Proposed Acquisition of an Animal Protein Domain by Bacteria

Peer Bork; Russell F. Doolittle

Proceedings of the National Academy of Sciences of the United States of America, Vol. 89, No. 19 (Oct. 1, 1992), 8990-8994.

Stable URL:
http://links.jstor.org/sici?sici=0027-8424%2819921001%2989%3A19%3C8990%3APAOAAP%3E2.0.CO%3B2-B

Proceedings of the National Academy of Sciences of the United States of America is currently published by National Academy of Sciences.

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 8990–8994, October 1992
Evolution

Proposed acquisition of an animal protein domain by bacteria

(horizontal gene transfer/mobile domains/fibronectin type III)

PEER BORK* AND RUSSELL F. DOOLITTLE†

*European Molecular Biology Laboratory, Meyerhofstrasse 1, 6900 Heidelberg, Federal Republic of Germany; Max-Delbrück Centre for Molecular Medicine, Robert Rössle-Strasse 10, 1115 Berlin-Buch, Federal Republic of Germany; and †Center for Molecular Genetics, University of California, San Diego, La Jolla, CA 92093-0634

Contributed by Russell F. Doolittle, June 24, 1992

ABSTRACT **A systematic screen of a protein sequence data base confirms that the fibronectin type III (Fn3) domain is widely distributed among animal proteins and occurs also in several bacterial carbohydrate-splitting enzymes. The motif has yet to be identified in proteins from plants or fungi. All indications are that the bacterial sequences are much too similar to the animal type to be the result of conventional vertical descent. Rather, it is likely that the bacterial units were initially acquired from an animal source and are being spread further by horizontal transfers between distantly related bacteria.**

The fibronectin type III (Fn3) module is widely spread among contemporary animal proteins, having so far been identified in more than 50 different proteins, not counting species redundancies (1–3). These include many extracellular proteins (for example, fibronectin itself), some intracellular proteins (titin, twitchin), and the extracellular domains of many kinds of membrane-receptor protein (Table 1). Depending on the protein, the module may occur alone or in numerous nonidentical forms. The motif has a characteristic sequence of 90–100 amino acid residues typified by a pattern of well-defined structural features (4, 5).

Recently, sequences resembling the Fn3 domain were reported in a bacterial chitinase (6) and two cellulases (7), leading to the supposition that the domain originated before the divergence of prokaryotes and eukaryotes (6). If the supposition were true, it would be expected that the module would occur in some proteins of most contemporary organisms. We now report that a systematic searching regimen has turned up additional occurrences of the Fn3 unit in a wide variety of animal proteins. The survey did not uncover any occurrences in plants or fungi, but it did retrieve four more bacterial enzymes (for a total of seven). Surprisingly, the bacterial Fn3 sequences are no more different from the animal sequences than the latter are from each other. Our analysis of the sequence relationships leads us to the conclusion that the bacterial occurrences are likely the result of a single gene having been acquired from an animal source long after the divergence of prokaryotes and eukaryotes.

METHODS

Data Base Screening. The screening process for potential Fn3 domains took two forms. First, 79 sequences labeled Fn3 in the Swiss-Prot sequence data base (8) were compiled as a learning set and properly aligned, and a consensus pattern was established by a previously described scheme (9). The screening was conducted iteratively, new candidates being added to the learning set progressively to increase the sensitivity (10). Deviations from the pattern (mismatches) were penalized as a function of the degree of conservation at a given position, as were deletions and insertions. The behavior of the random background of unrelated sequences was

Table 1. Summary of known proteins containing Fn3 modules

Category	Proteins (organisms)
A. Receptor protein kinases	Sevenless* (fruit fly), Ros1* (human), eph (human), eck (human), elk (human), axl (human), hek (human), cek5 (chicken), insulin-like growth factor receptor* (human), insulin receptor* (human, rat, mouse), insulin-related receptor* (human, frog, fruit fly)
B. Receptor protein phosphatases	rptp-β1 (human), rptp-δ1 (human), rptp-My (human), dptp* (fruit fly), dptp 99A (fruit fly), dptp 10DA* (fruit fly), leukocyte-related antigen (human, fruit fly), CD45 antigen† (human, mouse, rat)
C. Neural adhesion molecules	Contactin* (chicken), caml L1 (mouse), F1* (chicken), neural cell adhesion molecule* (human, rat, mouse, chicken, frog), fasciclin II (grasshopper), neuroglian (fruit fly), Tag1* (rat)
D. Adhesive matrix proteins	Fibrin–fibrinogen-related protein (human), fibronectin (human, bovine, rat, chicken), tenascin (human, mouse, chicken), undulin (human), collagen VI-α3 (human, chicken), collagen VII† (human), collagen XII (human)
E. Cytokine receptor-like	Growth hormone receptor (human, pig, rat, rabbit), erythropoietin receptor (human, mouse), prolactin receptor (human, rat, mouse, rabbit), granulocyte/macrophage colony-stimulating factor receptor β chain (human), interleukin 6 receptor subunit gp130 (human), granulocyte colony-stimulating factor receptor* (human), natural killer cell colony-stimulating factor p40 subunit (human), interleukin 2 receptor β chain (human, mouse), interleukin 3 receptor (mouse), interleukin 4 receptor (mouse), interleukin 5 receptor (mouse), interleukin 6 receptor (human), interleukin 7 receptor (human)
F. Cytoplasmic muscle proteins	Myosin light chain kinase (chicken), muscle protein C (chicken), twitchin (nematode), twitchin-related (fruit fly), titin (rabbit)
G. Unclassified	Integrin β4 chain (human), Kallmann syndrome protein* (human), 45-kDa antigen† (tapeworm), oncogen b† (tapeworm), oncogene a† (tapeworm)
H. Prokaryotes	Amylase/pullulanase† (*Clostridium thermohydrosulfuricum*), chitinase (*Bacillus circulans*), amylase-180* (Gram-positive alkaliphilic), depolymerase† (*Alcaligenes faecalis*), galacturonosidase† (*Erwinia chrysanthemi*), cellulase (*Cellulomonas flavigena*), endoglucanase (*Cellulomonas fimi*)

*Previous reports claimed either too many or too few Fn3 repeats according to our criteria.
†Repeats detected by our pattern search but to our knowledge not recognized by others.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: Fn3, fibronectin type III.

also taken into account. A corroborative screening was conducted with the PROFILE method (11). Neither method listed every authentic Fn3 unit ahead of all false positives, but a careful comparison of the two sets of results (Fig. 1) yielded a high-confidence list. Further verification was obtained by searching candidate sequences with a modified version of FASTA (12) in which sequences below a given threshold are filtered off (13). We also included a number of Fn3 sequences taken directly from the literature before their inclusion in Swiss-Prot.

Sequence Alignment. Sequences were aligned by the progressive method (14) after the initial determination of approximate relationships by a series of binary comparisons of the Needleman–Wunsch type (15). Sequences were entered in an order in accord with their approximate similarity, and multiple alignments were constructed.

Rates of Change. Approximate rates of sequence change were estimated by comparisons of the same protein from different species and depended on generally accepted divergence times for the major groups of vertebrate and invertebrate animals. The results were viewed in the light of a set of comparison proteins that do not contain Fn3 sequences (16).

Phylogenetic Trees. Phylogenetic trees were constructed from alignments by two independent procedures: the first a distance matrix method (17) that follows closely upon that of Fitch and Margoliash (18), the other a four-taxon character analysis referred to as PAPA (19) that is based on the analytical approach initially used by Cavender (20).

RESULTS

All told, more than 300 sequences were assembled that met the criteria for the Fn3 motif. These were found in 67 different proteins, not counting species redundancies (Table 1). Of these, 7 were extracellular enzymes excreted by assorted bacteria. The remaining 60 were all from animal proteins, including extracellular, intracellular, and membrane-spanning types. None were found in any plant or fungal sequences.

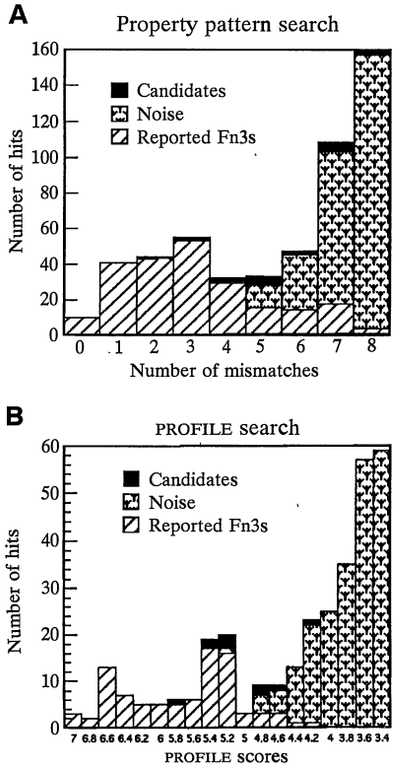

FIG. 1. Retrieval of Fn3 sequences from the Swiss-Prot data bank (December 1991) by the property pattern search (PROPAT) (*A*) and the PROFILE (*B*) method. The discrepancies in total number of hits are mostly due to the fact that the PROFILE program counted each protein only once, whereas PROPAT counted every individual Fn3 unit.

Rates of Change. Rates of change were estimated for those animal Fn3 sequences for which representatives are available from various species (Table 2). In both fibronectins and neural cell adhesion molecules (N-CAM) the Fn3 units are changing at an intermediate rate as judged by comparison with a representative set of 30 other proteins found in vertebrate animals (16). Cruder estimates were made for the Fn3 sequences found in growth hormone receptor and insulin receptor, in which cases comparisons were necessarily limited to mammals. More distant comparisons involving vertebrate and invertebrate animals were possible in two instances. Thus, titin, found in animal muscles, contains Fn3 repeats that are 40–45% identical with those found in the protein twitchin from nematodes and a twitchin-related protein from fruit flies. Similarly, the leukocyte-related antigen (LAR) proteins from human and fruit fly also contain Fn3 repeats whose resemblances are in the 40% range (Table 2).

All in all, fibronectin units are changing more slowly than some other animal proteins, including globins and albumins, but somewhat faster than mainstream metabolic enzymes such as enolase (Table 2). Given this moderate rate of change, and assuming constancy of change and vertical descent in even earlier times, then it ought to be possible to recognize Fn3 units wherever they may occur among the eukaryotes and perhaps in prokaryotes, although in the latter case it would be expected that the sequence resemblance would be marginal at best.

In fact, the prokaryotic Fn3 sequences identified so far are remarkably similar to the animal type, being up to 38% identical with some of the units found in animal proteins. Further, the units found in bacteria are all quite similar to each other and much more so than are the proteins in which they are embedded. Indeed, of the six enzymes depicted in Fig. 2, only the two amylases have detectable sequence similarity with each other, exclusive of the Fn3 segments. In contrast, some of these enzymes have recognizable counterparts in other bacteria that do not contain the Fn3 units. For example, the high molecular weight chitinase from *B. circulans* can be aligned with that from *Serratia marcescens*, which is devoid of Fn3 sequences (6). Similarly, the endoglucanase (cellulase) from *Cellulomonas fimi* can be aligned over parts of its sequence with cellulases from a wide variety of

Table 2. Percent identities of orthologously* related Fn3 units

Protein	Rodent/ rodent	Mammal/ mammal	Mammal/ bird	Mammal/ amphibian	Vertebrate/ invertebrate
Fn3 units					
Fibronectin	—	91.7	85.0	—	—
N-CAM	92.9	92.3	83.3	68.6	—
Insulin receptor	97.6	93.0	—	—	—
GH receptor	—	86.1	—	—	—
Titin–twitchin	—	—	—	—	(46.7)†
LAR	—	—	—	—	(39.5)†
Comparison proteins					
Enolase	—	94.0	92.2	87.0	72.1
Cytochrome *c*	100.0	91.3	87.5	82.7	82.0
β-/γ-Fibrinogen	—	87.0	81.6	70.9	(33.0)
Albumin	90.0	72.1	—	38.4	—

N-CAM, neural cell adhesion molecule; GH, growth hormone; LAR, leukocyte-related antigen.

**Orthologous*, as defined by Fitch (21), indicates direct lineage, in contrast to *paralogous*, which implies descent after gene duplication.

†May be paralogous.

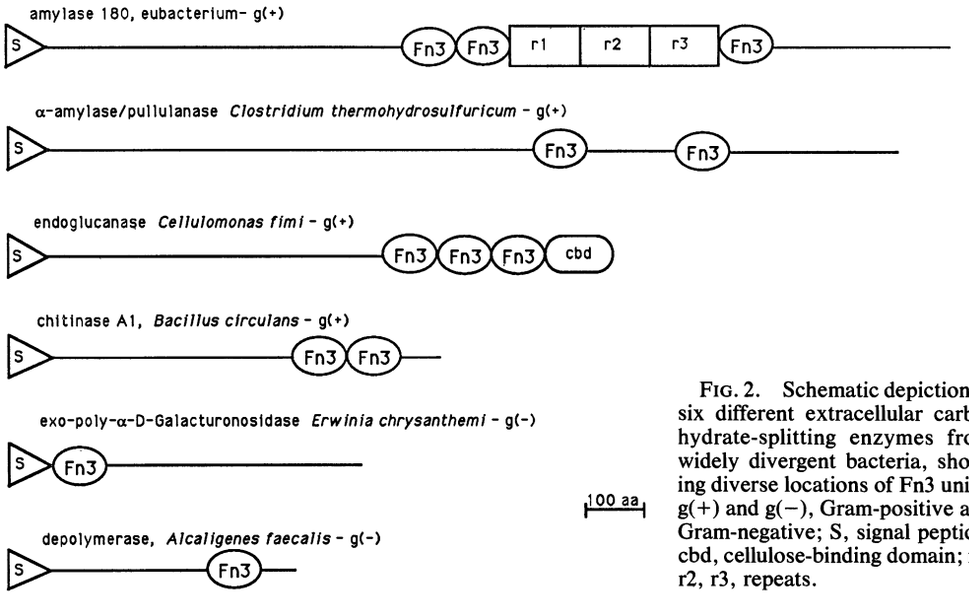

FIG. 2. Schematic depiction of six different extracellular carbohydrate-splitting enzymes from widely divergent bacteria, showing diverse locations of Fn3 units. g(+) and g(−), Gram-positive and Gram-negative; S, signal peptide; cbd, cellulose-binding domain; r1, r2, r3, repeats.

organisms—including plants—that do not have Fn3 modules (22).

Phylogenetic Trees. A computer-determined alignment was prepared of 39 Fn3 sequences, including representatives of all the major categories from animals (Table 1) and all 13 bacterial sequences (Fig. 3). Widely divergent animal sequences were chosen purposely in order that the deepest roots of the phylogeny would be apparent. The average pairwise resemblance for the 26 animal sequences used in the alignment was 20% identity.

The resulting phylogenetic trees have all the bacterial entries clustered as a single group and the animal sequences, for the most part, clustered according to their previously assigned categories (Table 1). The most reassuring aspect is that the trees made by two fundamentally different methods, albeit from the same alignment, are so similar (Fig. 4). The phylogenetic trees also showed that there has been a tendency for the Fn3 units from a given protein to cluster together. This was especially evident for units that are adjacent in the bacterial enzymes (Fig. 2), an indication of recent tandem duplication.

Because some of the pairs described in Table 2 were included in the alignment, it was possible to put an approximate time dimension on the trees. By extrapolation, the

```
D  UNH1    AITGPPTELITSEVT  ARSFMVNWT      HA P GNVEKYRVVYY P TRG  GKP  DE VV   VDGTSST  V            LKNLMSLTEYQIAVFAIYAHTAS EGLRGTETT
D  UNH2    LALPMASDLLLYDVT  ENSMRVKWD      AV P GASG YLILYA PLTEGLAGDE  KE MK   IGETHTD  IE           LSGLLPNTEYTVTVYAMFGEEAS DPVTGQETT
D  TNH1    SEVSPPKDLVVTEVT  EETVNLAWD      NE M RVTE YLVVYT PTHEG  GLE  MQ FR   VPGDQTS  TI           IQELEPGVEYFIRVFAILENKKS IPVSARVAT
D  TNH7    GETPNLGEVVVAEVG  WDALKLNWT      AP E GAYE YFFIQV QEADT VEAA  QN LT   VPGGLRS  TD           LPGLKAATHYTITIRGVTQDFST TPLSVEVLT
D  FNH9    TGLDSPTGIDFSDIT  ANSFTVHWI      AP R ATITGYRIRHH PEHFS  GRP  RE DR   VPHSRNS  IT           LTNLTPGTEYVVSIVALNGREES PLLIGQQST
D  FNH10   TVSDVPRDLEVVAAT  PTSLLISWD      AP A VTVRYYRITYG ETGGN  SPV  QE FT   VPGSKST  AT           ISGLKPGVDYTITVYAVTGRGDS PASSKPISI
D  FNH1    SSGPVEVFITETPSQ  PNSHPIQWN      APQP SHISKYILRWR PKNSV  GRW  KE AT   IPGHLNS  YT           IKGLKPGVVYEGQLISIQQYGHQ EVTRFDFTT
D  TNH2    TYLPAPEGLKFKSIK  ETSVEVEWD      PL D IAFETWEIIFR NMNK   EDE  GE IT   KSLRRPE  TS           YRQTGLAQEYEISLHIVKNNTRG PGLKRVTTT
F  TWD2    KPSPPNGPLQISDVH  KEGCHLKWK      R  P TDDGGTPIEYP QIDKLEPETGC WI PS   CRSTEPQ  VD           VTGLSPGNGYKPRVSAVNAEGES QPLVGDESI
F  TWE4    GVPGKPGRPEITDFD  ADRIDIAWE      PP H KDGGAPIEEYI VEVRDPDTKEW KE VK    RVPDTN  AS           ISGLKEGKEYQFRVRAVNKAGPG QPSEPSEKQ
F  TIR3    KPGIPTGPIKFDEVT  AEAITLKWG      PP K DDGGSEITNYI LEKRDSVMMKW VT CA   SAVQKTT  FR           VTRLHEGMEYTFRVSAENKYGVG EGLKSEPIV
F  TWD1    GVPTKPGVPEPTDWT  ANKVELAWP      EP A SDGGSPIQGYI VEVKDKYSPLWEKA LD   TNSPTPT  AT           VQGLIEGNAYQFRVVAVNIGGLS EPSDPSKIFT
F  MKC1    KPDPPAGTPCASDIR  SSSLTLSWY      GS S YDGGSAVQSYT VEIWNSVDNKW TD LT    TCRSTS  FN           VQDLQADREYKFRVRAANVYGIS EPSQESEVV
F  PCC2    APTSEPTHVVLEDVT  DTTATIKWR      PP E RIGAGGVDGYL VEWCREGSNEWVAA NT   ELVERCG  LT           ARGLPTGERLLFRVISVNMAGKS PPATMAQPVT
E  I2M1    RLVAPHSLQVL HID  TQRCNISWK      VSQV SHYIEPYL EF EARRRLLGHSWEDA SV   LSLKQRQQWLF           LEMLIPSTSYEVQVRVKAQRNNT GTWSPWSQPL
E  I7H1    KPEAPFDLSVIYREG  ANDFVVTFN      TSHLQKKYVKVLMHDV AYRQEKDENKWTHV NL   SSTKLT   LL           QRKLQPAAMYEIKVRSIPDHYFK GFWSEWSPSY
B  LRH1    ALPKPPIDLVVTETT  ATSVTLTWD           SGNSEPVTYYG IQYRAAGTEGPFQE VD   GVATTR   YS           IGGLSPFSEYAFRVLAVNSIGRG PPSEAVRART
E  GRBF    PPIGLNWTLLNVSLTGIHADIQVRWE      PPPNADVQKGWIVLEYE LQYKEVNETQW  K MM   DPVLSTS   VP           VYSLRLDKEYEVRVRSRQRSSEKYGEFSEVLYVT
B  CDH1    GSPGEPQIIFCRSEA  AHQGVITWN      PPQ   RSFHNFTLCYI KETEK       DC LN   LDKNLIK   YD           LQNLKPYTKYVLSLHAYIIAKVQRNNGSAAMCHF
C  NCHU    VQADTPSSPSIDQVEPYSSTAQVQFD      EPE   ATGGVPILKYK AEWRAVGEEVWHSKWYD   AKEASMEGIVT          IVGLKPETTYAVRLAALNGKGLGEISAASEFKT
G  IBH2    EVPSEPGRLAFNVVS  STVTQLSWA      EPAETNGEITAYEVCYG LVNDDNRPIGPMKKVLV   DNPKNRM   LL           IENLRESQPYRYTVKARNGAGWGPEREAIINLA
C  CMM5    ALFGKPDFGNISATAG ENYSVVSW        VPRKGQCNFRFHILFK ALPEGKVSPDHQPQPQY   VSYNQSS   YT           QWNLQPDTKYEIHLIKEKVLLHHLDVKTNGTGPV
G  45TW    EYPSLRDIFAWEPPT  SNSIGLTWQ      RHAFPGVEREVLTLKAV PTSEPNNTKT     AY   AKLGSGK   VT           LDGLKPNATYLVTATANISGDTILVLSNTFHTLA
G  KLH2    APPANLRLANSTVNSDGSVTVTIVWD      LPEEPDIPVHHYKVFWSWMVSSKSLVPTKKKRRKT   TDGFQNS   VI           LEKLQPDCDYVVKLQAITYWGQTRLKSAVKVSL
H  EPH1    GPPSAPRNLSFSAS   GTQLSLRWEPPADTGGRQDVRYSV RCSQCQGTAQDGGPCQPCGVGVHFSP  GARGLTTPAVH          VNGLEPYANYTFNVEAQNGVSGLGSSGHASTSVS
H  AXH1    VLPQQPRNLHLVSRQ  PTELEVAWTP     GLSGI YPLTHCTLQAVLSDDGMGIQAGEPDPPEEPLTSQASVPPHQLR          LGSLHPHTPYHIRVACT SSQGPSSWTH WLPV

H  DPA1    QAGSAPTGLAVTATT  STSVSLSWN           AVAN     ASSYGVYRNGS       K    VGSATAT AYT           DSGLIAGTTYSYTVTAVDPTAGESQPSAAVSATT
H  AME3    TSPSKPTDLTAIATA  HT VSLSWT           ASADD  VEVAGYKIYRDGV        E    IGVTEST TYT           DSGLTAETTYSYMVQAYDTSNNFSALSDELTIET
H  CHB2    QAPTAPTNLASTAQT  TSSITLSWT           ASTDN  VGVTGYDVYNGTA        L    ATTVTGT TAT           ISGLAADTSYTFTVKAKDAAGNVSAASNAVSVKT
H  CHB1    TAPSVPGNARSTGVT  ANSVTLAWN           ASTDN  VGVTGYNVYNGAN        L    ATSVTGT TAT           ISGLTAGTSYTFTIKAKDAAGNLSAASNAVTVST
H  AME2    EPAEAPENLRIADIT  DTTVTINWN           ASNGY    VTGYEVLRDGV        V    IGETTRT TFI           DTGLDADRTYTYTIVALGDGGQKSDPSEALEVTT
H  AME1    EPATTPKNLSVVNVT  ETTVTFEWD           QSDGY    VVEYEILRDED        V    VASTIRT TFT           DEDLNPDTTYTYSVVAVGEGGQKSAPSEALKVTT
H  AML2    TAPQPITDLKAVS G  NGKVDLSWS           VVDK     AVSYNIYRSTVKGGLYEKI      ASNVTQI TYT           DTEVTNGLKYVYAVTAVDNDGNESALSNEVEAYP
H  GLCF    TPPTTPGTPVATGVT  TVGASLSWA           ASTDA GSGVAGYELYRVQG    TTQTL    VGTTTAA AYI           LRDLTPGTAYSYVVKAKDVAGNVSAASAAVTFTT
H  GLCF    EPPTTPGTPVASAVT  STGATLAWA           PSTGD  PAVSGYDVLRVQG    TTTT     VVAQTTVPTVT           LSGLTPSTAYTYAVRAKNVAGDVSALSAPVTFTT
H  GLCF    VAPTVPGTPVASNVA  TTGATLTWT           ASTDSGGSGLAGYEVLRVSG    TTQT     LVASPTTATVA           LAGLTPATAYSYVVRAKDGAGNVSAVSSPVTFTT
H  CLCF    QAPSVPSGLTAGTVT  ETSVALSWT           ASTDN  VGVTGYDVYRNGS       K    VGSSSGT TYS           DTGLTAATAYQYSVAAKDAAGNVSQRSSALSVTT
H  AML1    ETPTAP VLQQPGIE  SSRVTLNWS           PSADD  VAIFGYEIYKSSSETGPFIKI     ATVSDSVYNYV           DTDVVNGNVYYYKVVAVDTSYN RTASNTVKATP
H  PXE1    ATAQAPQKLQIPTLSYDDHSVMLVWD           TPEDT  SNITDYQIYQNGQLIGLASQN     NDKNSPAKPYI..(21 res.)..VDGLKAGTDYQFTVRTVYADGTTSNDSNTVTTTT
```

FIG. 3. Computer-generated multiple alignment of 39 Fn3 sequences, including 13 taken from seven bacterial enzymes and 26 representative Fn3 sequences from assorted eukaryotic sources. Letters at beginnings of lines correspond to categories in Table 1. Boldface letters indicate highly conserved residues. UNH, human undulin; TNH, human tenascin; FNH, human fibronectin; TWD, fruit fly twitchin-like protein; TWE, nematode twitchin; TIR, rabbit titin; MKC, chicken myosin light chain kinase; PCC, chicken muscle protein C; I2M, mouse interleukin 2 receptor; I7H, human interleukin 7 receptor; LRH, human leukocyte-related antigen; GRBF, rabbit growth hormone receptor; CDH1, human CD45 antigen; NCHU, human neural cell adhesion molecule; IBH, human B4 integrin; CMM, mouse neural cell adhesion molecule; 45TW, tapeworm 45-kDa protein; KLH, Kallmann syndrome protein; EPH, human tyrosine kinase receptor; AXH, human AX tyrosine kinase; DPA, *Alcaligenes* depolymerase; AME, Gram-positive bacterium amylase; CHB, *Bacillus* chitinase; AML, *Clostridium* amylase; GLC, *Cellulomonas fimi* cellulase; CLCF, *Cellulomonas flavigena* cellulase; PXE, *Erwinia* galacturonosidase.

intracellular and extracellular types of Fn3 module diverged about a billion years ago. Both phylogenetic trees have the bacterial Fn3 units emerging more recently and certainly well after the divergence of prokaryotes and eukaryotes. This assumes the bacterial rate has been similar to that for animals (Table 2). If anything, the bacterial rate might be expected to be greater and the divergence time correspondingly more recent.

DISCUSSION

The first report of an Fn3 occurrence in bacteria was in a chitinase from *Bacillus circulans* (6); the motif was subsequently noted (7) also in the cellulases of *Cellulomonas fimi* and *C. flavigena* (22). We have now found the motif in several other published sequences of bacterial carbohydrases, including the sequences of a poly(3-hydroxybutyrate) depolymerase from *Alcaligenes faecalis* (23), an exo-poly-α-D-galacturonosidase from *Erwinia chrysamthemi* (24), a bifunctional α-amylase–pullulanase from *Clostridium thermohydrosulfuricum* (25), and a maltopentaose-producing amylase from an alkaliphilic Gram-positive bacterium (26). The resemblance to Fn3 units was not described in any of these latter reports (23–26).

Although the enzymes have the common functional feature of mobilizing metabolites from polymeric substances in the environment, the organisms involved are diverse and represent a broad distribution of both Gram-positive and Gram-negative bacteria. They appear to have little in common other than a dependence on these hydrolytic enzymes to obtain their food source, although it is significant that most of them are soil organisms. The Fn3 units occur in different locations and in different numbers in the bacterial enzymes (Fig. 2). The modular fashion in which many of these proteins are constructed and the possibility of genetic shuffling has been remarked upon by others (7).

We are aware that the sequences being compared in this analysis are of a most challenging sort. Thus, they are only 90–100 residues long and have resemblances as low as 9% identity, well below the perilous "twilight zone." Still, it is our carefully considered judgment that the presence of Fn3 units in some contemporary prokaryotes is the result of an unconventional gene transfer at some point in the past.

There are three principal observations that lead us to this conclusion. First, the bacterial sequences are much more similar to the animal sequences than would be expected for conventional vertical descent during the period since the prokaryote–eukaryote divergence (27). Convergent evolution leading to the degree of similarity observed seems unreasonable. Second, the Fn3 sequence occurs only sporadically in bacteria and in situations suggestive of a mobile domain; it is absent from many homologous enzymes. Third, the domain has yet to be found in a fungal or plant protein. All of these considerations are reinforced by sequence alignments and phylogenies generated by an objective computer regimen.

If the bacterial and animal Fn3 sequences had last shared a common ancestor 1.5–2.0 billion years ago at the time of the divergence of prokaryotes and eukaryotes, then it would be anticipated that such sequences would be found in most eukaryotes, including protists, fungi, plants, and animals. It might be expected, also, that the sequences would be found in most bacteria. Such is not the case. No Fn3-like sequences from plants or fungi were detected in our screening. There is a report (28) of a sequence bearing some, but not all, of the features of Fn3 in a trypanosomal enzyme, but the relationship was marginal and the sequence was not retrieved by either searching scheme (Fig. 1). This result can be viewed in two ways, either of which supports the case for horizontal transfer from animals to bacteria. First, the trypanosomal sequence may not be a homologue of Fn3, in which case the

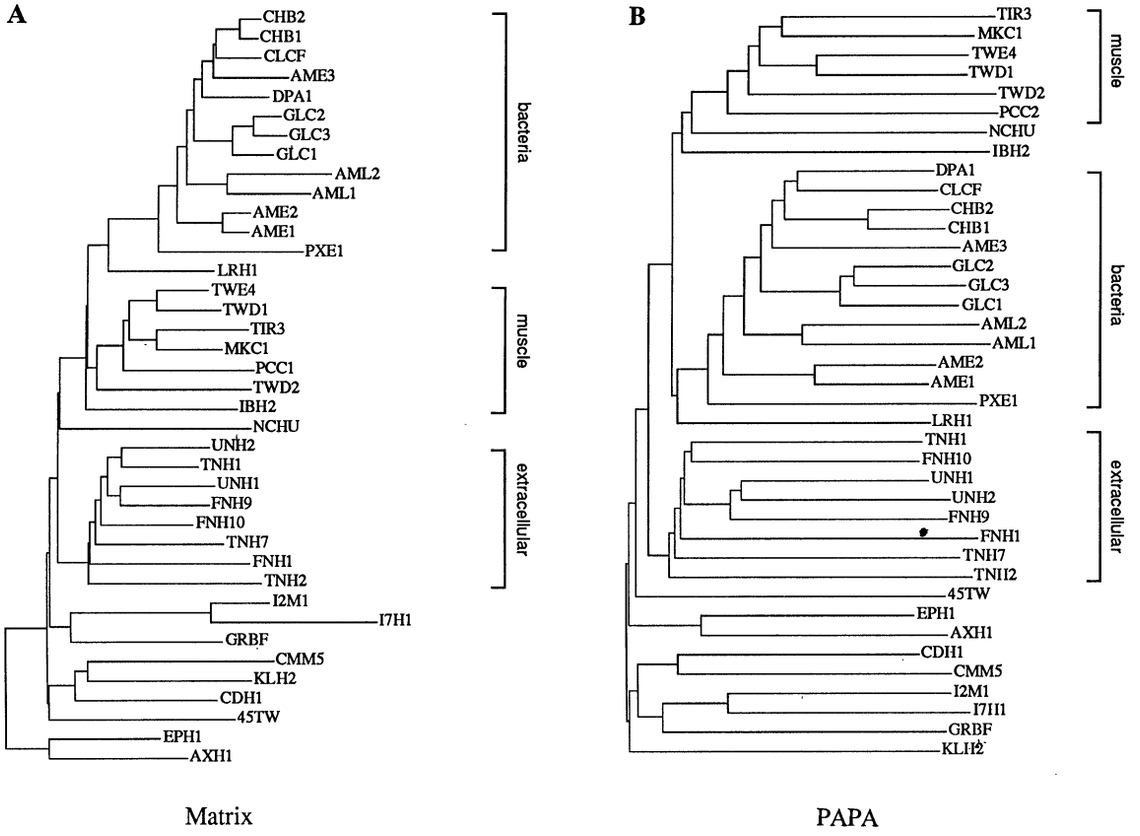

FIG. 4. Phylogenetic trees constructed from the sequence alignment shown in Fig. 3. (*A*) Constructed by a distance matrix method (17). (*B*) Constructed by a four-taxon character analysis method called PAPA (19). See legend to Fig. 3 for codes.

result is moot. Or, second, the sequence has changed so much that it was not recognized by the screening procedures. The latter would be all the more testimony to the fact that the bacterial sequences, which were readily retrieved, are overly similar to animal types and must not be the result of conventional vertical gene propagation.

Many fewer plant and fungal sequences have been reported than for animals, and it is conceivable that sampling biases have excluded the kinds of protein that would likely contain such segments. On the other hand, we estimate that the Fn3 unit occurs in about 2% of all animal proteins: more than 50 independent occurrences in about 2500 proteins, immunoglobulin variable regions and species redundancies aside. As such, this domain is at least as common as the epidermal growth factor (EGF) domain, another unit widespread among animal proteins but not, so far, found in plants or fungi. Several hundred fungal and about as many plant sequences have been reported. Even with great allowances for bias in sampling, it would be expected that at least one Fn3 candidate would have surfaced by now, if they occur at all. We can add that neither the Fn3 nor the EGF units occur anywhere in yeast chromosome III (30). As for prokaryotes, apart from the exceptions described in this article, the Fn3 unit has not yet been found in other bacteria, even though sequences accounting for more than a third of all *Escherichia coli* proteins have been reported.

Naturally occurring horizontal gene transfers are difficult to prove. If sequences are the only criterion, then the case must be made by showing that objective comparisons consistently misplace the entry in an accepted organismic phylogeny. In a sense, the challenge is less great the more distant the putative donor and acceptor in that the distinction between other comparable sequences from the two groups will be clearer. For the same fundamental reason, horizontal transfers between distantly related organisms can sometimes be documented by features like codon usage and general base composition. One very relevant case involves the transfer of a cellulase gene from *Erwinia*, a Gram-negative organism, to *Cellulomonas*, a Gram-positive one (31).

The broad diversity of the bacteria with carbohydrases containing Fn3 modules, as well as the fact that the sequence-based phylogeny of the bacteria is not in accord with accepted bacterial taxonomy, implies that the units are being spread by a series of horizontal transfers. Not only are the bacteria themselves diverse, but the occurrences are in related but different enzymes, and in different locations within the various enzyme sequences. It is as though the unit has been spread by rampant transformation or some broadly based plasmid in some local habitat. That this kind of flagrant transfer might not be exceptional is illustrated by a recent analysis of codon usage in *E. coli* that revealed a cohort of genes sharing the codon distribution profile of phages and which appear to have been imported by horizontal transfer (32).

Recent x-ray and NMR studies (33, 34) have shown that the three-dimensional structure of the Fn3 module is an all-β structure with a rendering not unlike the fold observed in immunoglobulins. Similar folding patterns have been found in proteins as diverse as the CD4 membrane protein (35) and the bacterial chaperone protein PapD (36). There is no recognizable sequence resemblance, however. By coincidence, an extracellular cellulase from *Clostridium thermocellum* has had its crystal structure determined recently, and an accessory immunoglobulin-like domain was found (29); the sequence does not resemble either immunoglobulins or Fn3 units. The question remains: are the similar folds observed in Fn3 units and these other structures the result of common ancestry, or is this a case of structural convergence? It is possible that these structures shared a common ancestor at a very early stage, but all sequence similarity has been eroded.

In the meantime, we must be cautious about assuming that the appearance of similar sequences in animals and bacteria implies an existence before the divergence of prokaryotes and eukaryotes. The Fn3 domain may very well have ancient roots, but its occurrence in bacterial extracellular carbohydrate-splitting enzymes appears to be the result of a much more recent gene acquisition from a eukaryote.

P.B. was supported by grants from the Bundesministerium für Forschung und Technik and the Deutsche Forschungsgemeinschaft, and R.F.D. was supported by grants from the U.S. National Institutes of Health.

1. Bork, P. (1991) *FEBS Lett.* **286,** 47–54.
2. Bork, P. (1992) *Curr. Opin. Struct. Biol.* **2,** 413–421.
3. Patthy, L. (1991) *Curr. Opin. Struct. Biol.* **1,** 351–361.
4. Patthy, L. (1990) *Cell* **61,** 13–14.
5. Bazan, J. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 6934–6938.
6. Watanabe, T., Suzuki, K., Oyanagi, W., Ohnishi, K. & Tanaka, H. (1990) *J. Biol. Chem.* **265,** 15659–15665.
7. Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C., Jr., & Warren, R. A. J. (1991) *Microbiol. Rev.* **55,** 303–315.
8. Bairoch, A. & Boeckmann, B. (1991) *Nucleic Acids Res.* **19,** Suppl. 1, 2247–2249.
9. Bork, P. & Grunwald, C. (1990) *Eur. J. Biochem.* **191,** 347–358.
10. Bork, P., Sander, C. & Valencia, A. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7290–7294.
11. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4355–4359.
12. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 3338–3342.
13. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
14. Feng, D. F. & Doolittle, R. F. (1987) *J. Mol. Evol.* **25,** 351–360.
15. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
16. Doolittle, R. F. (1992) *Protein Sci.* **1,** 191–200.
17. Feng, D. F. & Doolittle, R. F. (1990) *Methods Enzymol.* **183,** 375–387.
18. Fitch, W. M. & Margoliash, E. (1967) *Science* **155,** 279–284.
19. Doolittle, R. F. & Feng, D. F. (1990) *Methods Enzymol.* **183,** 659–669.
20. Cavender, J. A. (1978) *Math. Biosci.* **40,** 271–280.
21. Fitch, W. M. (1970) *Syst. Zool.* **19,** 99–113.
22. Meinke, A., Braun, C., Gilkes, N. R., Kilburn, D. G., Miller, R. C., Jr., & Warren, R. A. J. (1991) *J. Bacteriol.* **173,** 308–314.
23. Saito, T., Suzuki, K., Yamamoto, J., Fukui, T., Miwa, K., Tomita, K., Nakanishi, S., Odani, S., Suzuki, J.-I. & Ishikawa, K. (1989) *J. Bacteriol.* **171,** 184–189.
24. He, S. Y. & Collmer, A. (1990) *J. Bacteriol.* **172,** 4988–4995.
25. Melasniemi, H., Paloheimo, M. & Hemiö, L. (1990) *J. Gen. Microbiol.* **136,** 447–454.
26. Candussio, A., Schmid, G. & Böck, A. (1990) *Eur. J. Biochem.* **191,** 177–185.
27. Doolittle, R. F., Anderson, K. L. & Feng, D. F. (1989) in *The Hierarchy of Life*, eds. Fernholm, B., Bremer, K. & Jörnvall, H. (Elsevier, Amsterdam), pp. 73–85.
28. Pereira, M. E. A., Mejia, J. S., Ortega-Barria, E., Matzilevich, D. & Prioli, R. P. (1991) *J. Exp. Med.* **174,** 179–191.
29. Juy, M., Amit, A. G., Alzari, P. M., Poljak, R. J., Claeyssens, M., Beguin, P. & Aubert, J.-P. (1992) *Nature (London)* **357,** 89–91.
30. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992) *Protein Science*, in press.
31. Guisseppi, A., Aymeric, J. L., Cami, B., Barras, F. & Cveuzet, N. (1991) *Gene* **106,** 109–114.
32. Medigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. (1991) *J. Mol. Biol.* **222,** 851–856.
33. DeVos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992) *Science* **255,** 306–312.
34. Baron, M., Main, A. L., Driscoll, P. C., Mardon, H. J., Boyd, J. & Campbell, I. D. (1992) *Biochemistry* **31,** 2068–2073.
35. Ryu, S.-E., Kwong, P. D., Truneh, A., Porter, T. G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N.-h., Axel, R., Sweet, R. W. & Hendrickson, W. A. (1990) *Nature (London)* **348,** 419–426.
36. Holmgren, A. & Branden, C.-I. (1989) *Nature (London)* **342,** 248–251.