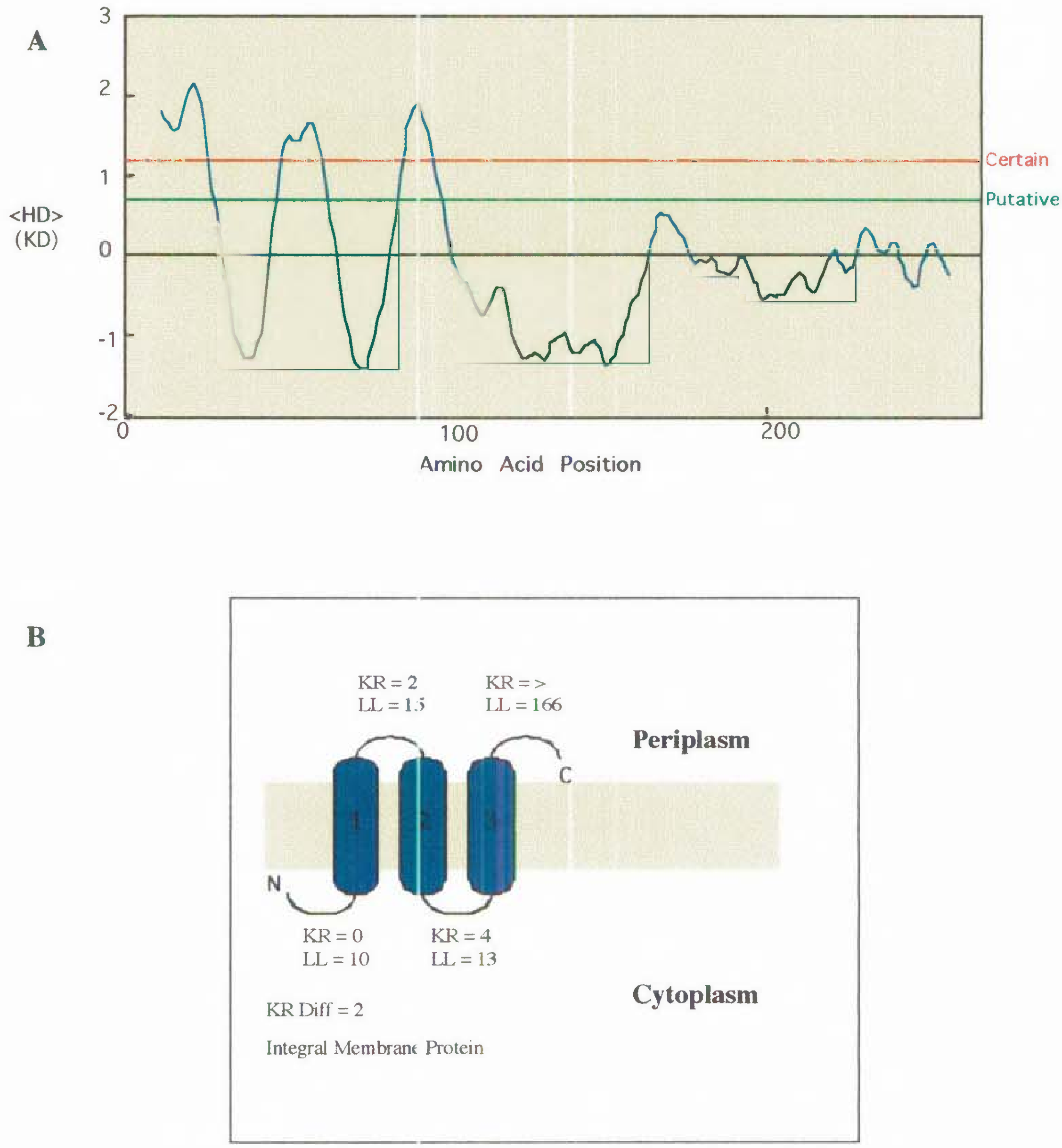


Figure 5.9: Hydrophobicity (<HD>) plot (**A**) and predicted topology (**B**) of the GepC protein using the computer program TopPred II (Claros, 1996). Hydrophilicity is calculated according to the algorithm of Kyte-Doolittle (KD) (Kyte & Doolittle, 1982). Transmembrane domains are numbered 1 to 3. Other features are as follows: LL indicates the number of amino acids in the loop; KR indicates the number of lysine and arginine residues respectively; KR Diff indicates the positive charge difference; N- and C- indicate the amino and carboxy termini of the protein; > indicates too long to be significant. Cytoplasmic and periplasmic faces are also indicated.

Thus, it is possible that GepC may also comprise part of the transporter complex, and may form a heterodimer with the GepE putative membrane-associated protein of the BPD ABC transporter protein.

In systems that consist of two membrane-associated protein components, the proteins are often similar to each other, and are thought to have arisen by gene duplication (Boos & Lucht, 1996). A ClustalW alignment of GepC and GepE (Figure 5.10) shows that GepC has 37% similarity and 17.3% identity to the GepE protein over 249 aa, and therefore the two putative membrane-associated proteins do share similarity. However, many of the conserved proline and glycine residues previously identified in GepE and even other GepE-related proteins (Figure 5.8) are not present at the same positions in GepC, nor was an EAA-loop identified in GepC.

```
                 v         v         v         v         v         v
GepC     MPIDVFCFAKECRVKAGARLAFQIDPKYNTIFDAQIVGFTYSFNSEQEIDRFDDDLPSDF
GepE     ----MIRFAAQFLLTAFGLCFLWQSVIWLTQTPPYILPSPLAVLQ---------CLYTDF
             .  ** .  . *             *     *.    .             * .**

GepC     SLIISAIFAAVFFAMITFLYPPSILILLGTAS-VIMFIQLLYERWSWERNQSPLCSFNPP
GepE     DVLWRHSRVTVLEIALSLG----IASVFGTATALILMINARLRRWLQP---LILVSQTMP
          ..      .*    ..      *  . ***. .*. *.    **        * * . *

GepC     VFLILSLLGAWSGALLGQYMFNYQRRQPRFKYLLWLVSGINFSVLFLLGINFDMRPPEPE
GepE     VYALAPILMLWFG--------------------------YGLTPKIIVTVLIAYFPITTA
         *. .  .*  * *                           . .   .. .     *

GepC     PVAIVQTHQPSSLTEQNTAPTDTAPQKPVTNTVSNQTEPSSTDISQNSSPTATAPDISLP
GepE     TFDGLQQTPPAYLRLAQTLGAN--PRQ----ILWRIRMP------------AALPHLASG
             .*   *. *   .*  .   *..     .     *            *. * ..
                                EAARALGASPWRIFFK------------IILPLA

GepC     AVSFFETPKVTPPESCYIVVQDFRDMEAAKTFAAEQALNNPDAPIRIFFTQKYKFAVTNG
GepE     LRVGAAMAPIG------AIIGEYVGGSDGLGYLMQYGINRS----QVALTFAALFVMTLL
                  .        .. ..        .  .  .*      ..  *    * .*

GepC     TLEISSAAAQLDQKIAAGELPPESYCLLKAGVREEVAR
GepE     TLAIYYGIDALFEKMVLLGNNGE---------------
         ** *      * .*.       *
```

Figure 5.10: ClustalW alignment of GepC and GepE proteins. Amino acids that are identical in both sequences are indicated by asterisks; conserved amino acid residues are indicated by a period; 10 aa intervals are shown (v). The conserved glycine and proline residues are shown in red. The EAA-loop is shown in red. GepC has 37% similarity and 17.3% identity respectively to GepD.

Hence, it seems unlikely, though not impossible that the GepE and GepC proteins are involved in the formation of a heterodimeric membrane-associated protein complex. It is therefore equally possible that the GepE protein forms a homodimer. It is also notable that, in most cases, those BPD systems that do share similarity to the Gep proteins similarly have only one gene encoding the membrane-associated protein (Figure 5.11), so perhaps this specific subgroup of transporters forms a homodimeric membrane-associated protein complex.

Hydrophobicity plots of GepD, GepF, GepG proteins, and topology predictions subsequently made, indicated that the GepD, GepF and GepG proteins are not membrane proteins (data not shown).

5.2.5 A bacterial BPD-ABC importer or exporter?

In addition to having an ATP-binding domain and a membrane-associated protein, bacterial ABC importers have a periplasmic binding protein that binds the substrate, and then presents it to the import complex (membrane-associated protein or permease) in the inner membrane. In addition, the ATP-binding domain and the membrane-associated protein are encoded by different polypeptides. In contrast, the bacterial ABC exporters do not have a periplasmic substrate-binding protein, and the ATP-binding domains and the membrane-associated protein are encoded by the same polypeptide in almost all cases (Boos & Lucht, 1996; Fath & Kolter, 1993). Based on nucleic acid homologies, GepD comprises the ATP-binding domain and GepE comprises the membrane-associated protein. Hence, GepD and GepE are most likely to comprise a binding protein-dependent ABC importer complex.

The soluble periplasmic substrate-binding protein has high affinity for the substrate to be transported and for the membrane-associated protein dimer which interacts with the

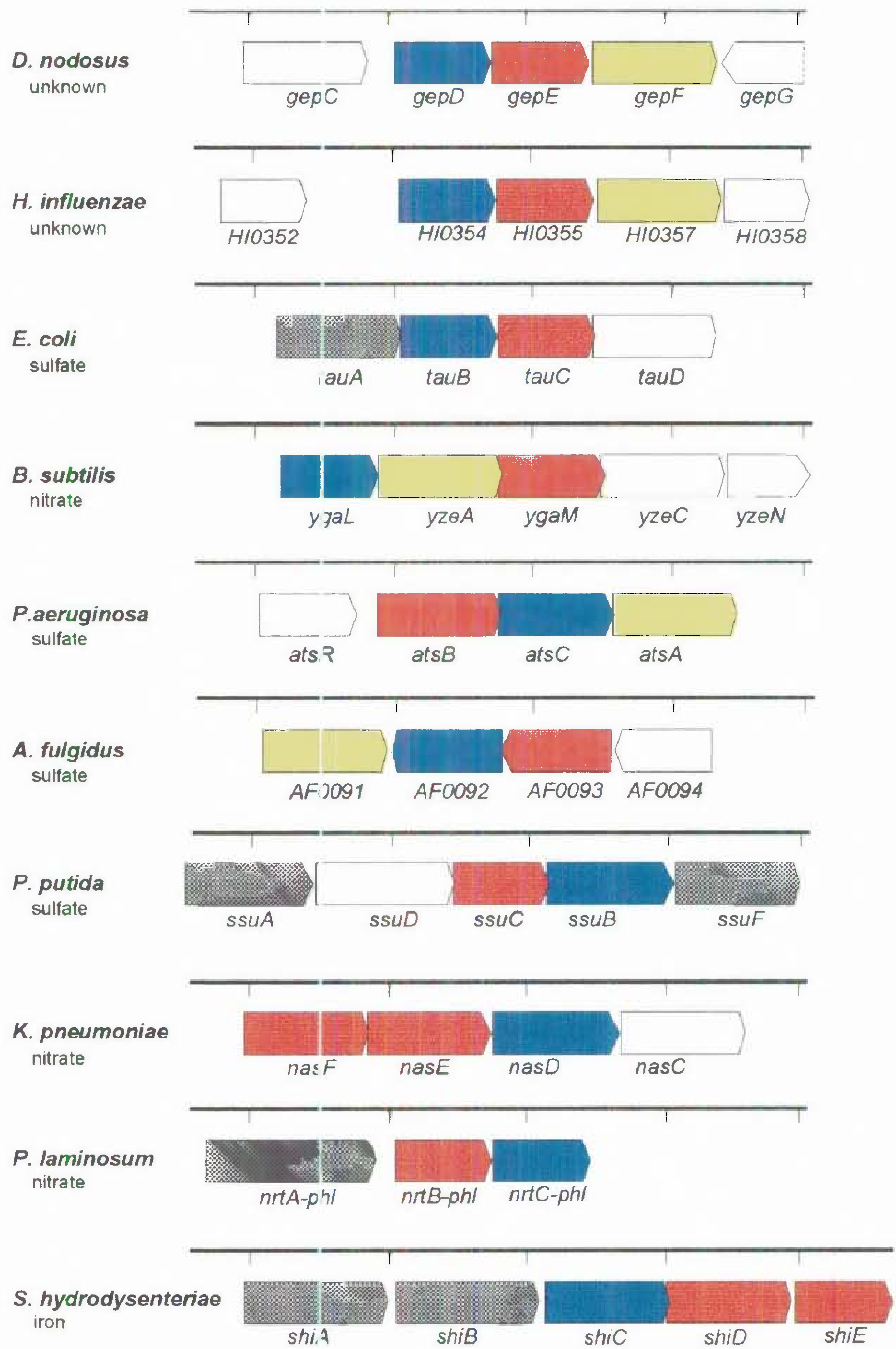


Figure 5.11: Organisation of selected bacterial ABC importer operons with similarity to the putative ABC importer operon present downstream of *intB* in the *D. nodosus* strain A198. The orfs are colour-coded according to which component of a BPD ABC transport system they encode as follows: ATPase containing an ATP-binding cassette (blue); membrane-associated protein (red); periplasmic binding proteins (grey); proteins related to GepF (yellow); genes encoding unrelated proteins are unshaded. The organism of origin and the substrate that is putatively imported is indicated at the left of the corresponding orfs. The marks on the scale lines are 1 kb apart.

both the periplasmic substrate-binding protein and the peripherally membrane-associated ATP-binding domains which are located on the cytoplasmic side of the inner membrane (Boos & Lucht, 1996).

The polypeptides that comprise the BPD ABC transporter systems need to be produced in stoichiometric amounts (Boos & Lucht, 1996) and consequently the structural genes encoding the components of the BPD transporters are almost invariably organised as operons in an effort to co-ordinate their expression. Sequences matching the –35 and –10 *E. coli* promoter consensus sequence were identified upstream of *gepC* and upstream of *gepDEF* respectively (Figure 5.4). Thus, *gepC* appears to be transcribed separately from *gepD* and *gepE* genes which are organised in an operon like arrangement and are thus likely to be expressed co-operatively.

In the majority of systems analysed to date, the periplasmic binding-protein is encoded by the first gene in the operon (Boos & Lucht, 1996). Although *gepD* and *gepE* show most similarity to BPD ABC importers, no gene encoding a putative periplasmic binding protein has been identified adjacent to *gepD* and *gepE*. It has been noted that despite the structural resemblance as determined from X-ray crystallographic studies, the sequence homology between periplasmic binding proteins is limited (Boos & Lucht, 1996), so it is possible that the *D. nodosus* periplasmic binding-protein is not related to other periplasmic binding proteins.

In addition, there are examples of systems where the periplasmic binding proteins are encoded in a separate transcriptional unit. In *E. coli*, the *fepB* gene which encodes a protein which binds enterobactin-Fe complexes and is located upstream from *fepDGC* which are transcribed from a different promoter (Elkins & Earhardt, 1989; Shea & McIntosh, 1991). For some systems, two different binding proteins are able to interact with the same membrane-associated protein, and in such systems the alternative binding proteins are not encoded by the operon but may be transcribed independently from genes located elsewhere in the genome. Examples include *(i)* the *S. typhimurium hisJ* (histidine) binding-protein and *E. coli argT* (arginine, lysine, ornithine) binding-proteins (Higgins & Ames, 1981); *(ii)* proteins encoded by *cysP* and *sbp* of *E. coli* that are periplasmic binding proteins of thiosulfate and sulfate respectively (Hryniewicz *et al.*, 1990); and *(iii) E. coli livK* and *livJ* encoding leucine and threonine binding proteins (Landick & Oxender, 1985). Thus, it may be that the periplasmic protein/s that interacts with the GepD and GepE transporter complex is located elsewhere in the *D. nodosus* chromosome.

5.2.6 Identification of a sulfate uptake system?

BPD transport systems are important in the uptake of a variety of substances that are found at low concentrations in the environment. They are not expressed constitutively but instead are expressed when substrates for the transporters are present in the growth medium (Boos & Lucht, 1996; Fath & Kolter, 1993). Since most of the transporters that are closely related to the *gepDEF* system import sulfate or nitrate, it is probable that the *gepDEF* system may also transport sulfate or nitrate. In order to determine whether GepD and GepE do encode a putative transport protein that is involved in the transport of sulfur or nitrate, Northern blot analyses could be used to determine whether the expression of *gepD* and *gepE* increases in the presence of the putative substrates.

Anaerobic bacteria carry out a variety of redox reactions involving organic compounds, CO_2, molecular hydrogen and sulfur compounds, which are coupled to substrate-level phosphorylation. The reduction of CO_2 and of sulfate is carried out by strict anaerobes whereas nitrate reduction is carried out by aerobes when O_2 is not present (Gottshalk, 1986). After carbon, nitrogen is the next most abundant element in the cell (12-15% dry weight). The bulk of the nitrogen available in nature is in the inorganic form, either as ammonia (NH_3), nitrogen gas (N_2) or nitrate (NO_3^-). Although most bacteria are able to utilise ammonia as the sole nitrogen source, only some can utilise nitrate, and nitrogen gas (N_2) is only useful for nitrogen-fixing bacteria.

D. nodosus is a strict anaerobe and as such, *D. nodosus* is not likely to use nitrate for anaerobic respiration, since nitrate usually acts as an electron acceptor during anaerobic respiration only in facultative aerobes when O_2 is not present (Gottshalk, 1986). Although some bacteria use nitrate in assimilatory pathways, the first step in all cases involves the reduction of nitrate to nitrite. It has previously been reported that *D. nodosus* cannot reduce nitrate to nitrite (Beveridge, 1941). It is therefore unlikely that the putative *gepDE* encoded BPD importer is a nitrate transporter, assuming that the transporter is functional.

In contrast, a wide variety of microorganisms can utilise sulfate as a sulfur source, which is reduced to sulfide and then used for biosynthetic purposes or anaerobic respirations (Gottshalk, 1986). It is likely that *D. nodosus* uses sulfur in anaerobic respiration since it is a strict anaerobe which does produce hydrogen sulfide (Beveridge, 1941). In addition, sulfur is required for the biosynthesis of the amino acids cysteine and methionine, and a number of coenzymes including coenzyme A, biotin, α-lipoic acid (Brock & Madigan, 1991) and thiamine pyrophosphate (TPP) (Gottshalk, 1986), the prosthetic group of decarboxylases and transketolases. Hence it is most probable that *gepD* and *gepE* are involved in the import of sulfate.

5.2.7 *gepF* and thiamine biosynthesis

In the alignment of bacterial importer operons that share most similarity with the proteins encoded by *gepD*, *gepE* and *gepF* (Figure 5.11), periplasmic substrate-binding proteins are only clustered with the putative ATP-binding and membrane-associated subunit genes where a *gepF* homologue is not present. This suggests that the GepF protein may be functionally associated with the ABC importer functions.

GepF is located only 6 nt downstream of *gepE*. Similarly, in homologous BPD ABC import systems a GepF-like gene is adjacent to genes encoding a putative ABC transporter ATP-binding domain and ABC transporter membrane-associated subunit (Figure 5.11). In addition, since no putative promoter sequence was identified in the region immediately upstream of *gepF* it is likely that *gepF* is transcribed with *gepD* and *gepE*. These observations further support the hypothesis that GepF is linked with ABC transporter associated functions.

GepF is most similar to the putative thiamine biosynthesis protein (HI0357) of *H. influenzae* (Fleischmann *et al.*, 1995), and to several thiamine biosynthesis proteins from yeasts and fungi (Table 5.2). Studies in fission yeast indicate that the expression of the homologous thiamine biosynthesis gene is completely repressed at the transcriptional level in the presence of thiamine (Maundrell, 1990). Thus, in order to determine whether GepF is functionally similar to the *nmt1*-like thiamine biosynthesis genes, one could investigate the expression of *gepF* at the transcriptional level in the presence versus absence of thiamine in the growth medium.

No homologous thiamine biosynthesis protein has been identified in the *E. coli* genome or in the other bacterial genomes that have been completely sequenced to date (with the exception of *H. influenzae*). It has been noted that although thiamine is synthesised by both prokaryotes and eukaryotes, the pathway by which this is achieved is different (White & Spenser, 1996). This might suggest that some prokaryotes like *D. nodosus* and *H. influenzae* have thiamine biosynthesis pathways that are more similar to the eukaryotic pathway than to the biosynthesis pathway utilised by *E. coli*. Unfortunately, the early steps of thiamine biosynthesis, which include the synthesis of the pyrimidine ring and the thiazole subunit (prior to 1 and 3 in Figure 5.12), have not yet been elucidated (White & Spenser, 1996). The only other genes that have known functions that share aa identity with GepF are the extracellular copper response proteins from *M. byrantii*, and the

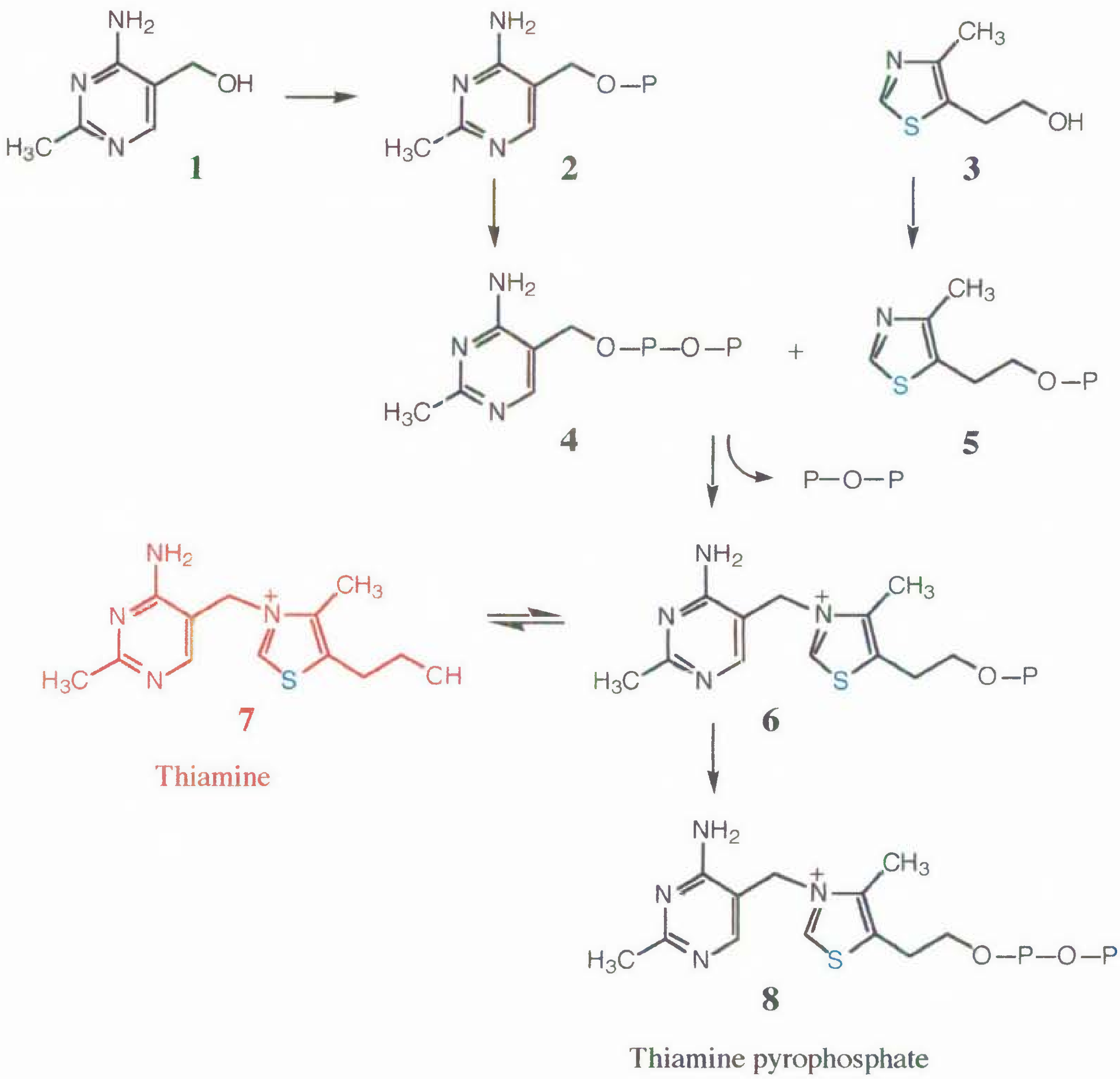


Figure 5.12: Summary of the final stages in the biosynthesis of thiamine (structure 7) and thiamine pyrophosphate (structure 8). Enzymes involved in the synthesis are as follows: hydroxymethylpyrimidine kinase (steps 1 to 2), hydroxymethylthiazole kinase (step 3 to 5), phosphomethylpyrimidine kinase (step 2 to 4), thiamine phosphate kinase (step 6 to 8), thiamine kinase (step 7 to 6), thiamine pyrophosphokinase (step 7 to 8), thiamine phosphate phosphatase (step 6 to 7), thiamine phosphate pyrophosphorylase (steps 4 + 5 to 6) (White & Spenser, 1996).

Synechococcus cyanate periplasmic binding-protein CynA. Although GepF has similarity to these periplasmic binding-proteins, GepF has significantly higher similarity to thiamine biosynthesis proteins (Table 5.2).

The biosynthesis of thiamine requires sulfur (Figure 5.12, prior to step 3) as a component of the thiazole ring within the thiamine molecule (White & Spenser, 1996). Hence, the import of sulfate is likely to be coupled with thiamine biosynthesis, which may explain why *gepD* and *gepE*, thought to be involved in sulfate import, are located next to *gepF*, which is proposed to be required for thiamine biosynthesis.

There is no evidence to suggest that genes encoding sulfate BPD importer proteins are present in multiple copies in bacterial genomes, although it is known that BPD importer proteins for different sulfur sources are present in bacterial genomes. Examples include the *E. coli tauABCD* operon (van der Ploeg *et al.*, 1996) which imports taurine, and *cysATW* which utilises periplasmic binding proteins Sbp and CysP for the import of sulfate and thiosulfate respectively (Hryniewicz *et al.*, 1990; Jacobson *et al.*, 1991; Sirko *et al.*, 1990).

It is interesting that the periplasmic binding protein for sulfate in *E. coli* (Hryniewicz *et al.*, 1990) is not adjacent to the genes encoding the ATP-binding domain and membrane-associated protein, and furthermore, this seems to be a characteristic shared with sulfate transporters that are related to GepD and GepE proteins (Figure 5.11). In addition, this might also suggest that as in *E. coli*, in *D. nodosus* and other microorganisms, sulfate and thiosulfate import may be mediated by the same ATP-binding domain and permease protein, but two separate binding proteins located elsewhere in the genome.

5.2.8 Southern blot analysis of the sequences downstream of the *intB* gene in seventeen strains of *D. nodosus*

In order to investigate the prevalence and arrangement of the *intB* element and adjacent sequences in *D. nodosus*, the genomic DNA from seventeen strains of *D. nodosus* was digested with *Eco*RI, *Hin*dIII and *Eco*RI/*Hin*dIII, and analysed in Southern blot experiments in which seven probes were utilised (Figure 5.13). The strains of *D. nodosus* analysed included virulent strains A198, 1311, B1006, G1220, H1215, D1172, intermediate strain AC3577, and benign strains C3052, 819, 1169, 2483, 1493, 3138, 1469, 1311A, H1204 and AC390. The data generated from these Southern blot experiments has been tabulated in Appendix 5.

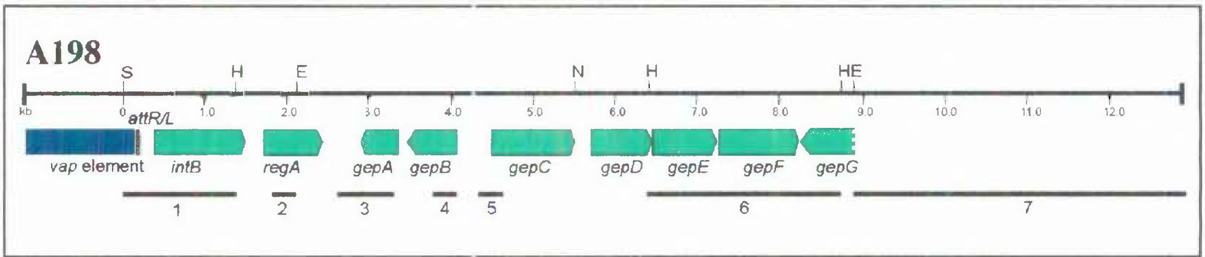


Figure 5.13: Restriction map of sequences downstream of *intB* from *D. nodosus* virulent strain A198. The DNA fragments utilised as probes in Southern blot analyses are shown below the corresponding DNA sequence and numbered 1 to 7. Other features shown include: putative attachment site, *att* (red rectangle) and *vap* element sequences is located at the left of *intB* gene. Restriction sites shown include *Sac*I (S), *Hin*dIII (H), *Eco*RI (E) and *Nru*I (N). The numbers indicate the distance in kb.

D. nodosus strain A198 contains two copies of the *intB* gene. One of these copies of *intB* hybridises strongly to the *intB* probe (Figure 5.13, probe 1), and corresponds to a whole copy of the *intB* gene which is adjacent to the genes *regA*, and *gep* genes *A-G*. The second copy of *intB* that is present in strain A198 corresponds to a partial copy of the *intB* gene, called $intB_N$, which is not adjacent to *regA* and the *gep* genes.

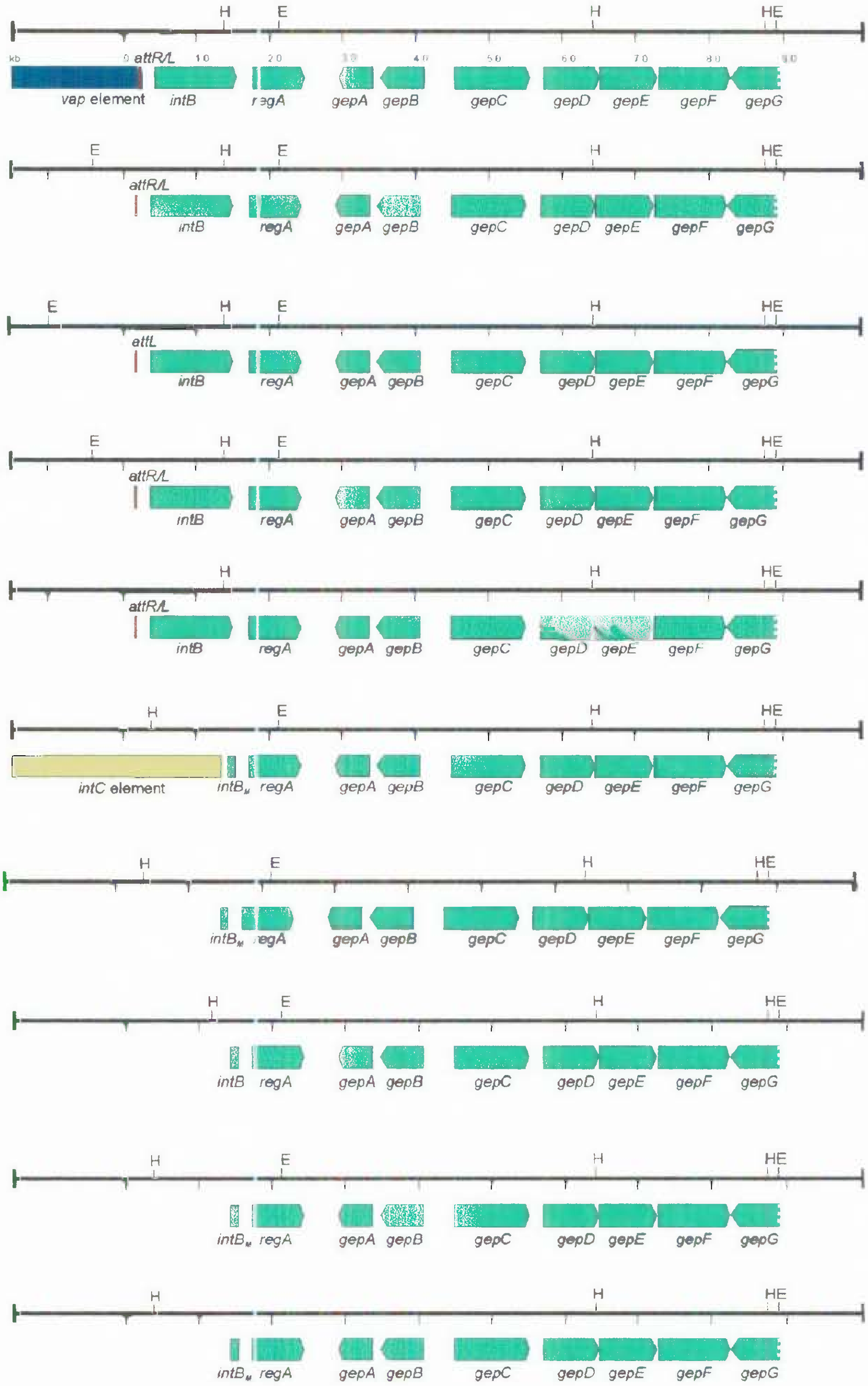
All seventeen strains of *D. nodosus* analysed in this work contained at least one copy of the *intB* gene or part thereof. Only two strains of *D. nodosus* (B1006 and G1220) contained a single copy of the *intB* gene (probe 1, Figure 5.13). In eleven strains (A198, C305, 1311, 1311A, 819, 1169, 2483, 1493, 3138, 1469 and D1172) two copies of *intB* gene were detected by hybridisation to the *intB* probe, whilst in another four strains (AC3577, H1204, H1215, AC390) the *intB* probe hybridises to three or more copies of the *intB* gene (Appendix 5). Thus, in general the *intB* gene is found in multiple copies in the *D. nodosus* genome.

Results from these Southern blot analyses have been divided into three sections based on the hybridisation to *intB* and to *regA*: probes as follows: copies of *intB* that are adjacent to *regA* (Section 5.2.8.1); copies of *intB* that are not adjacent to *regA* (Section 5.2.8.2); and other copies of *intB* (Section 5.2.8.3).

5.2.8.1 Copies of *intB* that are adjacent to *regA*

A copy of the *intB* gene, or part thereof, is adjacent to *regA* in almost all strains analysed (Figure 5.14). Based on co-hybridisation to *intB* and *regA* probes *D. nodosus* strains can be divided into at least three groups. Group 1 strains (strains A198, 1311, 1311A, AC3577, D1172) contain a complete copy of the *intB* coding region which is followed by *regA* and *gepA-G* in consecutive order. Group 2 strains (strains C305, 819, 1469, 1493, 3138) hybridise weakly to the *intB* probe, indicating that the copy of the *intB* gene that is present is not complete, and restriction patterns suggest that this partial copy is probably like the copy of $intB_M$ which is present in strain C305 (Section 1.6.8.1). This hypothesis should be confirmed using PCR to amplify the putative copies of $intB_M$ in these strains, followed by sequencing of the amplified products if necessary. The putative copy of $intB_M$ present in these strains of *D. nodosus* is immediately followed by genes *regA* and *gepA-G* in consecutive order (Figure 5.7). Group 3 strains (strains H1215, H1204, 1169, AC390, 2483, G1220 and B1006) do not contain *gepA*, and although *gepB-G* are present and contiguous in these strains they are not adjacent to the copy of *regA*, nor to the copy of the *intB* gene where it is present (Figure 5.14).

A198



1311

1311A

AC3577

D1172

C305

819

3138

1469

1493

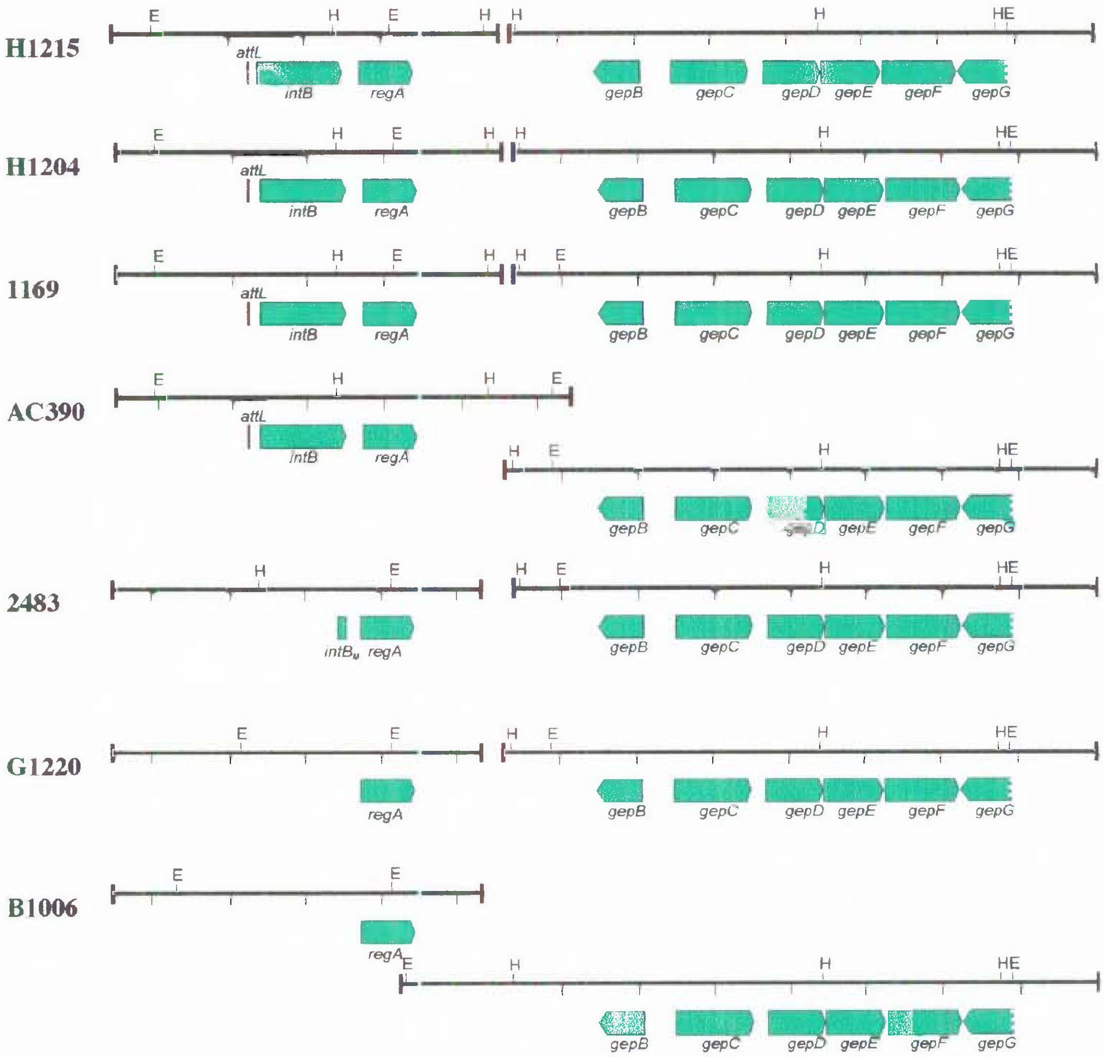


Figure 5.14: Restriction map of the *intB* element and adjacent sequences in which *intB* is adjacent to *regA*, in seventeen strains of *D. nodosus* as determined by Southern blot analyses. Restriction sites shown include *Eco*RI (E) and *Hin*dIII (H). For strain A198 *Sac*I (S) and *Nru*I (N) sites are also shown. The sequences that flank the left end of the *intB* element are also show where known and include: the *vap* element (blue) or the *intC* element (yellow). Putative attachment sites are indicated as *attL* (left) or *attR* (right) respectively. Note that in B1006 and G1220, there is only one copy of *intB* that is not adjacent to *regA*, hence the arrangement shown above for the two aforementioned strains.

In addition, in all strains analysed *gepB-G* are contiguous with sequences at least 17 kb to the right of *gepG*, since probe 7 (Figure 5.13) hybridises to the same *Eco*RI (>23.1 kb) and *Hin*dIII (17.7 kb) fragments in all strains of *D. nodosus* analysed (Appendix 5).

Since the *intB* gene is adjacent to the *attL* site in the *D. nodosus* genome, it is not surprising that the *intB* gene would hybridise to different sized restriction fragments, in different strains of *D. nodosus*, dependent upon which sequences flank the left-end of this element. In the strains of *D. nodosus* analysed in this work, four different restriction patterns are observed upstream of the copy of the *intB* gene, indicating that there are at least four different positions for this copy of the *intB* gene in the *D. nodosus* genome:

(i) it was previously known that in strain A198, this copy of *intB* is integrated next to the *attR* of *vap* region 3 (Figure 5.14) (Bloomfield *et al.*, 1997b). Since no other strain of *D. nodosus* analysed contains a copy of *intB* which hybridises to a 5.1 kb *Eco*RI, and a 3.8 kb *Hin*dIII fragment (Appendix 5), results suggest that *intB* gene sequences are not adjacent to *vap* region 3 in any other strain;

(ii) in *D. nodosus* strain C305, $intB_M$ is located next to *intC* element sequences (Bloomfield, 1997), and the $intB_M$ probe hybridises to a 5.8 kb *Hin*dIII fragment and an 8.8 kb *Eco*RI fragment. Similarly a faint 5.8 kb *Hin*dIII fragment is present in strains 819, 1469, 1493 and 2483 (Appendix 5), suggesting that the $intB_M$ gene and associated sequences are immediately adjacent to a truncated copy of the *intC* element, as they are in strain C305. In strain 3138 the *intB* probe also hybridises to an 8.8 kb *Eco*RI fragment, however the *Hin*dIII fragment is slightly smaller (5.2 kb), suggesting that there may be a small deletion within the *Hin*dIII fragment in this strain or a restriction site polymorphism. Alternatively, in strain 3138 this copy of the *intB* gene could be integrated in a different position in the genome;

(iii) in strains 1311, AC3577 and D1172 the *intB* gene hybridises to 2.5 kb *Eco*RI, 10.2 kb *Hin*dIII and 1.9 kb *Hin*dIII/*Eco*RI fragments. No other strain analysed hybridises to these restriction fragments, indicating that in these strains there is a third position for the copy of *intB* that is associated with *regA*;

(iv) in strains 1311A, H1215, H1204, 1169 and AC390 the *intB* gene hybridises to 3.3 kb *Eco*RI, 3.6 kb *Hin*dIII and 2.7 kb *Eco*RI/*Hin*dIII fragments, indicating that in these strains the copy of the *intB* gene is found in a fourth position in the *D. nodosus* genome.

In strains (G1220, B1006) there is no copy of the *intB* gene that is associated with *regA* in the genome. From these results it appears that at least two of four possible integration sites for the *intB* gene are as yet unknown. Consequently, the identification of these other two sites at which the *intB* gene sequences are found in *D. nodosus* were the subject of further investigations that are discussed in Chapter 7 (Section 7.2.8).

In most strains of *D. nodosus* a single copy of the *regA* gene (probe 2, Figure 5.8) is present in the genome, however in five strains (1311, 1311A, AC390, D1172, H1215) two copies of *regA* are present (Appendix 5). The Southern blot results indicate that where there is a second copy of *regA* present in the *D. nodosus* genome it is not adjacent to a copy of *intB*, but is instead located elsewhere in the *D. nodosus* genome. Since these alternative positions in the genome are unknown they have not been included on the maps that have been constructed for each strain (Figure 5.14).

Although *intB* is present in multiple copies in the genome of most strains of *D. nodosus*, and *regA* genes are present in multiple copies in some strains analysed in this work, the sequences located downstream from the *intB* and *regA* genes in strain A198, including *gepA-G*, are never present in multiple copies.

In seven of the seventeen strains analysed (41%) the *gepA* gene is absent, although *gepB*, *gepC*, *gepD*, *gepE*, *gepF* and *gepG* genes are present in consecutive order in all of the strains analysed (Figure 5.14). The *gepA* gene is absent in both virulent (B1006, G1220, H1204, H1215) and benign (1169, 2483, AC390) isolates of *D. nodosus* and so there is no direct correlation between the presence or absence of *gepA* and virulence. Furthermore, the absence of *gepA* in several strains of *D. nodosus* suggests that the *gepA* gene is dispensable.

In all cases, where *gepA* is present in the *D. nodosus* genome, it is located between the *regA* gene and the *gepB* gene. In contrast, where the *gepA* gene is absent, *gepB-G* genes are not immediately adjacent to *regA* (Figure 5.14). There is no evidence to suggest that the *gepA* gene is an independently mobile sequence (Table 5.1) since it shows no similarity to transposases, and is not flanked by repeated sequences.

5.2.8.2 Copies of *intB* that are not adjacent to *regA*

Copies of *intB* that are not adjacent to a copy of *regA* can be divided into six groups (Figure 5.15). In strain A198, $intB_N$ hybridises to a 2.2 kb *Eco*RI and a 6.8 kb *Hin*dIII fragment and is located downstream of *vap* region 2. There are three other strains, 1311, 1311A, and D1172 which hybridise to restriction fragments of identical size, indicating that a copy of $intB_N$ is likely to be integrated in the same position as in strain A198 (Group A).

In strain C305, a copy of *intB* probe hybridises to 13.8 kb *Eco*RI, 4.3 kb *Hin*dIII fragments, which corresponds to a copy of $intB_N$ which is adjacent to a copy of *pnpA*. Since *D. nodosus* strains C305, 1169, 2483, 1493 and 1469 hybridise to restriction fragments of a similar size, they are also appear to contain a copy of $intB_N$ in the same position of the *D. nodosus* chromosome (Group B).

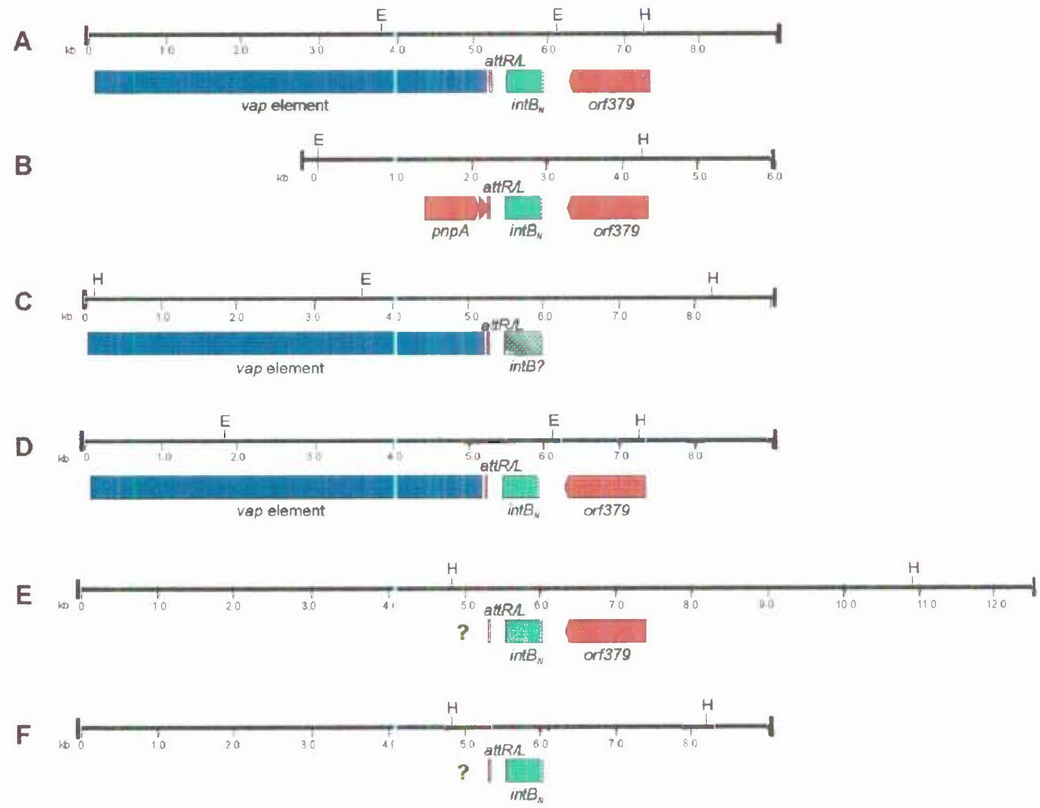


Figure 5.15: Restriction map of the *intB* sequences which are not adjacent to a copy of *regA*, in seventeen strains of *D. nodosus* as determined by Southern blot analyses. Restriction sites shown include *Eco*RI (E) and *Hin*dIII (H). The arrangement of *intB*$_N$ sequences in strains A198, 1311, 1311A, D1172 (A); in strains C305, 1169, 2483, 1493, 1469 (B); in strains H1215, H1204 (C) (it is unknown whether this copy of the *intB* gene is a partial or complete copy, and so it has been distinguished by a spotted green box); and strains B1006, G1220 (D); in strain 819 (E); strain 3138 (F). Putative attachment sites are indicated as *attL* (left) or *attR* (right) respectively. Numbers indicate the distance in kb. The position of the *intB*$_N$ gene in strains A198, B1006, G1220, H1204 and H1215 (Bloomfield *et al.*, 1997b) and strain C305 (Shaw, 1997) was determined previously, and so the position of *intB*$_N$ relative to *pnpA* and the *vap* element is indicated in those cases. It has been assumed that strains with an identical restriction pattern are integrated in the same position in the *D. nodosus* genome. Where sequences adjacent to *intB*$_N$ are unknown it is indicated (?).

Although it is known that in strains H1204 and H1215 there is a copy of the *intB* gene integrated downstream from the *vap* region (Bloomfield *et al.*, 1997b), it is unknown whether this copy of the *intB* gene corresponds to a copy of *intB*$_N$ or some other derivative of the *intB* gene (Group C).

Similarly, strains B1006 and G1220 (Group D) contain only single copies of the *intB* gene, and since the *intB* probe hybridises weakly to a 4.0 kb *Eco*RI fragment, this copy of the *intB* gene is likely to be a partial copy. A copy of the *intB* gene was previously found to be adjacent to the single copy of the *vap* region which is present in these strains (Bloomfield *et al.*, 1997b).

In *D. nodosus* strains 819 (Group E) and 3138 (Group F) the copy of *intB* that is not adjacent to *regA* is found on unique *Hin*dIII fragments of 5.6 kb and 3.3 kb respectively. The position of either of these copies in the *D. nodosus* genome is unknown. It is interesting that in strain 3138, the *intB* probe hybridises to the same size *Hin*dIII fragment (3.3 kb) as original laboratory strains of *D. nodosus* strain C305 (C3051) (Section 1.6.8.2), suggesting that strain 3138 may contain the sequences that separated *pnpA* and *intB*$_N$ in the strain C3051 that have been lost from the genome of current laboratory strain of C305 (C3052) (Section 1.6.8.2). This is the subject of further discussion in Chapter 6.

5.2.8.3 Other copies of *intB*

Although most strains of *D. nodosus* have two copies of the *intB* gene, there are some strains in which the number of copies exceeds two, including strains AC3577, AC390, H1215 and H1204. Copies in excess of two have not been discussed in this work because it is unknown where these copies are located, how intact they are, or whether they are attributable to cross hybridisation to a related integrase gene which is present in only some strains. In addition, since strains AC3577 and AC390 do contain more than two copies of the *intB* gene, and none of these have been previously associated with a *vap* region, it is not possible to determine which *Eco*RI, *Hin*dIII, and *Hin*dIII/*Eco*RI fragments are associated, and thus even the second copies of *intB* present in these strains were not discussed in Section 5.4.2.2.

The large degree of restriction fragment length polymorphisms evident in Southern blot analyses in which the *intB* gene was utilised as a gene probe (Appendix 5), suggests that this gene could be a useful probe for DNA fingerprinting of different *D. nodosus* isolates in epidemiological studies.

5.3 Discussion

The identification of a putative attachment site (*att*), a putative integrase gene (*intB*), a putative regulator (*regA*) with similarity to a regulator present on extrachromosomal elements, and the similarity of *gepA* to putative protein *rteC* from a *Bacteroides thetaiotomicron* conjugative transposon, in work done previously (Bloomfield, 1997) suggested that these genes may have comprised part of an integrated genetic element (Table 5.1).

In this work, sequences 4.3 kb downstream of the first 201 nt of the *gepC* coding region in *D. nodosus* strain A198 were isolated (Section 5.2.1) and determined (Section 5.2.2). These sequences contained five potential open reading frames called *gepC*, *gepD*, *gepE*, *gepF* and *gepG* respectively. Analysis of the potential proteins encoded by these genes suggests that *gepC* encodes a integral-membrane protein, *gepD* and *gepE* encode a putative sulfate binding-protein dependent ABC importer, and *gepF* encodes a protein required for the early steps of thiamine biosynthesis. GepG did not show similarity to proteins of known function (Section 5.2.3).

Subsequent Southern blot analyses of these regions that are adjacent to *intB*, *regA*, indicated that in all seventeen strains of *D. nodosus* analysed *gepB-G* are present in single copy, are clustered together, and are contiguous to sequences at least 17 kb to the right of *gepG* in the chromosome (Figure 5.13). It is most probable, that *gepB*, *C*, *D*, *E*, *F* and *G* do not correspond to genes of an integrated *intB* element, primarily because *gepC*, *gepD* and *gepF* appear to encode housekeeping genes, that would be absolutely required by a microorganism. Furthermore, *gepC*-, *D*-, *F*-like genes are, in general, present in single copies in bacterial genomes (Boos & Lucht, 1996).

Although there are many ABC exporter complexes that are encoded by autosomal elements, there are very few examples of periplasmic binding protein-dependent ABC importer proteins. The best characterised of these is the *A. tumefaciens* plasmid pTi15955 which encodes a mannopine import system (Oger *et al.*, 1998). The secretion of mannopine by plant tumors induces conjugal transfer of octopine-mannityl-opine-type Ti plasmids. Other examples of periplasmic binding-protein dependent systems include the *nodLMN* genes that induce nodulation in *Rhizobium leguminosarum* encoded by plasmid pRL1J1 (Surin & Downie, 1988), and that encoded by the *Rhizobium* sp. plasmid pNGR234a, both of which have unknown functions (Freiberg *et al.*, 1997). Thus, although it is possible that the *intB* element of *D. nodosus* carries a binding-protein dependent ABC importing system, it seems unlikely.

One of the most significant variations between different strains of *D. nodosus* is that in strains in which *regA* and *gepB* are not adjacent, the *gepA* gene alone is missing (Figure 5.14). The presence of the gene *gepA* in some strains of *D. nodosus*, and its absence in other strains of *D. nodosus* is consistent with the idea that the *gepA* gene comprises part of an integrated genetic element, and the absence of the *gepA* gene could be indicative of strains in which deletions of the integrated element had occurred. However, all of those genes that are located immediately downstream of *gepA* in strain A198 are present in other strains of *D. nodosus*, irrespective of whether they contain a copy of the *gepA* gene or not.

The simplest explanation is that the *intB* element consisted of *intB*, *regA* and *gepA*, and that in some strains sequences may have been inserted between *regA* and *gepB*. Such an insertion may have resulted in the concomitant deletion of the *gepA* gene rather than the repositioning of the *gepA* adjacent to the insertion site, and the separation of *regA* and *gepB* in the *D. nodosus* genome.

However it is unlikely that the *intB* element was so small, given the size of related *D. nodosus* elements (*vap* element, *intC* element) which are at least 7 kb in length, containing many more than three genes. Therefore, *intB*, *regA* and *gepA* may be only part of a primordial genetic element, which in strain A198 is immediately followed by genes *gepB-G* which are not part of the *intB* element. It is possible that in those strains in which *regA* and *gepB* are not adjacent, a new sequence integrated between *regA* and *gepB* or a DNA rearrangement separated *gepB* from *regA*.

It is also possible that there may have originally been related genetic elements that contain the genes *intB* and *regA*. One of these 'related-elements' may contain *gepA*, whilst the other does not. The analysis of other genetic elements in *D. nodosus*, including the *vap* element and the *intC* element suggests that various derivatives of these elements exist in *D. nodosus*, and the differential divergence of sequences within the *vap* element has been reported previously (Cheetham *et al.*, 1995b). *vap* region 1 contains a copy of *vapE* whilst *vap* region 3 contains a similar but divergent gene, called *vapE'*, however both *vap* elements contain an identical copy of *vapD*. Since it would be expected that the genes within two *vap* elements would diverge equally from ancestral genes in a related element, it is likely that *vap* region 1 and *vap* region 3 constitute different but related elements.

It is apparent from the maps of the *intB* elements and adjacent regions (Figure 5.14) that there is some divergence in the sequences that separate *regA* and *gepB* in group 3 strains, to the right of *regA* and to the left of *gepB*, since the restriction patterns vary.

Long-range PCR could be utilised to amplify the putative sequences between *regA* and *gepB* in Group 3 strains (H1215, H1204, 1169 AC390, 2483, G1220 and B1006) of *D. nodosus*. If amplification of the intervening sequences were successful, analyses of them would give insight into the evolution of the *intB* element in different strains of *D. nodosus*, and may reveal whether the genes *intB*, *regA* and *gepA* are the remnants of a primordial integration event.

In summary, results suggest that *gepB-gepG* are not part of an integrated genetic element, and instead are almost certainly housekeeping genes, which are clustered together and contiguous with sequences to the right of *gepG* in all strains studied. In contrast, there is a great deal of variation in sequences located to the left of *gepB* in different strains of *D. nodosus*, consistent with the hypothesis that sequences left of *gepB* comprise part of an integrated genetic element or are indicative of some DNA rearrangement event, in all strains analysed. Results suggest that in Group 1 and Group 2 strains only a truncated copy of a primordial *intB* element consisting of *intB*, *regA* and *gepA* remains, whilst in Group 3 strains another genetic element separates *regA* and *gepB* genes.