J Mol Evol (1994) 39:631–643

Journal of Molecular Evolution

Tracing the Spread of Fibronectin Type III Domains in Bacterial Glycohydrolases

Elizabeth Little, Peer Bork, Russell F. Doolittle

Center for Molecular Genetics, University of California, San Diego, La Jolla, CA 92093-0634, USA

Received: 23 April 1994 / Accepted: 1 July 1994

Abstract. The evolutionary spread of 22 fibronectin type III (Fn3) sequences among a dozen bacterial enzymes has been traced by searching databases with the non-Fn3 parts of the enzyme sequences. Numerous homologues were found that lacked the Fn3 domains. In each case the related sequences were aligned, phylogenetic trees were constructed, and the occurrences of Fn3 units on the trees were noted. Comparison with phylogenetic trees prepared from the Fn3 segments themselves allowed inferences to be made about when the Fn3 units were shuffled into their present positions.

Key words: Fibronectin type III — Bacteria — Glycohydrolases — Phylogeny — Horizontal gene transfers

Introduction

The fibronectin type III domain (Fn3) is a very common constituent of animal proteins, occurring in roughly 2% of known sequences (Bork and Doolittle 1992). The unit has a distinctive motif (Patthy 1990; Bazan 1990) that has been shown by experiment to be a seven-stranded β sandwich similar to the immunoglobulin fold (Baron et al. 1992; de Vos et al. 1992; Leahy et al. 1992). The domain occurs in intracellular, extracellular, and membrane-spanning proteins, often in multiple copies. Not infrequently, they are set off genetically by introns, and their broad distribution in animal proteins is often regarded as an example of "exon shuffling," although it should be noted that the units often contain an internally situated intron also.

The discovery of a protein in bacteria that contained two obvious Fn3 sequences initially led to the proposal that this structure must have been present in the common ancestor of prokaryotes and eukaryotes (Watanabe et al. 1990). So far the domain has not been observed in plants, fungi, or protists, however. Moreover, subsequent events have revealed that among the bacteria, Fn3 sequences are found only in one particular class of proteins: extracellular glycohydrolases (Gilkes et al. 1991; Bork and Doolittle 1992; Hansen 1992). The glycohydrolases are extremely diverse in their own right, having been already classified into at least 45 families (Henrissat 1991; Henrissat and Bairoch 1993). Even among these, the distribution of Fn3 domains is remarkably sporadic. This unusual distribution, along with other considerations, led us to propose that bacteria actually acquired the gene for the domain from an animal source (Bork and Doolittle 1992). Beyond that, it seems clear that DNA segments encoding the domain are being shuttled around bacterial genomes in a manner that suggests module shuffling (Meinke et al. 1991a; Gilkes et al. 1991; Bork and Doolittle 1992). If—as appears to be the case—these rearrangements are evolutionarily recent, then they would be examples of intron-independent "exon shuffling."

The goal of the present study was to trace the spread of the Fn3 unit among the bacteria and, if possible, to pinpoint the initial acquisition. In our previous study (Bork and Doolittle 1992) we searched databases with

Correspondence to: R.F. Doolittle

Table 1. Distribution of 22 prokaryotic Fn3 units

Fn3 No.	Enzyme	Organism	Gram	No. units
F1	Polygalacturonosidase	*Erwinia chrysanthemi*	−	1
F2,3,4	Endoglucanase B	*Cellulomonas fimi*	+	3
F5,6	Endoglucanase D	*Cellulomonas fimi*	+	2
F7	Cellulase	*Cellulomonas flavigena*	+	1
F8,9	Chitinase A	*Bacillus circulans*	+	2
F10	Chitinase D	*Bacillus circulans*	+	1
F11	Exochitinase	*Streptomyces olivaceoviridis*	+	1
F12	Chitinase 63	*Streptomyces plicatus*	+	1
F13,14	Pullulanase 1	*Clostridium thermohydrosulfuricum*	+	2
F15,16	Pullulanase 2	*Clostridium thermohydrosulfuricum*	+	2
F17,18	Pullulanase	*Clostridium thermosulfurogenes*	+	2
F19,20,21	Amylase A-180	Alkaliphilic eubacteria	+	3
F22	PHB depolymerase	*Alcaligenes faecalis*	−	1

the intent of identifying all possible occurrences of Fn3 sequences. Now we have inverted the process, the object being to find homologous proteins that lack the Fn3 domains. In most cases this meant searching those sequences that constitute the catalytic units of the various bacterial glycohydrolases. We constructed phylogenetic trees from those homologous segments in an effort to find where Fn3 units first made their appearance within a group. These trees were subsequently compared with trees based on the Fn3 sequences themselves. The results are consistent only with an unorthodox and apparently recent genetic transfer of Fn3 units among several bacterial groups.

Methods and Data Sources

Sequences were taken either from the Protein Identification Resource or from the on-line GenBank service available from NCBI. The PIR or GenBank identifier codes for sequences are provided in the appropriate figure legends. Three additional Fn3 units from bacteria came to light after the completion of this study and are referred to in the text but are not addressed in the tables or figures. The programs used for routine editing, inspection, and searching procedures have been described previously (Doolittle 1987). Typically, a sequence entry known to contain one or more Fn3 domains was edited to remove the Fn3 portions and the remainder searched against all three sequence collections available on the PIR (release 38). Candidates were retrieved and suitable segments were cut out for preliminary alignment. Patterns were then constructed from these alignments by the procedure of Rohde and Bork (1993) and the database was then rescreened in an effort to identify more distantly related members of each group.

After a reasonable set of related sequences had been assembled for most of the entries known to contain Fn3 units, final alignments were made (Feng and Doolittle 1990; Doolittle and Feng 1990).

Results

In our initial study (Bork and Doolittle 1992) we reported on 13 Fn3 units in seven different bacterial enzyme sequences. In the interval since, at least a dozen more bacterial Fn3 units have been found, nine of which have been used in this study. These were: three more from clostridial pullulanases (Mathupala et al. 1990; Burchhardt et al. 1991), two chitinases from different species of *Streptomyces* (Blaak et al. 1993; Robbins et al. 1992), one in a chitinase D from *Bacillus circulans* (Watanabe et al. 1992), and two in an endoglucanase D from *Cellulomonas fimi* (Meinke et al. 1993). In addition, we were able to add a putative ninth new entry by translating one of the clostridial pullulanases in an alternative reading frame different from the one reported by the authors (Burchhardt et al. 1991). All told, we scrutinized 22 Fn3 sequences from 12 different bacterial enzymes representing seven different bacterial genera; 20 of the 22 occur in gram-positive bacteria (Table 1). As an aid to following the course of evolution, we assigned each of the 22 individual Fn3 units an identifier code from F1 to F22 (Table 1 and Fig. 1).

Galacturonosidases

The first of the Fn3-containing sequences to be examined phylogenetically was the exo-poly-α-D-galacturonosidase found in *Erwinia chrysanthemi* (He and Collmer 1990). Ten related sequences were culled from the PIR database, including three fungal and four plant entries, as well as three additional bacterial enzymes. All were galacturonases of the sort assigned to family 28 by Henrissat (1991). With the exception of the *Erwinia* sequence, none contained any evidence of an Fn3 domain.

A phylogenetic tree was constructed from an alignment of a conserved segment of about 250 residues covering the catalytic region of these enzymes (Fig. 2). In this case the Fn3-containing sequence from *E. chrysanthemi* was the outlier, and as a result, a judgment could not be made as to whether the *Erwinia* lineage acquired the Fn3 module or whether it was lost on the way to all the others.

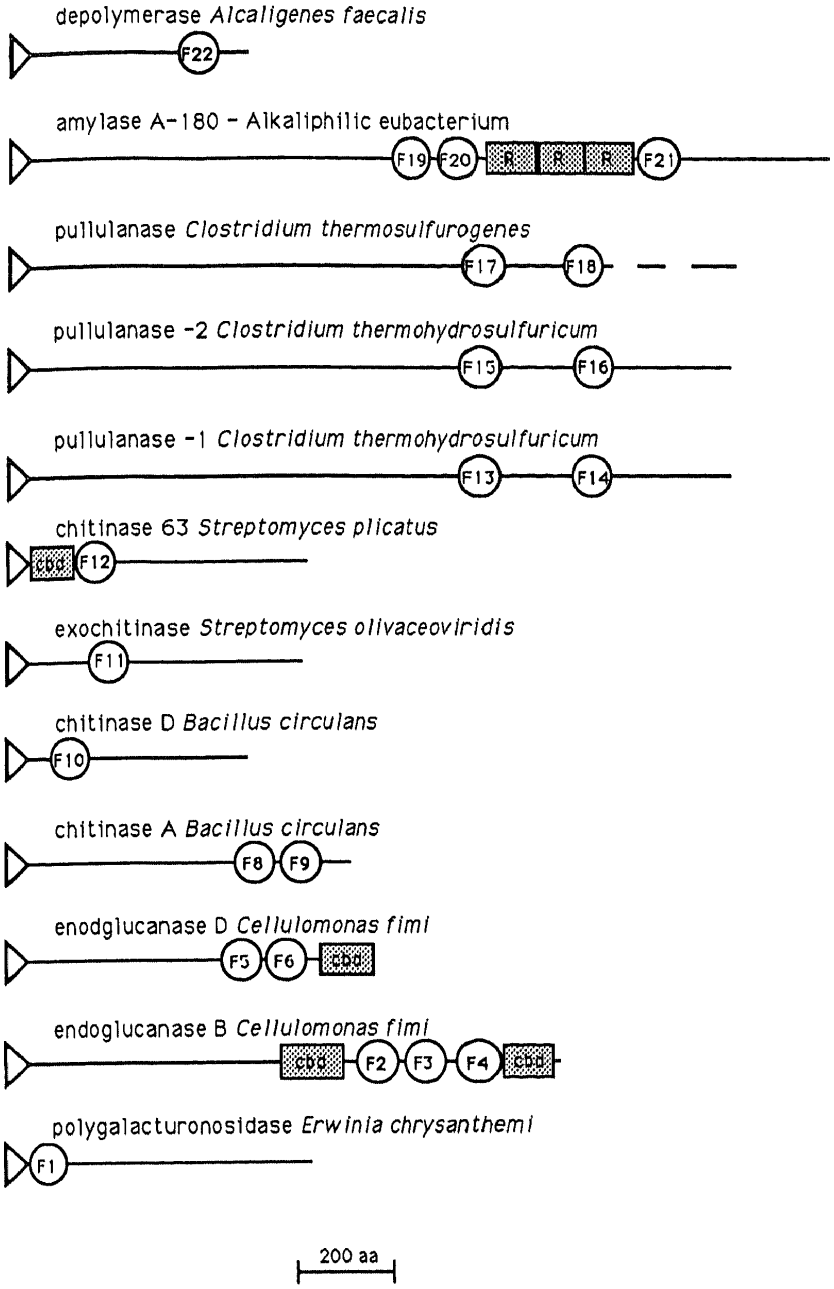

Fig. 1. Schematic depiction of 12 bacterial enzymes known to contain Fn3 units. The Fn3 units are denoted F1 through F22. (The unit denoted F7 occurs in an endoglucanase from *C. flavigena* that is not shown because the overall sequence of the enzyme is not yet reported.) The unit denoted *F16* and the *dashed line* following it in the pullulanase from *C. thermosulfurogenes* were identified by translating an alternative frame from the one initially reported (Burchhardt et al. 1991). Some modules denoted by other workers to be carbohydrate-binding domains are denoted by the *boxes* containing the designation *cbd*. An obvious tandem repeat in the amylase A from alkaliphilic bacteria has been designated by boxed *Rs*. The terminal *triangles* denote signal peptides.

Cellulases

The second Fn3-containing sequence to be traced was a cellulase from *Cellulomonas fimi*. The enzyme, which was described as an endoglucanase B (Meinke et al. 1991b), contains three Fn3 units (F1, F2, and F3). A fragment from a cellulase from another species, *C. flavigena*, was noted to have a similar sequence (Meinke et al. 1991c). We retrieved nine sequences related to the cellulase portions, including seven from prokaryotes and two from eukaryotes. None of the nine contained identifiable Fn3 domains. The enzymes in this group belong to Henrissat's family 9. The phylogenetic tree that emerged from an alignment of a 330-residue homologous segment revealed that in this case the Fn3 domains must have been acquired late in the diversification (Fig. 3).

Another cellulase in which Fn3 units were found recently involves an endoglucanase D from *C. fimi* (Meinke et al. 1993). We found five obviously related sequences, all of them from bacteria, none of which had Fn3 domains (Fig. 4). These enzymes are members of Henrissat family 5. Interestingly, the two Fn3 units (F5 and F6) are situated similarly to three units in the endoglucanase B (Fig. 3). except that a carbohydrate-binding domain has been translocated. It is notable also that homologous enzymes from *Clostridia* and *Bacilli* lack Fn3 units.

Chitinases

Two different Fn3-containing chitinases have been found in *B. circulans* (Watanabe et al. 1990, 1992). By coincidence, in both instances closely related chitinases have been reported in *Streptomyces*, one in *S. plicatus* (Rob-

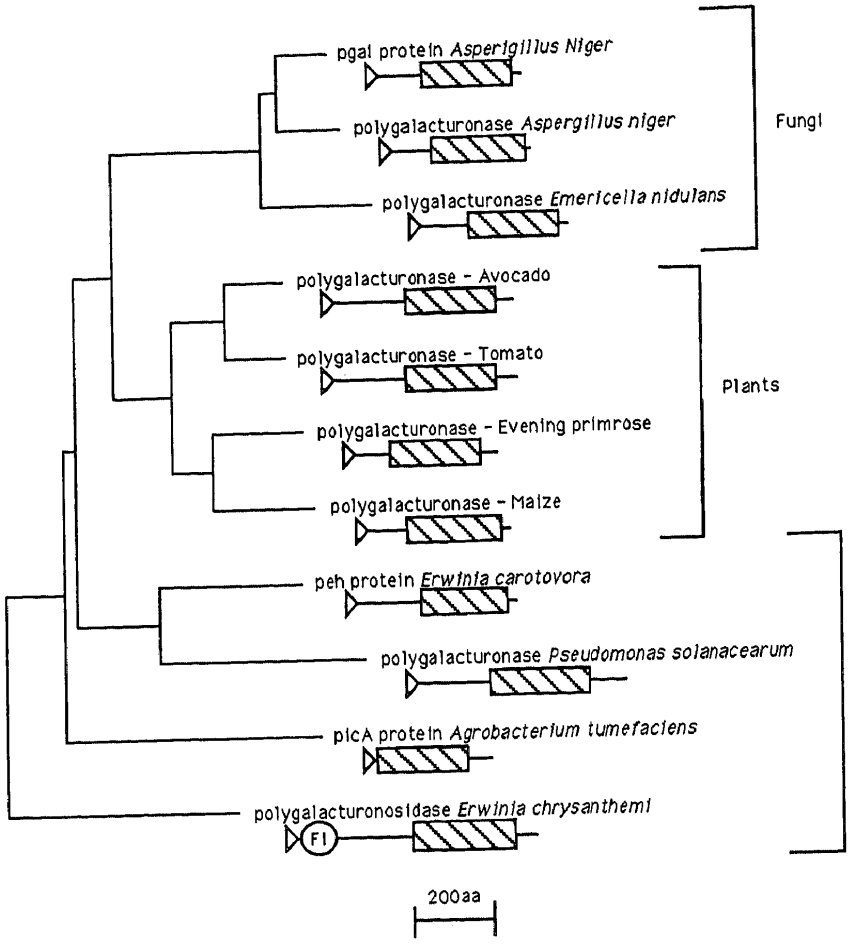

Fig. 2. Phylogeny of 11 polygalacturonase sequences (Henrissat family 28), only one of which contains a Fn3 sequence (F1). The tree is based on the alignment of the shaded regions (ca. 250 amino acids). The 11 sequences have the following PIR codes (top to bottom): S17980, S16303, S24156, S28072, A25534, JQ0992, S16998, S11773, A44508, S40364, and A36715.

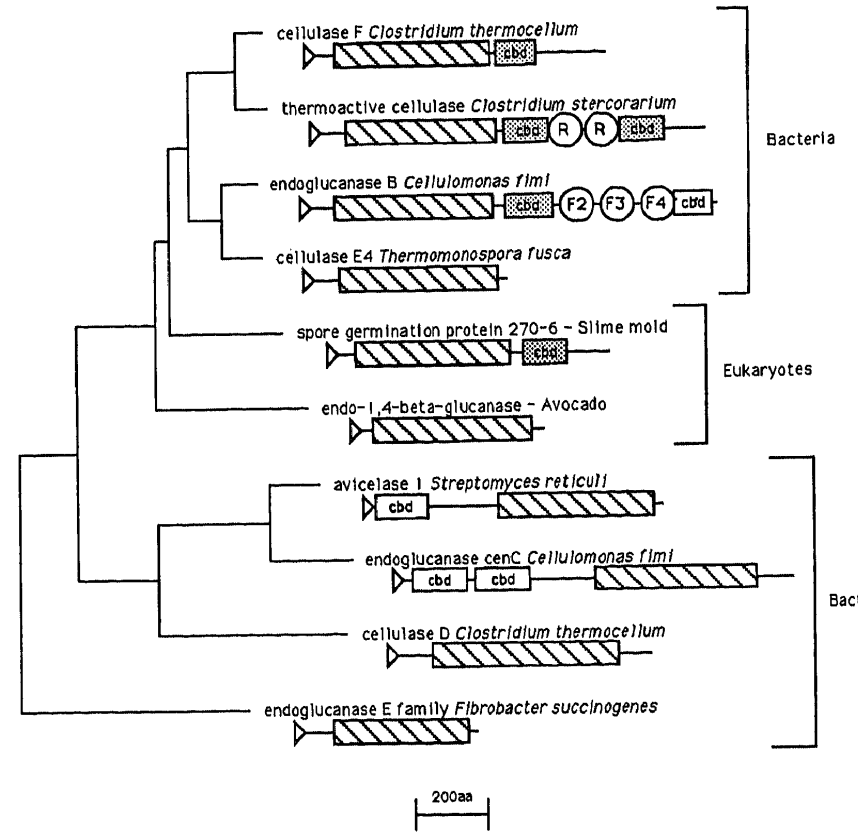

Fig. 3. Phylogeny of ten cellulases (also called endoglucanases) (Henrissat family 9), only one of which contains Fn3 units (F2, F3, and F4 in the endoglucanase B from *C. fimi*). Several of the enzymes contain modules thought to be carbohydrate-binding domains (cbd). An obvious tandem repeat in the cellulase from *C. stercorarium* is denoted by the circled *R*s. (This repeat is different from the one noted in Fig. 1, however.) PIR codes (top to bottom): S15727, S12021, A39199, B42360, A35621, S11946, S21398, S15271, CZCLDM, and A39416.

bins et al. 1992) and the other in *S. olivaceoviridis* (Blaak et al. 1993). In the case of one of the chitinases, the Fn3 domains occur in different places and different numbers in *B. circulans* and *Streptomyces plicatus* (Fig. 5). In the other (Fig. 6), they are situated similarly. Actually, three different species of *Streptomyces* have similar chitinases. Two of these, *plicatus* (shown in Fig. 5) and *lividans* (Fujii and Miyashita 1993; not shown), are more than

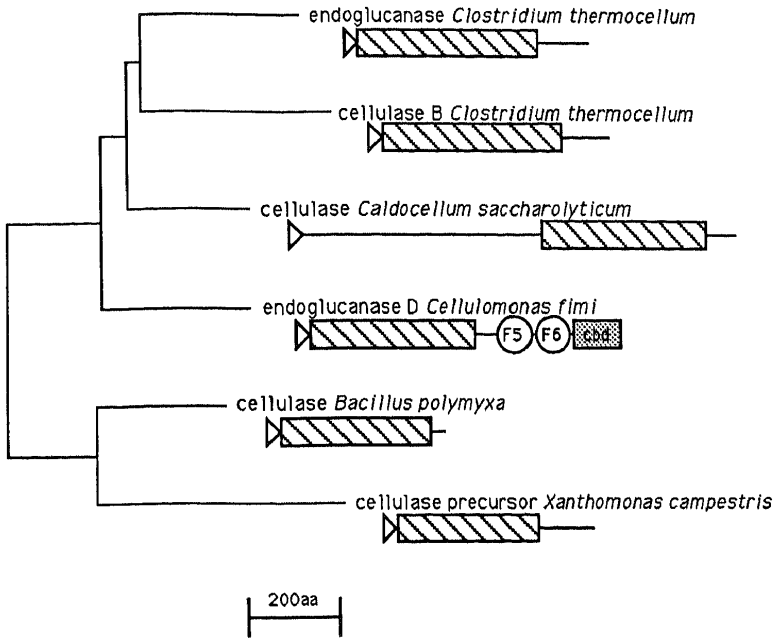

Fig. 4. Five sequences related to the endoglucanase D of *C. fimi* were retrieved from PIR. All were from bacteria, including *Clostridia* and *Bacilli*. All are members of Henrissat family 5. The PIR call numbers are (top to bottom): S31381, CZCLBM, S02711, B47093, A35136, and JH0158.

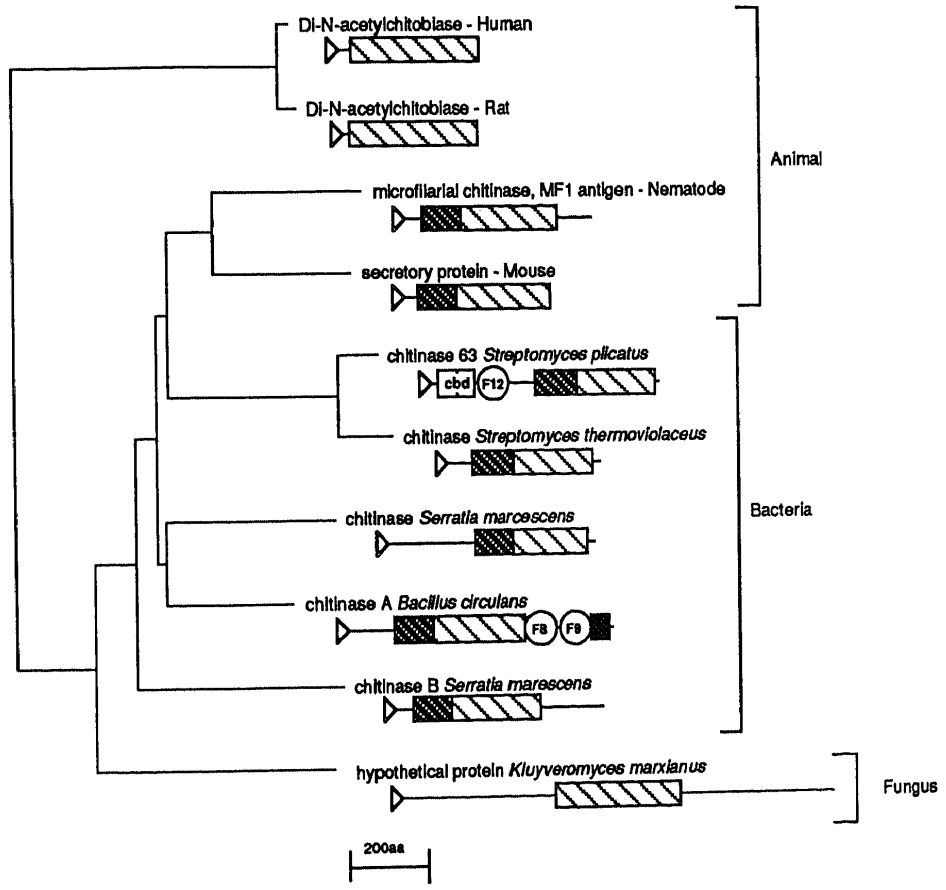

Fig. 5. Phylogeny of ten chitinases from assorted eukaryotes (Henrissat family 20) and prokaryotes (Henrissat family 18). Only the enzymes from *B. circulans* and *S. plicatus* contain Fn3 units (F7 and F10). Phylogenetic tree based on alignment of the *shaded regions* (ca 300 amino acid residues). The more *darkly shaded regions* denote regions that are readily aligned with the sequences shown in Fig. 6. The *very tightly shaded box* on the carboxy-terminal side of F9 is a region that is 62% identical with the similarly shaded region in Fig. 6 that occurs in the amino-terminal side of F10. PIR codes (or GenBank accession number) (top to bottom): S27959, S27882, A38221, S27879, A25090, GenBank accession number L07762, A38368, JH0573, S04856 and S22786.

90% identical, including the Fn3 domain and a carbohydrate-binding domain. Remarkably, a third species (*S. thermoviolaceus*) lacks both the Fn3 and carbohydrate-binding domains (Fig. 5), even though its catalytic domain is 74% identical to the others (Tsujibo et al. 1993). The enzyme from *Serratia marcescens* also lacks these domains, as do several eukaryotic homologues (Fig. 5). Interestingly, the bacterial enzymes in this group are members of Henrissat family 18, but the eukaryotic enzymes are from family 20.

In the case of the other chitinase (Fig. 6), a member of Henrissat family 19, the domains (F10 and F11) are situated in similar locations in *Bacillus* and *Streptomyces* (Fig. 6). Remarkably, however, the 46-residue segment on the amino-terminal side of F10 (from *B. circulans*) is 62% identical with the segment on the carboxy-terminal side of F9 (also from *B. circulans*) and must have been independently shuffled into position (Figs. 5 and 6).

Like the other chitinases, numerous members of family 19 lack the Fn3 domain. Moreover, although these

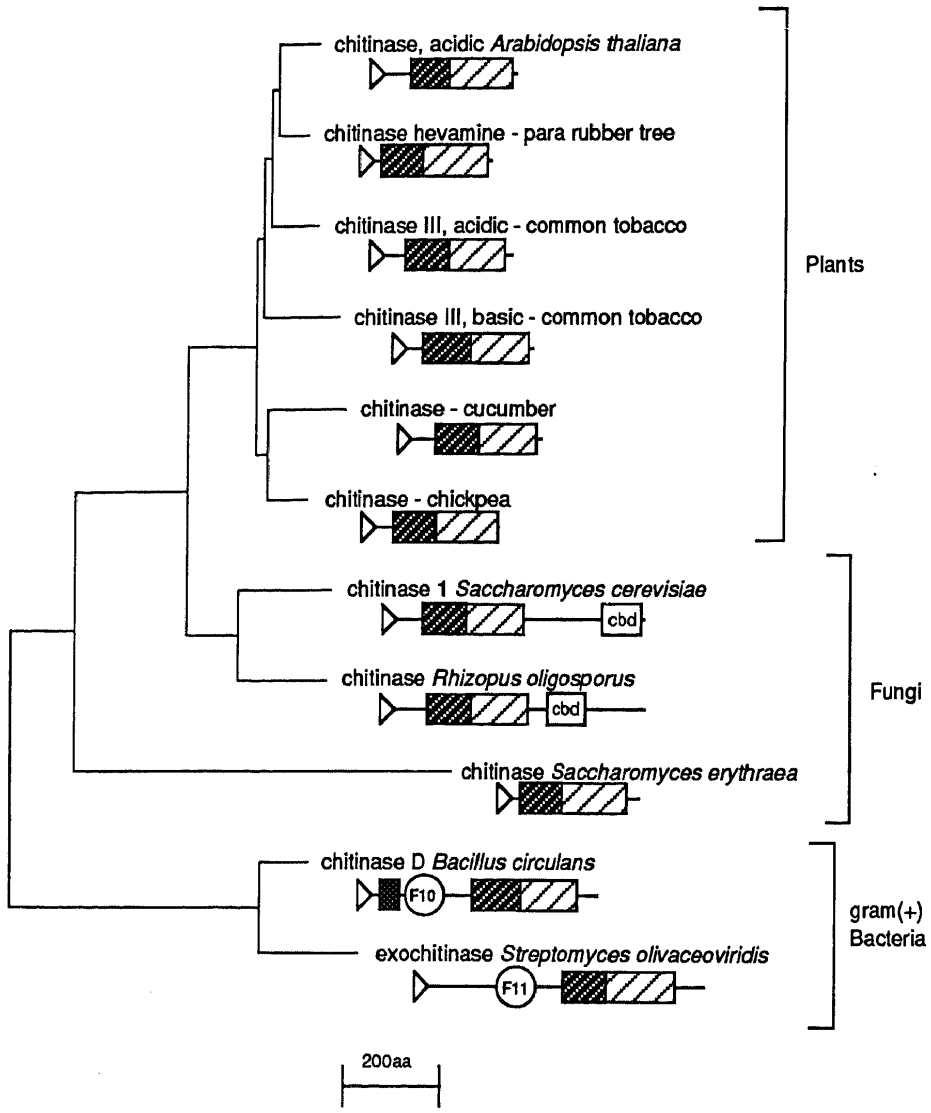

Fig. 6. Phylogeny of 11 chitinases from assorted prokaryotes (Henrissat family 19) and eukaryotes. Only the enzymes from *B. circulans* and *S. olivaceoviridis* contain Fn3 units. Although these sequences have been classified as a different set of glycohydrolases from those shown in Fig. 5 (Henrissat 1991), the *darker cross-hatched regions* are clearly homologous with the other family. See legend to Fig. 5 for explanation of very tightly shaded box next to F10. PIR codes (top to bottom): A45511, S17205, S23544, S23545, A31455, S31763, A41035, S27418, JX0076, A41961, and S32039.

chitinases fall into two separate groups, they have segments (darkly shaded in Figs. 5 and 6) that are obviously homologous, and we were able to construct a large phylogenetic tree that depicts their common origin (not shown).

Amylase-Pullulanase Group

The next Fn3-containing sequence traced was the bifunctional α amylase-pullulanase from *Clostridium thermohydrosulfuricum* (Melasniemi et al. 1990). This 800-residue-long sequence contains two Fn3 domains (F13, F14) separated by about 200 residues. All told, we identified seven other obviously related enzymes, all from bacteria and all either α amylases or pullulanases (Henrissat family 13). Two of these were from closely related *Clostridia* and had Fn3 units at the same locations as the initial sequence (although in the case of one of these, the second Fn3 [F18] became apparent only after we translated the GenBank sequence into an alternative frame). Recently, also, we uncovered another bacterial pullulanase with two Fn3 units at the same locations. The apparently unpublished sequence (Matur and Zeikus, unpublished, GenBank accession number L07762) was reported to be from *Thermoanaerobacter saccharolyticum;* the sequence is very similar to that of the pullulanase from *C. thermosulfurogenes* and is not shown in Fig. 7. The five other amylopullulanases we identified did not contain Fn3 units. (Alternative reading frames were checked when possible.)

An approximately 225-residue segment from each was used in the alignment. The phylogenetic tree (Fig. 7) indicates that pullulanases in ancestral species did not have Fn3 units, nor did related enzymes in *Clostridia.* The degree of similarity between the two Fn3 units suggested that they were the result of a tandem duplication (Table 2).

α-Amylase Group

The next candidate for characterization was an amylase from a eubacterium loosely described as an alkaliphilic gram-positive bacterium (Candussio et al. 1990). In this case an approximately 200-residue segment was used for the alignments and phylogenetic tree (Fig. 8). All told, nine other α amylases with homologous sequences were

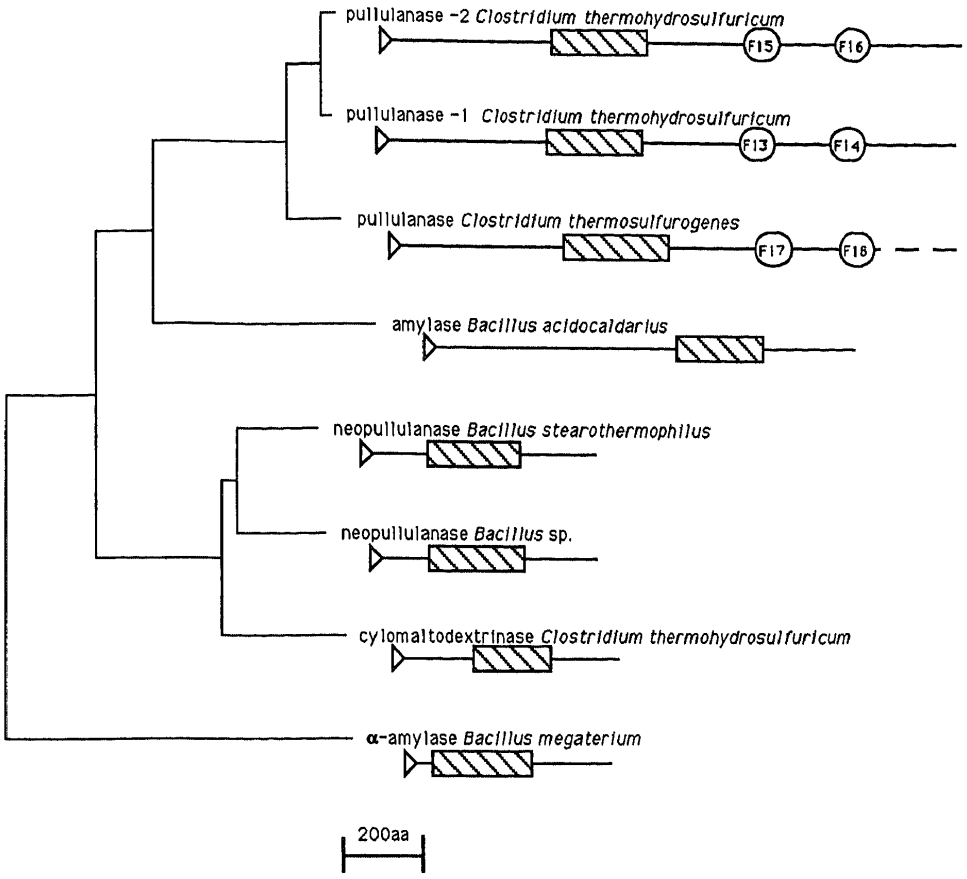

Fig. 7. Phylogeny of eight bacterial amylo-pullulanases (Henrissat family 13), three of which are known to contain pairs of Fn3 units. In the case of the enzyme from *C. thermosulfurogenes* (GenBank accession number M57692) we identified the second Fn3 unit (F18) by examining alternative reading frames to those translated by the original authors (Burchhardt et al. 1991). PIR codes (top to bottom): A44765, S28669, S27545, S18118, A37008, JS0673, A42950, and S01031.

found, including three from fungi; all are members of Henrissat family 13. With the exception of the alkaliphilic eubacterium, none of the amylases have Fn3 units.

The Fn3 units in the alkaliphilic eubacterium are situated on either side of three 99-residue repeats. Candussio et al. (1990) noted that these same repeats also occur twice in a β amylase from *B. polymyxa* (Uozumi et al. 1989) and once in a β amylase from *B. circulans* (Siggens 1987). Neither of the latter enzyme contains identifiable Fn3 units, however.

Polyhydroxybutyrate Depolymerase

The final Fn3-containing enzyme to be searched was a poly(3-hydroxybutyrate) depolymerase from *Alcaligenes faecalis* (Saito et al. 1989). Although searching revealed some potential relatives among a family of α amylases found in *Bacilli,* the resemblances were too marginal to undertake the construction of phylogenetic trees.

Fn3 Relationships

The 22 Fn3 segments from the 12 Fn3-containing enzymes were aligned (Fig. 9), together with a eukaryotic Fn3 sequence (FN10) as an outlier. The branching order was determined by both matrix and character-based methods (Fig. 10A,B). The topologies were viewed from the perspective of both the contributing species and the glycohydrolase family in an effort to follow the trail of the bacterial Fn3 lineage.

Most often the Fn3 units were the only parts of enzymes appearing in a cluster with any significant resemblance. For example, the Fn3 segment (F22) from the polyhydroxybutyrate depolymerase of *Alcaligenes faecalis* is ca. 50% identical with the Fn2 unit in the chitinase from *S. olivaceoviridis* (and also unit F7 in the cellulase from *C. flavigena*), but there is no detectable similarity anywhere else in these sequences. This seems a clear case of horizontal transfer, although it must be pointed out that the two different phylogenetic methods have slightly different groupings (Fig. 10A,B). Although the gram-negative organism is the likely receiver in either case, the donor could be any one of three gram-positives.

The fact that Fn3 sequences from the same enzyme are almost always more similar to each other than they are to other bacterial Fn3 units suggests that tandem duplication of these units is common. Indeed, the degree of similarity of adjacent units, sometimes exceeding 70% (Table 2), argues that the duplications are likely very recent. We have indicated the positions of obvious tandem duplications on the phylogenetic trees (Fig. 10).

G+C Contents

The G+C content of DNA can sometimes be an indicator of the source of bacterial genes. Of particular interest here is the fact that the gram-positive bacteria are ordi-

Table 2. Percent identities and distances between adjacent Fn3 segments in bacterial enzymes

Enzyme	Organism	%ID	d^a
Chitinase A	*Bacillus circulans* (F8,F9)	72.7	18.9
Endoglucanase D	*Cellulomonas fimi* (F5,F6)	73.0	21.7
Amylase A-180	Alkaliphilic eubacterium (F19,F20)	58.1	29.1
Endoglucanase B	*Cellulomonas fimi* (F2,F3,F4)	64.1	30.4
		66.7	34.0
		60.9	35.3
Pullulanase 1	*Clostridium thermohydrosulfuricum* (F13,F14)	37.8	107.3

[a] $d = -\ln S_{eff} \times 100$, where $S_{eff} = S_{real} - S_{rand}/S_{iden} - S_{rand}$ (Feng and Doolittle, 1987)

narily classified as G+C-poor or G+C-rich. The former include *Bacillus* and *Clostridia* and the latter *Streptomyces* and *Cellulomonas*. Accordingly, we examined the G+C content of the Fn3 segments and, as controls, the regions that flanked them, in search of indications of a change in ancestry. In most cases, the G+C content corresponded reasonably well with that of the host genome. This suggests that at least some of the Fn3-encoding segments have been in place long enough to acquire the character of the host organism. The greatest departures occur with the DNA encoding the Fn3 unit in the gram-negative *A. faecalis* (F22) and the various Fn3 units for *B. circulans* (Table 3). It should noted, however, that there is some uncertainty about the G+C content of the latter (Holt 1984).

Discussion

The sporadic occurrence of Fn3 units in bacterial glycohydrolases is underscored by the fact that the majority of related enzymes, including all those from eukaryotes, lack the domain entirely. Moreover, the Fn3 sequences are, for the most part, highly similar to each other, even though the sequences in which they are embedded show little or no resemblance.

Phylogenies based on the bacterial Fn3 sequences reveal three categories of relationship. First, clusters near the tips are frequently adjacent segments that have apparently been generated by tandem duplication. There are five of these situations, all but one of which involve greater than 50% identical sequences (Table 2). Second, there are those similar segments that are in different enzymes but from the same or closely related species. These relationships likely reflect duplication within individual genomes. A good example involves the F8–F9 and F10 domains in two different chitinases in *B. circulans*. Third, there are clear examples of very similar sequences in different enzymes in very distantly related genera. For example, the Fn3 units in the endoglucanase D of *C. fimi* are 63% identical with the units in a chitinase from *B. circulans*, and the Fn3 unit in the PHB depolymerase of the gram-negative *A. faecalis* is approximately 50% identical with three different enzymes from three different gram-positive organisms. That horizontal transfers have occurred between these bacteria seems inescapable.

A Scenario

Consider the following scenario and its consequences. A soil bacterium is transformed by the incidental acquisition of a fragment of animal DNA from a decaying organism (Fig. 11). The incorporation of a domain-encoding segment of that DNA into a gene encoding an extracellular glycohydrolase confers some natural advantage on the individual and encourages the propagation of the lineage. In time the lineage may become the major strain or even take over the entire species or genus. Moreover, recombinational events within that organism's own genome lead to tandem duplications whereby the acquired domain is present in more than one place in a single protein or even in other proteins, likely but not necessarily homologous.

But because the major genera were already in place when the initial propagation was begun, transfer to other genera, presumably rare even though vector-mediated, would likely involve only one (or a contiguous set of) domainal unit(s). As a consequence, a new "clonal" diaspora may occur in the recipient genus. In any subsequent comparison of the sequences of all these domains, they would be expected to cluster by organism and not by protein type.

Indeed, this appears to be the case for many of the Fn3 sequences observed in bacteria so far. For example, the Fn3 sequences found in two different endoglucanases (B and D) cluster by their host organism (*C. fimi*), even though the parent enzymes do not. The same is true for the Fn3 units found in the two different chitinases of *B. circulans* (Fig. 10).

Mechanistic Aspects

The most likely avenue for the initial acquisition of a eukaryotic "gene" by bacteria would seem to be transformation. Soil bacteria must encounter vast amounts of eukaryotic DNA, and even the most unlikely of opportunities should occur eventually. In contrast, the rapid spread from one bacterial group to another is more likely to occur by transduction by phages or plasmids, as is well

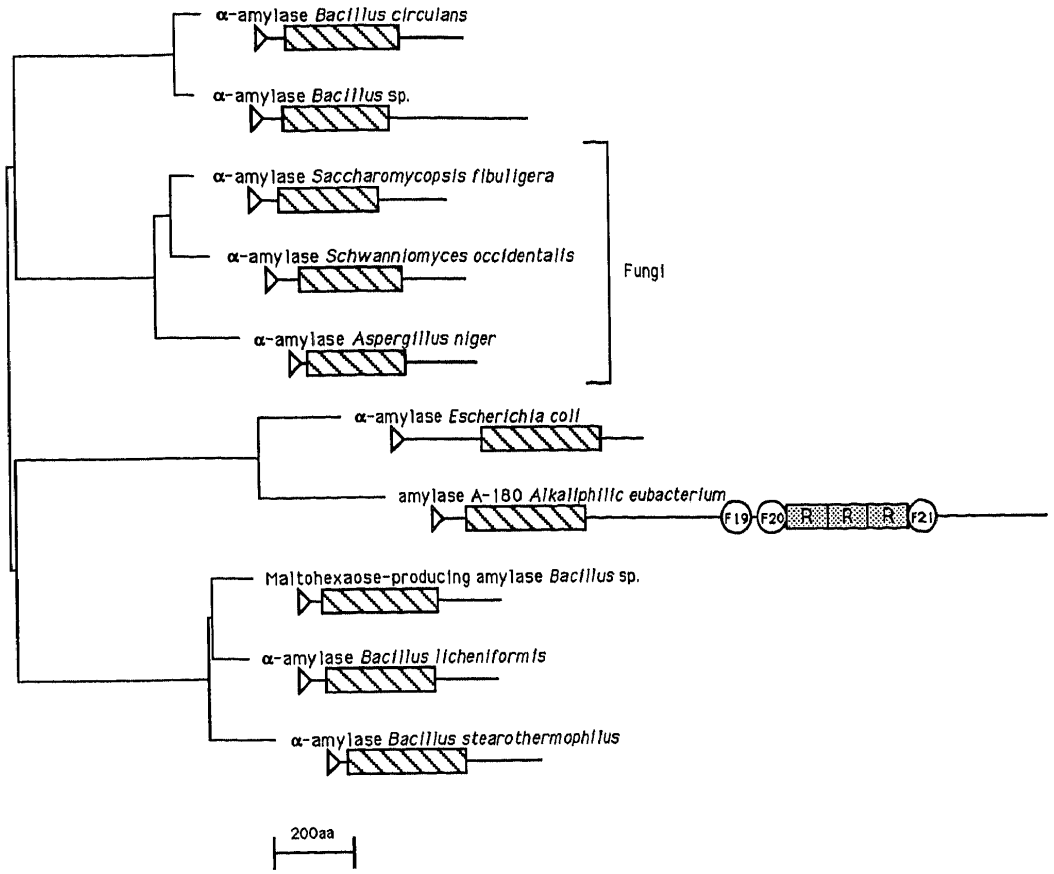

Fig. 8. Phylogeny of ten amylase sequences from assorted fungi and bacteria (Henrissat family 13), only one of which has identifiable Fn3 units (F17, F18, and F19). The *boxes* containing *R*s denote 99-residue repeats that are also observed in amylases from other *Bacilli* (Candussio et al. 1990). The latter do not contain the Fn3 units, however. PIR codes (top to bottom): ALBSK, A34648, ALBYAF, S23355, A35282, S23807, S10789, A27705, B24549, and ALBSXF.

```
F8 (chbc)     TAPSVPGNARSTGV  TANSVTLAWNASTD  NVGVTGYNV YN    GA   NLATSVTGT TAT          ISGLTAGTSYTFTIKAKDAAGNLSAASNAVTVST
F9 (chbc)     QAPTAPTNLASTAQ  TTSSITLSWTASTD  NVGVTGYDV YN    GT   ALATTVTGT TAT          ISGLAADTSYTFTVKAKDAAGNVSAASNAVSVKT
F10 (chbc)    TPPTVPAGLTSSLV  TDTSVNLTWNASTD  NVGVTGYEV YR    NG   TLVANTSTT TAV          VTGLTAGTTYVFTVKAKDAAGNLSAASTSLSVTT
F7 (cecf)     QAPSVPSGLTAGTV  TETSVALSWTASTD  NVGVTGYDV YR    NG   SKVGSSSGT TYS          DTGLTAATAYQYSVAAKDAAGNVSQRSSALSVTT
F5 (cecf)     TAPSVPTGLTAGTP  TATSVPLTWTASTDTGGSGVTGYEV YR    GS   TLVARPTGT SHT          VTGLSAATAYTFTVRAVDAAGNVSAASAPVGVTT
F6 (cecf)     TAPTAPTGLRAGTP  TASTVPLTWSASTDTGGSGVAGYEV YR    GT   TLVGTTTAT SYT          VTGLAADSAYTFSVRAKDGAGNTSAASAAVTART
F2 (cecf)     TPPTTPGTPVATGV  TTVGASLSWAASTD AGSGVAGYEL YRVQGTTQ  TLVGTTTAA AYI          LRDLTPGTAYSYVVKAKDVAGNVSAASAAVTFTT
F4 (cecf)     VAPTVPGTPVASNV  ATTGATLTWTASTDSGGSGLAGYEV LRVSGTTQ  TLVASPTTA TVA          LAGLTPATAYSYVVRAKDGAGNVSAVSSPVTFTT
F3 (cecf)     EPPTTPGTPVASAV  TSTGATLAWAPSTG  DPAVSGYDV LRVQGTTT  TVVAQTTVP TVT          LSGLTPSTAYTYAVRAKNVAGDVSALSAPVTFTT
F12 (chsp)    VAPSAPGTPTASNI  TDTSVKLSWSAATD  DKGVKNYDV LR    DG   ATVATVTGT TYT          DNGLTKGTDYSYSVKARDTGDQTGPASGSVKVTT
F11 (chso)    QPPAPPTGLRTGSV  TATSVALSWSPVTG     ATGYAV YR    DG   VKVATASGT SAT          VTGLTPDTAYAFQVAAVNGAGE SAKSATVTATT
F21 (amae)    TSPSKPTDLTAIA   TAHTVSLSWTASAD  DVEVAGYKI YR    DG   VEIGVTEST TYT          DSGLTAETTYSYMVQAYDTSNNFSALSDELTIET
F22 (dpaf)    QAGSAPTGLAVTAT  TSTSVSLSWNAVAN     ASSYGV YR    NG   SKVGSATAT AYT          DSGLIAGTTYSYTVTAVDPTAGESQPSAAVSATT
F20 (amae)    EPAEAPENLRIADI  TDTTVTINWNASNG    YVTGYEV LR    DG   VVIGETTRT TFI          DTGLDADRTYTYTIVALGDGGQKSDPSEALEVTT
F19 (amae)    EPATTPKNLSVVNV  TETTVTFEWDQSDG    YVVEYEI LR    DE   DVVASTIRT TFT          DEDLNPDTTYTYSVVAVGEGGQKSAPSEALKVTT
F13 (puct)    TAPQPITDLKAVS   GNGQVDLSWSAVDR      AVSYNI YRSTVKGGLYEKIASNVTQI TYI          DTDVTNGLKYVYSVTAVDSDGNESALSNEVEAYP
F15 (puct)    TAPQPITDLKAVS   GNGKVDLSWSVVDK      AVSYNI YRSTVKGGLYEKIASNVTQI TYT          DTEVTNGLKYVYAVTAVDNDGNESALSNEVEAYP
F17 (puct)    TAPQAPSNVVVTS   GNGKVDLSWLQSDG      ATGYNI YRSSVEGGLYEKIASNVTET TFE          DANVTNGLKYVYAISAIDELGNESGISNDAVAYP
F16 (puct)    ETPTAP VLQQPGI  ESSRVTLNWSPSAD  DVAIFGYEI YKSSSETGPFIKIATVSDSVYNYV          DTDVVNGNVYYYKVVAVDTSYN RTASNTVKATP
F14 (puct)    VEPPTALGLQQPGI  ESSRVTLNWSLSTD  NVAIYGYEI YKSLSETGPFVKIATVADTVYNYV          DTDVVNGKVYYYKVVAVDTSFN RTASNIVKATP
F18 (puct)    IKPTAP YLNQPGT  ESSRVSLTWNPSTD  NVGLYDYEI YRS  DGGTFNKIATVSNEVYNYI          DTSVINGVTYNYKVVAVDLSFN RTESNVVTIKP
F1 (gaec)     ATAQAPQKLQIPTLSYDDHSVMLVWDTPED  TSNITDYQI YQNGQLIGLASQNNDKNSPAKPYI.(21 res).VDGLKAGTDYQFTVRTVYADGTTSNDSNTVTTTT

FN10 (hum)    TVSDVPRDLEVVAA  TPTSLLISWDAPA     VTVRYYRITYGETGGNSPVQEFTVPGSKS TAT          ISGLKPGVDYTITVYAVTGRGDSPASSKPISINY
```

Fig. 9. Computer alignment (Feng and Doolittle 1987) of 22 different bacterial Fn3 sequences analyzed in the present study. See Table 1 and Fig. 1 for locations of F1–F22. The *first two letters* of the *four-letter designations* denote the type of enzyme (*ce* = cellulase; *pu* = pullulanase; *ch* = chitinase; *am* = amylase; *ga* = galacturonosidase; *dp* = depolymerase). The *third* and *fourth letters* of the *four-letter designations* denote the genus and species and can be inferred from the information in Table 1. *Emboldened* residues correspond to the Fn3 consensus observed in animal proteins, an example of which is shown at the bottom (FN10 = human fibronectin unit 10).

known for advantageous genes such as those conferring resistance to antibiotics. Indeed, conjugation between gram-positive and gram-negative bacteria is not uncommon (Heinemann 1991).

There is also the possibility of assistance by or collaboration with transposable elements. We have examined the DNA sequences of the bacterial Fn3-encoding sequences and their flanking regions for obvious repeats and inverted repeats, but we have not been able to identify any consistent pattern.

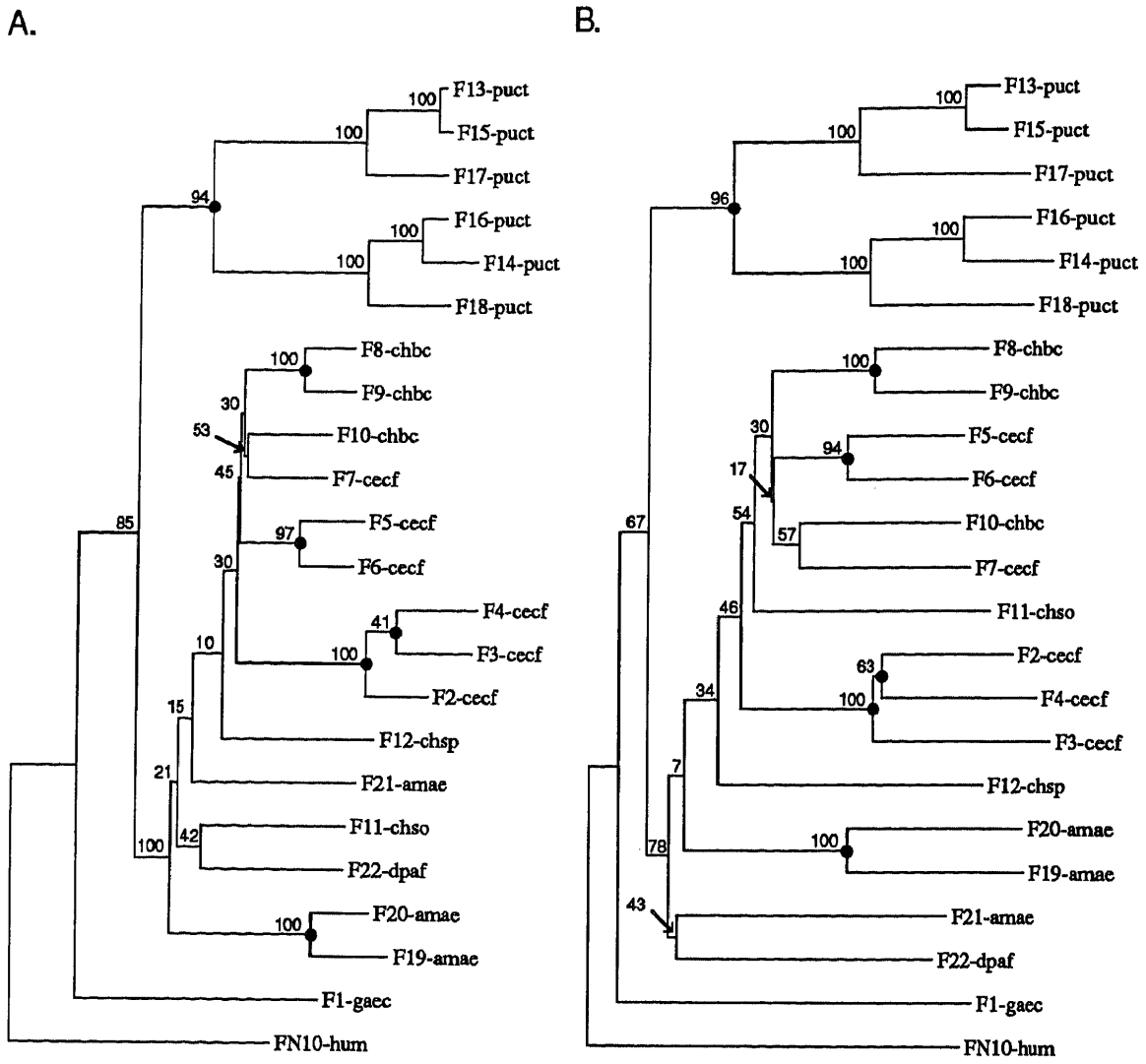

Fig. 10. **A** Phylogenetic tree of the 23 Fn3 sequences shown in Fig. 9 as determined by a distance matrix method (Feng and Doolittle 1990). The *solid circles* indicate nodes at which tandem duplications have occurred. The *numbers* indicate results of 100 bootstraps. **B** Phylogenetic tree of the same 23 Fn3 sequences shown in Fig 9 but determined by a character-based parsimony procedure (Doolittle and Feng 1990). *Solid circles* indicate tandem duplications; *numbers* are bootstrap results.

Determining the Chronology of Events

There are two ways we can get some indication of the chronology of events. For one, we can presume that the closer G+C content is to the host organism's regular genomic level, the longer an insertion has been in place, if it came from an organism with a significantly different G+C content. A second way of estimating the time has to do with tandem duplications: the more different the members of a tandem set, the longer they have been in place. By both of these criteria it would seem that the Fn3 unit must have been in *Clostridia* for a long time relative to the others. Consistent with this notion, *Clostridia* lie at or near the root of the Fn3 phylogeny (Fig. 10).

By the same criteria, the *B. circulans* units appear to be the most recently transferred, the donor organism in this case most likely being a member of the *Cellulomonas* genus (Fig. 11). Somewhat lower down in the phylogeny, and presumably longer ago, *Streptomyces* must have received an Fn3-encoding gene. The (matrix) phylogeny makes it very likely that some member of the *Streptomyces* genus subsequently passed a similar gene to the gram-negative *Alcaligenes* (Fig. 10).

More About Time

The time of the occurrence would naturally have to have been since the emergence of animals, well within the last billion years. Of course, it may have been much more recent. If we scale the rate of change relative to how fast Fn3 units are changing in vertebrates, then most of the change would have occurred during the last 500 million years. But it may have been even more recent. Certainly it must have been since the radiation of bacteria into their current major genera. Indeed, most of the exchanges appear to have occurred since the split of two clostridial species (Fig. 10). This could make the acquisition and spread very recent events, perhaps within the last ten million years.

Table 3. G + C content of Fn3-encoding bacterial DNA[a]

Fn3 Unit	Organism G + C	Fn3 G + C	Flanking G + C
Clostridia F-13	0.35–0.37	0.36	0.36
Clostridia F-14		0.38	
Clostridia F-15	0.35–0.37	0.39	0.36
Clostridia F-16		0.38	
Clostridia F-17	0.33	0.33	0.34
Clostridia F-18		0.35	
Bacillus F-8	0.36[b]	0.51	0.47
Bacillus F-9		0.54	
Bacillus F-10		0.52	0.49
Alkaliphilic F-19	0.30–0.55	0.41	0.39
Alkaliphilic F-20		0.44	
Alkaliphilic F-21		0.40	
Cellulomonas F-2	0.71–0.72	0.75	0.73
Cellulomonas F-3		0.76	
Cellulomonas F-4		0.75	
Cellulomonas F-5	0.71–0.72	0.77	0.70
Cellulomonas F-6		0.77	
Streptomyces F-11	0.69–0.78	0.76	0.71
Streptomyces F-12		0.69	0.71
Alcaligenes F-22	0.56–0.59	0.68	0.65
Erwinia F-1	0.57	0.51	0.42

[a] G + C content taken from *Bergey's Manual of Systematic Bacteriology* (Holt, 1984)
[b] Various strains range from 0.32 to 0.61

Other Interkingdom Transfers

A priori, it might be expected that the transfer of genetic material between prokaryotes and eukaryotes in the wild might be very uncommon. In a recent minireview, Smith et al. (1992) concluded that, although some past claims of such transfer are likely mistaken, some others seem to have stood the test of time. Nonetheless, the natural acquisition of eukaryotic genetic material by bacteria seems so remarkable that an extraordinary amount of effort ought to be invested to obtain proof. In the case of the proposed Fn3 transfer, the following evidence supports the case:

1. The distribution of the domain is so far restricted to animals and soil bacteria.
2. The bacterial sequences are more similar to some animal Fn3 sequences than are some animal Fn3 sequences to each other.
3. Even with allowance for the smaller number of occurrences identified so far, the diversity of the bacterial Fn3 sequences is significantly less than that observed among animals.
4. In every case of a bacterial occurrence, homologous enzymes have been found that lack the Fn3 domain.
5. Phylogenetic trees based on the bacterial Fn3 sequences are only consistent with a series of intergeneric transfers. The trees also reveal a number of recent tandem duplications.

Other Reported Transfers Among Bacteria

Intergeneric transfers among bacteria are not uncommon (see, e.g., Heinemann 1991; or Parkinson and Kofoid 1992), and indeed the genomes of many bacteria have a mosaic character reflective of past horizontal transfers (Medigue et al. 1991; Smith 1992). Soil bacteria may be especially susceptible to such exchanges. Whittam (1992) notes that, whereas enteric bacteria tend to be clonal in their evolution, soil bacteria have heterogeneous genomes reflective of great amounts of recombination.

These exchanges are not limited to closely related bacteria. Mazodier and Davies (1991) have reviewed exchanges between distantly related genera, including transfers between gram-positive and gram-negative bacteria (Trieu-Cuot et al. 1985; Brisson-Noel et al. 1988). In a particularly relevant case, Guiseppe et al. (1991) have reported a likely transfer of a cellulase gene from *Erwinia chrysanthemi* to *Cellulomonas uda.*

Other Shuffled Domains in Bacteria

This is not the first report of domains or modules being shuffled around in bacterial proteins. The A subunit of the *E. coli* Uvr DNA repair system has been reported to have a modular construction, some of the domains of which occur in active-transport proteins (Doolittle et al. 1986). Even more pertinent to the discussion at hand, carbohydrate-binding domains have been found in scattered occurrences in both bacterial and eukaryotic carbohydrases of the same ilk as contain Fn3 domains (Gilkes et al. 1991; Meinke et al. 1991a). Recently, also, an accessory domain of unknown function has been found in various locations in a family of enzymes that includes sialidases and galactose oxidase (Bork and Doolittle 1994). Clearly, protein domains can be evolutionarily mobile in bacterial genomes. One major driving force is likely their ability to stand on their own in a structural sense.

Possible Functions of Fn3 Domains

There must be some natural advantage to the presence of the Fn3 domain in so many extracellular carbohydrate-splitting enzymes. One possibility is that they are involved in binding to the cellulosome, a unique high-molecular-weight complex to which cellulases and other glycohydrolases bind (Salamitou et al. 1992; Fujino et al. 1993; Wang et al. 1993). Another is that they may bind polysaccharides directly. It is known that one or more of the Fn3 units in fibronectin is involved in binding heparin (Ingham et al. 1993). In any event, the function must be accessory and not essential, given the absence of the domain in so many homologues.

In conclusion, the degree of similarity of Fn3 se-

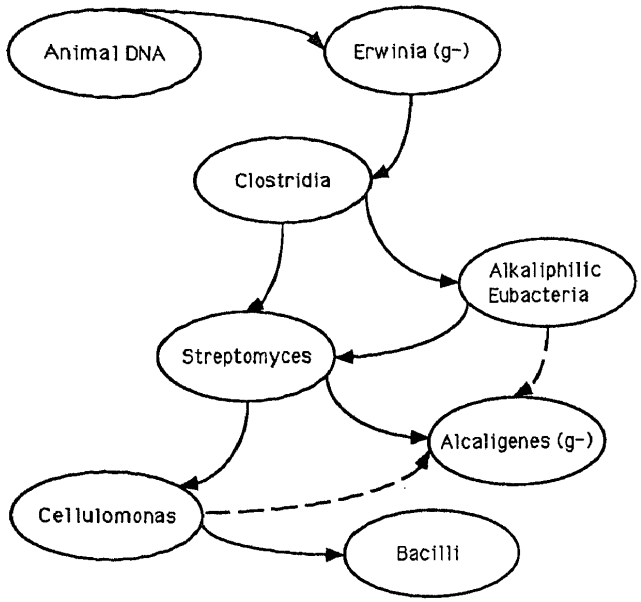

Fig. 11. Possible scenario of intergroup transfer of Fn3 units as suggested by the phylogeny shown in Fig. 10A,B. Other schemes are possible, and only the most recent transfers can be proposed with any certainty.

quences in bacteria and animals, coupled with the absence of such sequences in protists, fungi, and plants, strongly suggests interkingdom transfer. Fn3 units occur sporadically among bacterial glycohydrolases but never in the corresponding eukaryotic enzymes. They are a monophyletic group emerging from one set of animal sequences. In this regard, the most similar animal sequence to the bacterial set is an Fn3 unit found in the human leucocyte antigen designated LAR, its sequence being about 30% identical to the bacterial sequences (Bork and Doolittle 1992). Phylogenetic trees based on the bacterial Fn3 sequences are radically different from trees based on adjacent sequences. Of the 25 known occurrences, 23 are in gram-positive bacteria. One of the gram-negative occurrences involves an organism previously implicated in the transfer of a cellulase to a gram-positive *Cellulomonas* from the gram-negative *Erwinia chrysanthemi.* In the other case, the G+C content of the Fn3 encoding region from the gram-negative organism (*A. faecalis*) is closer to that of the G+C-rich gram-positive potential donor (*S. plicatus*). Its amino acid sequence is also most similar to an Fn3 unit found in a gram-positive organism. The correlation with extracellular glycohydrolases suggests the possibility of these domains binding either to cellulosomes or to polysaccharides. Like previously noted carbohydrate-binding domains in these same enzymes, the evolutionary history is consistent with module shuffling.

Acknowledgements. We are please to acknowledge the help and encouragement of our colleague Da Fei Feng, and in particular his assistance in performing the bootstrap analyses on phylogenetic trees. We are also grateful to Karen Anderson for help with the manuscript. This work was supported by N.I.H. grant HL-26873.

References

Baron M, Main AL, Driscoll PC, Mardon HJ, Boyd J, Campbell ID (1992) ^{1}H NMR assignment and secondary structure of the cell adhesion type III module of fibronectin. Biochemistry 31:2068–2073

Bazan JF (1990) Structural design and molecular evolution of a cytokine receptor superfamily. Proc Natl Acad Sci USA 87:6934–6938

Blaak H, Schnellmann J, Walter S, Henrissat B, Schrempf H (1993) Characteristics of an exochitinase from *Streptomyces olivaceoviridis,* its corresponding gene, putative protein domains and relationship to other chitinases. Eur J Biochem 214:659–669

Bork P, Doolittle RF (1992) Proposed acquisition of an animal protein domain by bacteria. Proc Natl Acad Sci USA 89:8990–8994

Bork P, Doolittle RF (1994) The *Drosophila* kelch motif is derived from a common enzyme fold. J Mol Biol 236:1277–1288

Brisson-Noël A, Arthur M, Courvalin P (1988) Evidence for natural gene transfer from gram-positive cocci to *Escherichia coli.* J Bacteriol 170:1739–1745

Burchhardt G, Wienecke A, Bahl H (1991) Isolation of the pullulanase gene from *Clostridium thermosulfurogenes* (DM 3896) and its expression in *Escherichia coli.* Curr Microbiol 22:91–95

Candussio A, Schmid G, Böck A (1990) Biochemical and genetic analysis of a maltopentaose-producing amylase from an alkaliphilic Gram-positive bacterium. Eur J Biochem 191:177–185

deVos AM, Ultsch M, Kossiakoff AA (1992) Human growth hormone and extracellular domain of its receptor—crystal structure of the complex. Science 255:306–312

Doolittle RF (1987) Of URFs and ORFs. A primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA

Doolittle RF, Feng D-F (1990) Nearest neighbor procedure for relating progressively aligned amino acid sequences. In: Doolittle RF (ed) Molecular evolution: computer analysis of protein and nucleic acid sequences. Academic Press, New York, pp 659–669

Doolittle RF, Johnson MS, Husain I, Van Houten B, Thomas DC, Sancar A (1986) Domainal evolution of a prokaryotic DNA-repair protein and its relationship to active-transport proteins. Nature 323: 451–453

Feng D-F, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25:351–360

Feng D-F, Doolittle RF (1990) Progressive alignment and phylogenetic tree construction of protein sequences. In: Doolittle RF (ed) Molecular evolution: computer analysis of protein and nucleic acid sequences. Academic Press, New York, pp 375–387

Fujii T, Miyashita K (1993) Multiple domain structure in a chitinase gene (*chiC*) of *Streptomyces lividans.* J Gen Microbiol 139:677–686

Fujino T, Béguin P, Aubert J-P (1993) Organization of a *Clostridium thermocellum* gene cluster encoding the cellulosomal scaffolding protein CipA and a protein possibly involved in attachment of the cellulosome to the cell surface. J Bacteriol 175:1891–1899

Gilkes NR, Henrissat B, Kilburn DG, Miller RC Jr, Warren RAJ (1991) Domains in microbial β-1,4-glycanases: sequence conservation, function, and enzyme families. Microbiol Rev 55:303–315

Gräbnitz F, Rücknagel KP, Seiss M, Staudenbauer WL (1989) Nucleotide sequence of the *Clostridium thermocellum bglB* gene encoding thermostable β-glucosidase B: homology to fungal β-glucosidases. Mol Gen Genet 217:70–76

Guiseppi A, Aymeric JL, Cami B, Barras F, Creuzet N (1991) Sequence analysis of the cellulase-encoding *celY* gene of *Erwinia chrysanthemi:* a possible case of interspecies gene transfer. Gene 106:109–114

Hansen CK (1992) Fibronectin type III-like sequences and a new domain type in prokaryotic depolymerases with insoluble substrates. FEBS Lett 305:91–96

He SY, Collmer A (1990) Molecular cloning, nucleotide sequence, and marker exchange mutagenesis of the exo-poly-α-D-galacturonosidase-encoding *pehX* gene of *Erwinia chrysanthemi* EC16. J Bacteriol 172:4988–4995

Heinemann JA (1991) Genetics of gene transfer between species. Trends Genet 7:181–185

Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 280:309–316

Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J (1993) 293:781–788

Holt JG, ed (1984) Bergey's manual of systematic bacteriology, volumes 1–4. Williams & Wilkins, Baltimore, MD

Ingham KC, Brew SA, Migliorini MM, Busby TF (1993) Binding of heparin by type III domains and peptides from the carboxy terminal Hep-2 region of fibronectin. Biochemistry 32:12548–12553

Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP (1992) Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethyionyl protein. Science 258:987–991

Mathupala S, Saha BC, Zeikus JG (1990) Substrate competition and specificity at the active site of amylopullulanase from *Clostridium thermohydrosulfuricum.* Biochem Biophys Res Commun 166:126–132

Mazodier P, Davies J (1991) Gene transfer between distantly related bacteria. Annu Rev Genet 25:147–171

Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coil* speciation. J Mol Biol 222:851–856

Meinke A, Gilkes NR, Kilburn DG, Miller RC Jr, Warren RAJ (1991a) Multiple domains in endoglucanase B (CenB) from *Cellulomonas fimi:* functions and relatedness to domains in other polypeptides. J Bacteriol 173:7126–7135

Meinke A, Braun C, Gilkes NR, Kilburn DG, Miller RC Jr, Warren RAJ (1991b) Unusual sequence organization in CenB, an inverting endoglucanase from *Cellulomonas fimi.* J Bacteriol 173:308–314

Meinke A, Gilkes NR, Kilburn DG, Miller RC Jr, Warren RAJ (1993) Cellulose-binding polypeptides from *Cellulomonas fimi:* endoglucanase D (CenD), a family A β-1,4-glucanase. J Bacteriol 175: 1910–1918

Melasniemi H, Paloheimo M, Hemiö L (1990) Nucleotide sequence of the α-amylase-pullulanase gene from *Clostridium thermohydrosulfuricum.* J Gen Microbiol 136:447–454

Parkinson JS, Kofoid EC (1992) Communication modules in bacterial signaling proteins. Ann Rev Genet 26:71–112

Patthy L (1990) Homology of a domain of the growth hormone/prolactin receptor family with type III modules of fibronectin. Cell 61:13–14

Robbins PW, Overbye K, Albright C, Benfield B, Pero J (1992) Cloning and high-level expression of chitinase-encoding gene of *Streptomyces plicatus.* Gene 111:69–76

Rohde K, Bork P (1993) A fast, sensitive pattern-matching approach for protein sequences. CABIOS 9:183–189

Saito T, Suzuki K, Yamamoto J, Fukui T, Miwa K, Tomita K, Nakanishi S, Odani S, Suzuki J-I, Ishikawa K (1989) Cloning, nucleotide sequence, and expression in *Escherichia coli* of the gene for poly(3-hydroxybutyrate) depolymerase from *Alcaligenes faecalis.* J Bacteriol 171:184–189

Salamitou, Tokatlidis K, Béguin, Aubert J-P (1992) Involvement of separate domains of the cellulosomal protein S1 of *Clostridium thermocellum* in binding to cellulose and in anchoring of catalytic subunits to the cellulosome. FEBS Lett 304:89–92

Siggens KW (1987) Molecular cloning and characterization of the beta-amylase gene from *Bacillus circulans.* Mol Microbiol 1:86–91

Smith JM (1992) Analyzing the mosaic structure of genes. J Mol Evol 34:126–129

Smith MW, Feng D-F, Doolittle RF (1992) Evolution by acquisition: the case for horizontal gene transfers. TIBS 17:489–493

Trieu-Cuot P, Gerbaud G, Lambert T, Courvalin P (1985) *In vivo* transfer of genetic information between gram-positive and gram-negative bacteria. EMBO J 4:3585–3587

Tsujibo H, Endo H, Minoura K, Miyamoto K, Inamori Y (1993) Cloning and sequence analysis of the gene encoding a thermostable chitinase from *Streptomyces thermoviolaceus* OPC-520. Gene 134: 113–117

Uozumi N, Sakurai K, Sasaki T, Takekawa S, Yamagata H, Tsukagoshi N, Udaka S (1989) A single gene directs synthesis of a precursor protein with β- and α-amylase activities in *Bacillus polymyxa.* J Bacteriol 171:375–382

Wang WK, Kruss K, Wu JHD (1993) Cloning and DNA sequence of the gene coding for *Clostridium thermocellum* cellulase S_S (CelS), a major cellulosome component. J Bacteriol 175:1293–1302

Watanabe T, Suzuki K, Oyanagi W, Ohnishi K, Tanaka H (1990) Gene cloning of chitinase A1 from *Bacillus circulans* WL-12 revealed its evolutionary relationship to *Serratia* chitinase and to the type III homology units of fibronectin. J Biol Chem 265:15659–15665

Watanabe T, Oyanagi W, Suzuki K, Ohnishi K, Tanaka H (1992) Structure of the gene encoding chitinase D of *Bacillus circulans* WL-12 and possible homology of the enzyme to other prokaryotic chitinases and class III plant chitinases. J Bacteriol 174:408–414

Whittam TS (1992) Sex in the soil. Curr Biol 2:676–678