred four times throughout the protein sequence. Figure 3 gives the amino acid sequence of each repeat, and Fig. 4 diagrams their lengths and locations within the protein chain. Comparisons of the individual A and C repeats with consensus A and C repeat sequences, respectively, indicated that the repeats contained 48 to 78% identity. The statistical significance of the comparisons was judged by a Monte Carlo shuffle analysis for sequence similarity (10). In this analysis sequence comparisons may be statistically significant, of probable significance, of possible significance, or not significant. Comparisons of A1 through A5 with the consensus sequence were statistically significant; the comparison with A6, the partial repeat, was of possible significance. Comparisons of C1 through C3 with the consensus sequence were statistically significant; the comparison with C4 was of possible significance. A comparison of the repeats at the nucleic acid level indicated that 36 to 69% of the bases were matched when one A repeat was compared with another A repeat, 29 to 65% of the bases were matched when one C repeat was compared with another C repeat, but only 17 to 42% of the bases matched when the A1 repeat was compared with any of the C repeats.

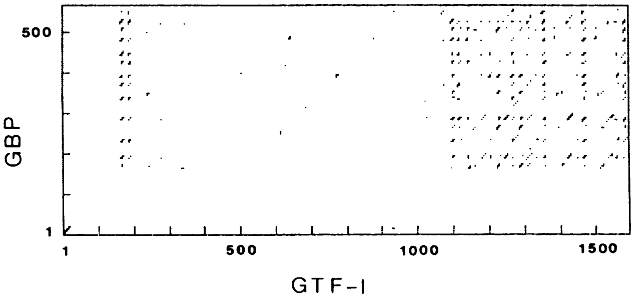


FIG. 5. DIALON plot comparing the amino acid sequences of GBP from *S. mutans* and GTF-I from *S. downei*. The right side of the plot shows the similarity in A and C repeats between the two proteins. The left side of the plot shows the 11-amino-acid sequence of GTF-I, which is repeated throughout GBP.

Homology studies. The deduced amino acid sequence of the GBP was compared with published sequences for GTFs and FTFs from mutans group streptococci. Regions of homology were observed between GBP and the gene products of *gtfI* of *Streptococcus downei* (3, 25) and *gtfB* and *gtfC* of *S. mutans* (20, 23) but not with the product of *ftf* of *S. mutans* (19). The homologies with the products of *gtfI*, *gtfB*, and *gtfC* were all of a similar pattern. Portions of the central and C-terminal regions of GBP were homologous to the C-terminal portions of the GTFs and to a small portion near the N termini (Fig. 5).

The basis of the homology to the C-terminal ends of the GTFs appeared to be the A and C repeats of GBP. Figure 6A and B show the similarity between the GBP A and C repeats, respectively, and repeated regions in the protein products encoded by *gtfI*, *gtfB*, and *gtfC*. The percent identity of the segments in the GTFs to the consensus sequence of GBP A repeats (WYYKGADGKRVTGWQTIDGKQYYFDQDGS QVKG) ranged from 38 to 58% (five were statistically significant, nine were probably significant, and one was possibly significant, as determined by a Monte Carlo analysis for sequence similarity); the percent identity between a GBP C repeat consensus sequence (DGKIYFFDPDSGEV VKNRFV) and regions within the GTFs ranged from 40 to 60% (five were significant, five were probably significant, and one was possibly significant by a Monte Carlo analysis). No portions of the GBP amino acid sequence appeared to be homologous to the B repeat regions previously observed in GTF-I.

Portions of the A repeats in GBP were also found to have homology with a consensus sequence repeated within two pneumococcal autolysins, an amidase encoded by the *lytA* gene of *Streptococcus pneumoniae* and a muramidase encoded by the *cpl* gene of the bacteriophage Cp-1 (5). The first 16 amino acids of the autolysin consensus sequence (GWVKIGDGWYYFDNSGAMATN) contained 10 amino acids identical to an internal 16-amino acid-portion of the GBP A repeat consensus sequence.

The short region of homology seen near the N termini of GBP and each of the GTFs had the sequence DGKWYYK KADG, beginning at residue 166 in GBP. This short sequence possessed weak homology with the second half of the A repeats and recurred 15 times. In each of the GTFs at least 6 of the amino acids were identical to those in GBP, with the G---Y sequence being conserved in all cases. In GTF-I (encoded by *gtfI*), the homologous fragment began at position 164, in GTF-I (encoded by *gtfB*) it began at position 166, and in GTF-SI (encoded by *gtfC*) it began at position 191.

Functional analysis. The region of GTF-I containing A and B repeats had been hypothesized to possibly be involved in glucan binding (3). Mooser and Wong (12) have also identified a domain of GTF-S from *Streptococcus sobrinus* that binds glucan. The amino acid compositions of the repeat regions of the GBP (amino acids 170 to 563) were compared

A

Protein & Location	Amino Acid Sequence
GBP 169	W Y Y K K A D G Q - - L A T G W Q I I D G K Q - L Y F N - Q D G S Q V K G
GBP 264	W F Y L G E D G K - - A A I G W R T I G G K K - Y Y F D - T N G V Q V K G
GBP 349	W F Y M G A D G I - - G V T D W Q K I D G M D - Y Y F E P S S G I Q V K G
GBP 425	- Y Y F G A D G S R K D L T G W Q I I D G K T - Y Y F K - D D - H S I K A
GBP 504	W Y Y Y G S D G K R - - V S G W Q T I D G K R - Y Y F S - Q D - E K T K G
GBP 544	T I D G K E - Y T F D - K D
GTF-I 1100	- Y Y F G Q D G Y - - M V T G A Q N I K G S N - Y Y F L - A N G A A L R N
GTF-I 1163	W R Y F K N - G V - - M A L G L T T D G H V - Q Y F D - K D G V Q A K D
GTF-I 1227	W Y Y L G K D G V - - A V T G A Q T - G K Q H L Y F E - A N G Q Q V K G
GTF-I 1292	W F Y L G K D G A - - A V T G A Q T I K G Q K - L Y F K - A N G Q Q V K G
GTF-I 1406	W V Y V - K S G K - - V L T G A Q T I - G N Q R V Y F K - D N G H Q V K G
GTF-I 1519	W L Y V - K D G K - - V L T G L Q T V - G N Q K V Y F D - K N G I Q A K G
GTF-B 1096	W Y Y F D N N G Y - - M V T G A Q S I N G V N - - Y F L - S N G L Q L R D
GTF-B 1160	W R H F - N N G E - - M S V G L T V I D G Q V - Q Y F D - E M G Y Q A K G
GTF-B 1224	W L Y L G E D G A - - A V T G S Q T I N G Q H - L Y F R - A N G V Q V K G
GTF-B 1289	W F Y F D N N G Y - - A V T G A R T I N G Q L - L Y F R - A N G V Q V K G
GTF-B 1354	W F Y F D N N G Y - - A V T G A R T I N G Q H - L Y F R - A N G V Q V K G
GTF-B 1419	W F Y F D N N G Y - - A V T G A R T I N G Q H - L Y F R - A N G V Q V K G
GTF-C 1126	W Y Y F D N N G Y - - M V T G A Q S I N G A N - Y Y F L - S N G I Q L R N
GTF-C 1189	W R Y F G N - G I - - M A V G L T R V H G A V - Q Y F D - A S G F Q A K G
GTF-C 1253	W F L F D H N G V - - A V T G T V T F H G Q R - L Y F K - P N G V Q A K G

B

Protein & Location	Amino Acid Sequence
GBP 227	V N K I Y Y F D P D S G E L W K D R F V
GBP 312	Y G K K S F L D P D T G E A W T N R F V
GBP 388	D G K V Y Y L D E D S G Q V V K N R F G
GBP 479	D G - - Y F F D T - Q G Q F V T N R F V
GTF-I 1201	D G K V R Y F D Q H N G N A V T N T F V
GTF-I 1265	D G K L Y F Y D V D S G D M W T N T F I
GTF-I 1330	D G K I R Y Y D A Q T G E Q V F N K S V
GTF-I 1444	D G K L R Y Y D A N S G D Q A F N K S V
GTF-I 1557	D G K V R Y F D E N S G S M I T N Q W K
GTF-B 1198	D G K I R Y F D K Q S G N M Y R N R F I
GTF-B 1263	H G R I S Y Y D G N S G D Q I R N R F V
GTF-B 1328	Y G R I S Y Y D G N S G D Q I R N R F V
GTF-B 1393	H G R I S Y Y D G N S G D Q I R N R F V
GTF-C 1227	D G K L R Y F D R D S G N Q I S N R F V
GTF-C 1292	N G Y L R Y Y D P N S G N Q V R N R F V

FIG. 6. The amino acid compositions of the A (A) and C (B) repeats in GBP and GTFs. The amino acid sequences of the A and C repeats in the GBP and similar sequences in GTFs are compared. GTF-I is the *gtfI* gene product from *S. downei* MFe28, GTF-B designates the GTF-I product from *gtfB* of *S. mutans* GS-5, and GTF-C designates the GTF-SI product of *gtfC* in *S. mutans* GS-5. The numbers in the left-hand margin refer to the first amino acid of the repeat. The boxed residues correspond to amino acids which are conserved through the majority of the repeats.

with the amino acid compositions of the glucan-binding domain, as determined by Mooser and Wong (12), and with the repeat region of GTF-I (3). Figure 7 shows the similarity in the amino acid compositions of these three sources.

DISCUSSION

A suggested role for the GBP in the virulence of *S. mutans* is that it may contribute to cohesive plaque formation (17). Proteins with similar properties have been reported in *S. sobrinus* (9) and *Streptococcus cricetus* (2) and are also thought to contribute to adherence and accumulation of the organisms in plaque.

The molecular weight of the processed GBP, as determined from the deduced amino acid sequence and the hypothesized signal peptide cleavage site, is 59,039. The sequence contains a signal peptide and is highly hydrophilic, consistent with its extracellular location. There is no indication that GBP is linked to the cell wall, as its structure is quite different from that of the wall-associated protein (4). The size of the protein is smaller than that determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis, 74,000 (15), but the reading frame is well defined, containing a putative promoter region upstream from the initiation codon and potential stem-loop structures near the termination codon. There are tandem methionine residues at the start of the GBP, and it appears that the protein sequence begins at the second methionine, based on the spacing of the ribosome-binding site 8 bases upstream. The source of the FTF activity associated with GBP prepared from *S. mutans* remains an enigma. The sequence data presented in this report confirm previous observations that *gbp* and *ftf* genes are unrelated (1). The possibility exists that the two proteins interact and comigrate in sodium dodecyl sulfate-polyacrylamide gels, accounting for the altered mobility and higher molecular weights observed previously. Specific inactivation of each of the genes may help to resolve the problem.

The most striking feature of the amino acid sequence was the presence of two sets of repeats that spanned three-quarters of the length of the protein. The A repeats were similar to the A repeats identified in GTF-I of *S. downei* (3) (this strain was previously identified as *S. sobrinus*), and the GBP C repeats shared partial identity to regions between the A repeats. Both the A and C repeats of GBP overlapped reiterated sequences identified by Kuramitsu and colleagues

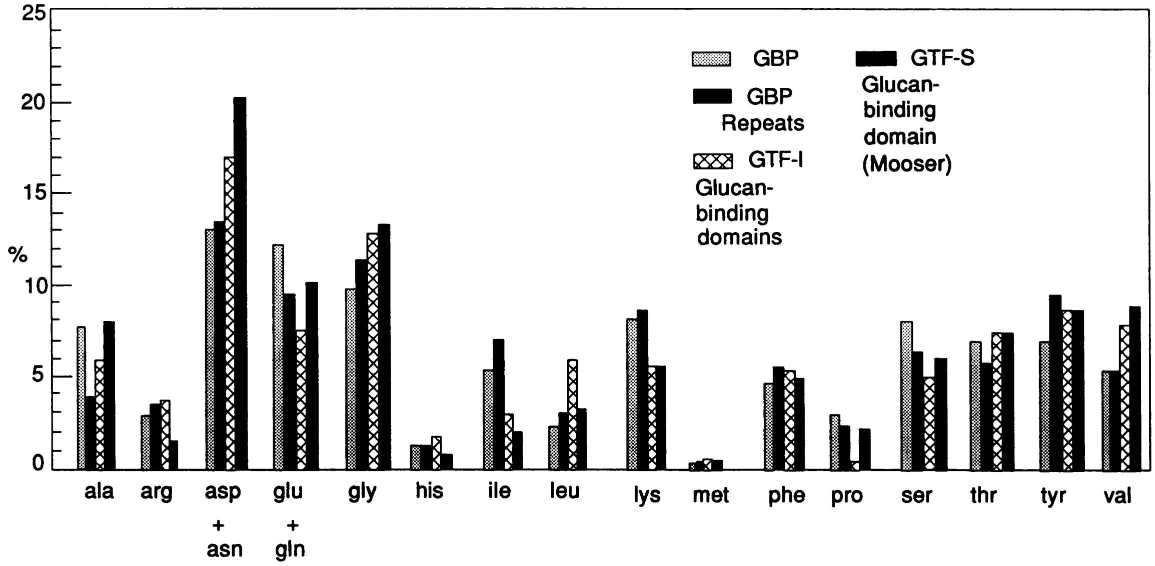


FIG. 7. Amino acid compositions of GBP and glucan-binding domains. The histogram illustrates the similarity in amino acid composition between the proposed glucan-binding domains in GBP, GTF-I, and GTF-S.

within the proteins encoded by *gtfB* (20) and *gtfC* (23), although the regions homologous to the GBP C repeats were consistently located between regions similar to the GBP A repeats. The position of the C repeats in relation to the A repeats differed from protein to protein and within a protein, leading us to designate the C repeats as separate rather than regarding both repeats as a single long repeat. The GBP did not have any regions homologous to the B repeats of GTF-I.

The significance, if any, of the 11-amino-acid sequence from GTF-I that is repeated 15 times in GBP with varying degrees of identity is uncertain; however, the beginning of the 11-amino-acid repeats coincides with the beginning of the A and C repeat regions. The significance of the A and C repeating units in the GBP and GTFs is unknown, although it is believed that they may function in glucan binding (3). In this regard the amino acid compositions of the proposed glucan-binding domains of GBP and GTF-I and the glucan-binding peptide of GTF-S were compared and found to be similar (Fig. 7). Portions of the GBP A repeats were also similar to a repeat region within two pneumococcal autolysins. The repeats in the autolysins were within the carboxy-terminal portion of each protein, and there is evidence that the carboxy-terminal half of each protein functions in substrate recognition (5). It would appear that the repeat domains have been duplicated and preserved among several genes in different species during evolution, indicating some importance in structure, function, or both.

When a consensus amino acid sequence of the A repeats of the GBP was compared with similar sequences in the GTFs, certain amino acids were found to be conserved throughout. These were most often glycine, but a tyrosine-phenylalanine pair was also conserved. Other amino acids that were conserved in almost all instances included threonine, tyrosine, asparagine, glutamine, and lysine. All of these amino acids, with the exception of phenylalanine, have been reported to be involved in hydrogen bonding in protein-saccharide complexes (14). Another common feature of carbohydrate-binding domains is a central beta-pleated sheet region bounded by helices (14). Analysis of the secondary structure of the GBP revealed that helical regions were present between A repeats but usually not within. While it is attractive to propose that the repeating units found in GBP and GTFs form a glucan-binding domain that is involved in binding to glucans (which also have repeating structures), the way in which this might occur is not yet apparent. In a study of deletion mutants of GTF-I, the smallest fragment observed to be retained by a glucan affinity column had a molecular weight of 65,000, although indirect evidence suggested that a peptide less than half that size would still have binding activity (3). The glucan-binding peptide found by Mooser and Wong (12) has a similar molecular weight, 60,500. On the other hand, Landale and McCabe (9) reported that a GBP with a molecular weight as small as 7,500 could still interact with glucan. It will be interesting to determine whether the number of repeat units correlates with the strength or specificity of the binding reaction. The way in which extracellular GBPs contribute to sucrose- or dextran-induced agglutination is not known. They do not fit the general concept of cell surface receptors by which bacteria bind to exogenous macromolecules, although GBP rapidly becomes bound to the surface of *S. mutans* on exposure to sucrose (1).

The sequencing of the *gbp* gene has laid the foundation on which further experimentation can build an understanding of the significance of the repeating regions in the GBP and GTFs and ultimately of the role of the GBP in *S. mutans* virulence.

ACKNOWLEDGMENTS

We thank David R. Lorenz for helpful suggestions and expertise. This research was supported by Public Health Service grant DE08191 from the National Institutes of Health. J.A.B. was supported by National Research Service award DE05545 from the National Institute of Dental Research.

LITERATURE CITED

1. **Aduse-Opoku, J., M. L. Gilpin, and R. R. B. Russell.** 1989. Genetic and antigenic comparison of *Streptococcus mutans* fructosyltransferase and glucan-binding protein. FEMS Microbiol. Lett. **59:**279–282.

1a. **Douglas, C. W. I., and R. R. B. Russell.** 1982. Effect of specific antisera on adherence properties of the oral bacterium *Streptococcus mutans*. Arch. Oral Biol. **27**:1039–1045.

2. **Drake, D., K. G. Taylor, A. S. Bleiweis, and R. J. Doyle.** 1988. Specificity of the glucan-binding lectin of *Streptococcus cricetus*. Infect. Immun. **56**:1864–1872.

3. **Ferretti, J. J., M. L. Gilpin, and R. R. B. Russell.** 1987. Nucleotide sequence of a glucosyltransferase gene from *Streptococcus sobrinus* MFe28. J. Bacteriol. **169**:4271–4278.

4. **Ferretti, J. J., R. R. B. Russell, and M. L. Dao.** 1989. Sequence analysis of the wall-associated protein precursor of *Streptococcus mutans* antigen A. Mol. Microbiol. **3**:469–478.

5. **Garcia, E., J. L. Garcia, P. Garcia, A. Arraras, J. M. Sanchez-Puelles, and R. Lopez.** 1986. Molecular evolution of lytic enzymes of *Streptococcus pneumoniae* and its bacteriophages. Proc. Natl. Acad. Sci. USA **85**:914–918.

6. **Henikoff, S.** 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28**: 351–359.

7. **Higgins, D. G., and P. M. Sharp.** 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73**:237–244.

8. **Kuehn, S., H. J. Fritz, and P. Starlinger.** 1979. Close vicinity of IS1 integration sites in the leader sequence of the *gal* operon of *Escherichia coli*. Mol. Gen. Genet. **167**:235–241.

9. **Landale, E. C., and M. M. McCabe.** 1987. Characterization by affinity electrophoresis of an α-1,6-glucan-binding protein from *Streptococcus sobrinus*. Infect. Immun. **55**:3011–3016.

10. **Lipman, P. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227**:1435–1441.

11. **Loesche, W. J.** 1986. Role of *Streptococcus mutans* in human dental decay. Microbiol. Rev. **50**:353–380.

12. **Mooser, G., and C. Wong.** 1988. Isolation of a glucan-binding domain of glucosyltransferase (1,6-α-glucan synthase) from *Streptococcus sobrinus*. Infect. Immun. **56**:880–884.

13. **Muller-Hill, B., L. Crapo, and W. Gilbert.** 1968. Mutants that make more *lac* repressor. Proc. Natl. Acad. Sci. USA **59**: 1259–1264.

14. **Quiocho, F. A.** 1986. Carbohydrate-binding proteins: tertiary structures and protein-sugar interactions. Annu. Rev. Biochem. **55**:287–315.

15. **Russell, R. R. B.** 1979. Glucan-binding proteins of *Streptococcus mutans* serotype c. J. Gen. Microbiol. **112**:197–201.

16. **Russell, R. R. B., D. Coleman, and G. Dougan.** 1985. Expression of a gene for glucan-binding protein from *Streptococcus mutans* in *Escherichia coli*. J. Gen. Microbiol. **131**:295–299.

17. **Russell, R. R. B., A. C. Donald, and C. W. I. Douglas.** 1983. Fructosyltransferase activity of a glucan-binding protein from *Streptococcus mutans*. J. Gen. Microbiol. **129**:3243–3250.

18. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.

19. **Shiroza, T., and H. K. Kuramitsu.** 1988. Sequence analysis of the *Streptococcus mutans* fructosyltranferase gene and flanking regions. J. Bacteriol. **170**:810–816.

20. **Shiroza, T., S. Ueda, and H. K. Kuramitsu.** 1987. Sequence analysis of the *gtfB* gene from *Streptococcus mutans*. J. Bacteriol. **169**:4263–4270.

21. **Stoker, N. G., N. F. Fairweather, and B. G. Spratt.** 1982. Versatile low-copy-number plasmid vectors for cloning in *Escherichia coli*. Gene **18**:335–341.

22. **Tartof, K. D., and C. A. Hobbs.** 1987. Improved media for growing plasmid and cosmid clones. Focus **9**:12.

23. **Ueda, S., T. Shiroza, and H. K. Kuramitsu.** 1988. Sequence analysis of the *gtfC* gene from *Streptococcus mutans* GS-5. Gene **69**:101–109.

24. **von Heijne, G.** 1983. Patterns of amino acids near signal sequence cleavage sites. Eur. J. Biochem. **133**:17–21.

25. **Whiley, R. A., R. R. B. Russell, J. M. Hardie, and D. B. Beighton.** 1988. *Streptococcus downei* sp. nov. for strains previously described as *Streptococcus mutans* serotype h. Int. J. Syst. Bacteriol. **38**:25–29.

26. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of M13mp18 and pUC19 vectors. Gene **33**:103–119.