Identification of interdomain sequences promoting the intronless evolution of a bacterial protein family
(protein evolution/albumin binding/recombination/module shuffling/domain structure)

MAARTEN DE CHÂTEAU* AND LARS BJÖRCK
Department of Cell and Molecular Biology, Section for Molecular Pathogenesis, Lund University, S-221 00 Lund, Sweden

Communicated by John Maynard Smith, University of Sussex at Brighton, Brighton, United Kingdom, April 15, 1996 (received for review November 16, 1995)

ABSTRACT In the evolution of eukaryotic genes, introns are believed to have played a major role in increasing the probability of favorable duplication events, chance recombinations, and exon shuffling resulting in functional hybrid proteins. As a rule, prokaryotic genes lack introns, and the examples of prokaryotic introns described do not seem to have contributed to gene evolution by exon shuffling. Still, certain protein families in modern bacteria evolve rapidly by recombination of genes, duplication of functional domains, and as shown for protein PAB of the anaerobic bacterial species Peptostreptococcus magnus, by the shuffling of an albumin-binding protein module from group C and G streptococci. Characterization of a protein PAB-related gene in a P. magnus strain with less albumin-binding activity revealed that the shuffled module was missing. Based on this fact and observations made when comparing gene sequences of this family of bacterial surface proteins interacting with albumin and/or immunoglobulin, a model is presented that can explain how this rapid intronless evolution takes place. A new kind of genetic element is introduced: the recer sequence promoting interdomain, in frame recombination and acting as a structureless flexibility-promoting spacer in the corresponding protein. The data presented also suggest that antibiotics could represent the selective pressure behind the shuffling of protein modules in P. magnus, a member of the indigenous bacterial flora.

Up-to-date numerous Gram-positive bacterial surface proteins have been described that share common functional traits and structural motifs (1). They are elongated, fibrous proteins anchored in the lipid bilayer of the cell membrane and the peptidoglycan of the cell wall by well-conserved regions. These regions, together with the equally conserved signal peptide, constitute the framework regions of these proteins. The part of the proteins that extend beyond the cell wall into the environment of the bacterium is built up of a distal variable NH2-terminal region followed by domains showing high and specific affinity for abundant host proteins such as human serum albumin (HSA) and IgG. Thus, these proteins arrest at least part of their function by binding host proteins to the surface of the bacterium. The functional consequences of these interactions remain unclear, but the coating of a microorganism with host proteins should influence the host–parasite relationship. The gene structure of one of these proteins, the albumin-binding protein PAB of Peptostreptococcus magnus, has been shown to contain a functional domain of 45 amino acid residues responsible for the binding of HSA. This domain is closely related to the HSA-binding domains of protein G, an IgG- and HSA-binding protein of group C and G streptococci (2–4). It was called the GA (protein G-related albumin-binding) module as it was shown to be mobile, i.e., shuffled between genes of different bacterial species, thereby representing the first contemporary example of module shuffling (5). Domains are regions of autonomous structure and these are usually found to be associated with a particular function, whereas a module is defined as a domain that has been found to be mobile. Multidomain proteins are usually the result of recombination events leading to different combinations of modules (6). With continuous genes, such as those of present day prokaryotes, the structural requirements of the proteins and the need to avoid frameshift mutations may restrict the target for successful gene recombination to very short DNA sequences. This is in contrast to the great length of introns in eukaryotic genes, often in the kilobase range, increasing the size of the target for recombination by up to three orders of magnitude. The question raised in the present work is how multidomain proteins evolve in intronless prokaryotes.

MATERIALS AND METHODS

Bacteria. P. magnus strains were clinical isolates from the Department of Clinical Microbiology, Lund University Hospital. The peptostreptococci were grown under strict anaerobic conditions at 37°C in Todd–Hewitt broth (Difco).

Binding Assay. Bacteria were suspended, heat killed (80°C, 5 min), and washed in phosphate buffered saline (PBS) containing 0.02% NaN3 and 0.5% Tween 20. Bacterial suspensions of different concentrations in a volume of 100 μl were mixed with 900 μl of Staphylococcus aureus strain L603 (108 bacteria per ml). Bacterial suspension (200 μl) was then incubated with 104 cpm of 125I-labeled HSA for 30 min. Cells were spun down and the radioactivity of the pellet was measured in a γ-counter and expressed as percentage of added radioactivity.

Proteins, Labeling, and Affinity Chromatography of Peptostreptococcal Proteins urPAB and PAB. HSA was purified from human plasma (7) and radiolabeled with 125I using the Bolton–Hunter reagent (Amersham). Culture medium from peptostreptococci (strains ALB1B and ALB8) expressing protein urPAB and protein PAB, respectively, were used as starting materials and subjected to affinity chromatography on HSA-Sepharose CL-4B (Pharmacia). Columns of Sepharose, coupled with 3–5 mg of HSA/ml packed gel, were equilibrated in PBS. The sample, in 0.01 M phosphate buffer (pH 7.5), was applied and the column was rinsed with PBS and then eluted with 0.1 M glycine buffer (pH 2.0). Eluates were immediately neutralized by the addition of one-tenth volume of unbuffered 1 M Tris. Eluted proteins were separated by SDS/12% PAGE and stained with Coomassie blue.

Preparation of Oligonucleotides and PCR Procedures.

Three oligonucleotides were synthesized, all of which include

Abbreviations: HSA, human serum albumin; SDR, short direct repeats; Tet, tetracycline.

Data deposition: The sequence reported in this paper has been reported in the GenBank data base (accession no. Z48975).

*To whom reprint requests should be addressed: Department of Cell and Molecular Biology, Section for Molecular Pathogenesis, Lund University, P.O. Box 94, S-221 00 Lund, Sweden. e-mail: Maarten.de Chateau@med.kemi.lu.se.

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.
restriction site linkers for cloning in the pH 389 vector (8): RXN0'-(dGGCAATTCGCGCACTAAGGATG-GCAATATTATGGC), RXN0''-(dGGCAATTCGCGCACTAGAAATATATAGAATTTT), RXC6-(dGAAATTCTCGAGTAATTTTATTTTTATG). These oligonucleotides were used as primers with genomic DNA from P. magnus strain ALB1B as template to generate inserts for sequencing vector pGEM-T. For sequencing experiments, two pUC-derived universal primers were also used.

Cloning and Sequencing. PCR products were purified by chloroform–phenol extraction and ethanol precipitation before direct ligation into the T/A-vector pGEM-T (Promega). Subsequently, the ligation mixes were transformed into competent *Escherichia coli* strain JM 109 cells. Clones were selectively grown on ampicillin-containing plates and screened by PCR and for the generation of the desired insert. Double-stranded DNA sequencing was performed with the dideoxynucleotide chain termination method (9), using T7 DNA polymerase (Pharmacia). Both strands of at least two separate PCR clones were sequenced.

Computer Searches and Homology Determinations. Computer searches and homology comparisons were made in databases with the Genetics Computer Group package, Program Manual for the GCG Package (10), and with Gene Works, Program Manual (11).

RESULTS AND DISCUSSION

Albunin-Binding Proteins from *Pepstopstrepococcus magnus*. Clinical isolates of the anaerobic Gram-positive bacterium *P. magnus* were tested for their ability to bind radiolabeled HSA. HSA binding was seen predominantly among isolates from deep wound infections (M.d.C., E. Holst, and L.B., unpublished data). At the highest bacterial concentration, 60% of added radioactivity was bound by strain ALB8, whereas isolates from a different origining infection (ALB1B) bound significantly less (Fig. 1). From the culture supernatant of ALB8, an HSA-binding protein was previously isolated by affinity chromatography on HSA–Sepharose and named protein PAB (5). In the same way, an HSA-binding protein was isolated from the culture supernatant of strain ALB5. This protein, here named protein uPAB (ur- is a prefix meaning origin/ predecessor), had an apparent molecular mass of 42 kDa on SDS/PAGE, which is 5 kDa less than the 47 kDa of protein PAB (Fig. 1). This size difference is equivalent to a GA module, the HSA-binding module found in protein PAB, as well as in protein G of group C and G streptococci. Slot blotting and Western blot experiments confirmed the HSA-binding capacity of protein uPAB (not shown). The difference between the two strains in binding activity and protein size prompted us to examine the structure of protein uPAB.

Cloning, Sequencing, and Structure of the Protein uPAB Gene. The NH₂ terminus of the isolated protein uPAB was sequenced by Edman degradation and the sequence (AEVPANEKYP) showed no homology to the NH₂ sequence of protein PAB. Despite this lack of homology a single fragment containing the gene coding for protein uPAB, uprb, could be generated by PCR with chromosomal DNA of *P. magnus* strain ALB1B as template. This was achieved by using oligonucleotides based on extragenic sequences of the gene for protein L (12), an Ig light chain-binding surface protein from *P. magnus* strain 312 (13). The entire PCR product was 1084 bp with one open reading frame (ORF) comprising 1035 bp (Fig. 2). The ORF translates into a protein of 345 amino acids, with the gene coding for protein PAB (pab) encoding 387 amino acids. The difference in size is thus 42 amino acid residues, which comes close to the size of the GA module (45 amino acid residues).

![Fig. 1. Binding curves of selected *P. magnus* strains and SDS/PAGE of HSA-binding proteins. (A) Dilution series of *P. magnus* strains were tested for the binding of radiolabeled HSA. ALB8 (●) is the strain expressing protein PAB, whereas ALB1B (○) expresses protein uPAB. Strain 312 (□) is the protein L-type strain, whereas strain 505 (△) binds neither albumin nor IgG. (B) Protein uPAB (lane A) and protein PAB (lane B), purified by affinity chromatography on HSA–Sepharose from growth medium of ALB8 and ALB1B, respectively. The proteins were separated by SDS/PAGE (12%) and stained with Coomassie blue.](image-url)

![Fig. 2. Nucleotide sequence and derived amino acid sequence of the uprb gene from *P. magnus* strain ALB1B. The sequence of 1084 nucleotides was obtained from PCR clones including the primer sequences RXN0' and RXC6. There is an ORF starting at position 13 coding for a protein of 345 residues. The first amino acid residues of the signal sequence (SS), the NH₂-terminal A domain, the C domain, and the GA module (uGA) are indicated. The ends of C and uGA are also given. The nucleotide sequence of the protein uprb gene appears in GenBank.](image-url)
When the two sequences are aligned the overall identity is 37% (Fig. 3). Protein urPAB has a highly homologous signal sequence of 26 amino acids, since the NH2-terminal sequence of the isolated protein begins at residue 27. The mature sequence has an NH2-terminal region of 137 residues followed by a domain that shows homology to the C domains of proteins L and PAB. Furthermore, the wall spanning (W) and membrane anchor (M) regions are also similar in the two proteins. Thus, the identity between proteins urPAB and L (Fig. 4) in these conserved regions (Sw, C, W, and M) is 53%, as compared with 32% between these regions in proteins urPAB and PAB. This difference in degree of identity is partly due to the fact that the added GA module in protein PAB resides in the W region. In the NH2-terminal region of the protein urPAB a stretch of 45 amino acids (uGA-ALBB) was found, showing 36% identity to the GA module in protein PAB (pos. 219-264) (Fig. 3). At the same position in protein PAB a similar homologous region (uGA-ALB8) was identified showing 36% identity to the previously described GA module. Because of its low (14-36% identity) homology to other GA modules this uGA domain was not detected in previous sequence analyses. The sequence of urpab reveals that the GA module has invaded this sequence to create pab, hereby making the strain more efficient in binding HSA. Assuming that the number of proteins displayed on the surface of the two strains is the same, the doubling in binding activity seen in the protein PAB strain as compared with the protein urPAB strain could thus be explained by the doubling of binding sites as the second GA module is added.

Sequence Analysis of a Group of P. magnus Genes. Only four genes of the anaerobic bacterial species P. magnus have been sequenced. Apart from protein urPAB the primary structures of proteins PAB and L (strain 312) and L (strain 3316) (18) have also been determined. The four genes share common features that imply that they have evolved from a common ancestor gene (see Fig. 4). The corresponding gene products are fibrous surface proteins anchored in the bacterial cell wall, but they are also released into the growth medium. They contain common, well-conserved framework regions (extra-genic sequences, signal peptide, wall- and membrane-spanning regions), but also various numbers and combinations of functional domains. The three-dimensional structures of the Ig light chain-binding B domains of protein L (19) and the GA module of protein PAB (20) have been determined. The DNA stretches corresponding to known (in the cases of B domains and GA modules) and assumed (in the case of C domains) structureless "spacer" sequences flanking the B, GA, and C domains in proteins PAB, L(312), and L(3316), were assembled and aligned (Fig. 5). This collection equalled 20 DNA sequences ranging from 18 to 72 nucleotides each. The aggregate size of the interdomain sequences is 852 nucleotides, which amounts to 13% of the total sequences. When optimally aligned these sequences showed a consensus sequence of 15 nucleotides (GAA AAA CCA GAA GAA), all of which were in-frame with the ORFs of the genes. Subsequently, the three proteins mentioned above and protein urPAB were searched for the consensus 15 nucleotides with the GCG program FINDPATTERNS. When allowing four mismatches, 15, 11, 40, and 48, in-frame hits were made in the coding sequences of the genes for protein urPAB, protein PAB, protein L(312), and protein L(3316), respectively. Ninety-three of these 114 sequences were found in DNA flanking the structural parts or in the NH2-terminal part of the wall-spanning region. Of these 93 sequences, 70 contained a proline. Only two of the remaining sequences, which were found within known or assumed structure bearing parts of the domains, contained a proline. Moreover, most of the overlapping matches (involving 2-4 consensus sequences) were found in interdomain regions.

Emergence and Evolution of Prokaryotic Protein Modules: Possible Mechanisms. A possible scenario for how the transfer of functional protein modules has come about starts with the massive exposure of the human commensal flora to antibiotics. During the last 5 decades this has stimulated the spread of conjugative plasmids carrying resistance determinants. Conjugation is a highly efficient mode of gene transfer. Conjugative transfer of the PAB type genes is mediated by the sex pilus, a proteinaceous fiber that is secreted by a special type IV secretion system. The pilus acts as an adhesion and transfer apparatus that tethers the donor cell to the recipient cell. The pilus also provides a conduit for the transport of macromolecules from the donor cell to the recipient cell. The pilus is composed of a number of distinct polypeptides, each of which is encoded by a separate gene. The genes encoding the pilus polypeptides are organized in clusters and are flanked by inverted repeat sequences.

Fig. 3. Schematic representation and protein sequence alignment of protein urPAB and protein PAB. The original GA module (uGA), the newly shuffled module (GA), and a domain of unknown function (C) are indicated. Recer sequences with no more than four mismatches at the nucleotide level are indicated by arrows in the upper part of the schematic and by boxes in the alignment. Identities are shaded. The protein sequence of protein PAB has accession number A53586 in the Protein Identification Resource database.
tive plasmids replicate at transfer by a rolling circle mechanism, by which single-stranded DNA enters the recipient (21). Besides being resistant to most restriction enzymes, which enables distant interspecies transfer, single-stranded DNA is more prone to take part in illegitimate recombination events (22). Thus, a conjugative plasmid could have recombined with the chromosome of a bacterium, thereby picking up a DNA fragment coding for a protein module. After integration of this DNA fragment into the plasmid, the plasmid could have been transferred to yet another bacterium and once again recombined. This time the DNA fragment encoding the module could be introduced into another gene resulting in a novel mosaic protein. The recombination of functional domains flanked by short direct repeats (SDRs) is directed to regions in the gene where repeated SDRs reside, i.e., the cell wall region or interdomain regions, to yield new functional hybrid proteins. The SDRs flanking the functional domains are structureless at the protein level and therefore named recer sequences; having a role to play in gene evolution as recombination sites, as well as in protein function as flexible spacers.

At least three of the studied *P. magnus* strains (ALB1B, ALB8, and 312) are known to be tetracycline resistant. Tetracycline is a broad-spectrum antibiotic second only to penicillin in worldwide use. It is established that peptostreptococci have acquired Tet resistance as early as can be traced back in frozen isolates, i.e., 1975 (23), and it has been hypothesized that commensals function as reservoirs for Tet resistance determinants (24). Among these determinants, TetM is the most common; this is also true among peptostreptococci (23). The TetM resistance determinant is found in many species and is carried by the natural plasmid isolate pCF10 (25). pCF10 is a conjugative plasmid originally isolated from *Enterococcus faecalis*, but has been shown to transfer to other Gram-positive bacteria as well (25). In the protein PAB gene, a region homologous to a region in pCF10 is found immediately downstream of the GA module (Fig. 3), supporting the notion that pCF10 or a related plasmid has participated in the transfer of this module.

The recer sequence is a specific example of an SDR. It is not homologous to any of the known recombination promoting sequences, like the X sequence (26). SDRs are 3–20 nucleotides in length and support illegitimate recombination events (22). In contrast to legitimate or general recombination, illegitimate recombination is neither dependent on longer homologies (30–70 nucleotides) or on any known recombination pathways, like *Rec A*. Instead, it seems to be a consequence of errors of enzymes that break and join DNA, or that replicate DNA (22). It has been shown that SDRs can promote both deletions and duplications as well as cointegrate formation (22). Close proximity between the SDRs seems to be of importance, because the likelihood for recombination increases if the distance between them is shorter (27). The likelihood for illegitimate recombination also seems to increase many-fold if the DNA is present in a single-stranded form (28), as is the case in rolling circle replication at conjugation (21). Thus, it has been shown that SDRs, not too far apart from each other, can stimulate the generation of mutations leading to insertion, deletion, or duplication of the genetic material lying between them. In the case of the recer sequences, this genetic material would then correspond to the different folded and functional modules to be shuffled.

The consensus recer sequence of 15 nucleotides corresponds to the pentapeptide Glu-Lys-Pro-Glu-Glu, which is highly charged and hydrophilic. This and the physical properties of the proline residue (29) should contribute to the absence of structure in this peptide. Structural studies on the Ig light chain-binding B domains of protein L (19) and the albumin-binding GA module of protein PAB (20), have revealed that the folded parts are flanked by flexible regions. The location
of four overlapping reer sequences within such a flexible stretch linking two folded Ig light chain-binding B domains of protein L(312) is illustrated in Fig. 6. The spacer often contains a proline residue that should promote the lack of structure and thereby interdomain flexibility, whereas when inserted into a functional domain, the spacer will probably disturb the fold of the domain. Such insertions are supposedly counter-selected, thereby guaranteeing domain integrity. This assumption is supported by the structure of different reer sequences. As mentioned above, 21 of the 114 reer sequences found in the four P. magnus genes were found within DNA sequences corresponding to structure bearing parts of the domains. However, in these 21 sequences, a proline was found only in 2 cases, both in the original GA modules of proteins PAB and urPAB. This is in marked contrast to the remaining 93 sequences found outside the folded parts of the domains, where a proline was present in 70 of the reer sequences. Similar sequences (Glu-Lys-Pro-Glu-Val) are also found as flexibility promoting spacers between the IgGFe-binding domains of streptococcal protein G (30). Also, in some M proteins, surface molecules and virulence determinants in Streptococcus pyogenes, reer-like sequences (Glu-Lys-Glu-Leu-Glu) flank repeated domains (31). Interestingly, linker sequences in eukaryotic zinc finger domains with a similar amino acid sequence (Thr-Glu-Glu-Lys-Pro) were shown to correspond to structureless peptide sequences linking the functionally and structurally defined zinc finger domains. It has also been proposed that these linkers at the gene level might promote the evolution of multi-fingered proteins (32). Finally, in some bacterial enzymes, interdomain peptides rich in proline, alanine, and charged amino acid residues have been shown by nuclear magnetic resonance to constitute flexible linkers (33).

Apart from being inserted between the folded parts of the modules, proline-containing reer sequences are also found within the wall-spanning regions of the P. magnus proteins.

![Fig. 5. Overview of reer sequences in a family of bacterial surface proteins. (A) The assembled structureless DNA sequences flanking B, GA, and C domains (equal segments 11-18 of protein L(312), 11-11 of protein L(3316), and 13 of protein PAB) were aligned and a consensus sequence of 15 nucleotides was found (shaded). (B) Schematically the locations of the 114 reer sequences, found after the primary assembly and definition of the consensus sequence in the four peptostreptococcal genes, are indicated by arrowheads. Multiple arrowheads indicate overlapping reer sequences. Gene structures are on scale, with the 5' ends to the left. Domains/modules are indicated.](https://example.com/)
depicted in Fig. 5. It has been suggested that proline residues, with their tendency to form bends in a peptide, may help these regions to intercalate inside the highly cross-linked peptidoglycan cell wall of Gram-positive bacteria and stabilize the protein within the cell wall (34, 35). Such a functional constraint would conserve the recer sequences and at the DNA level make these regions well suited for recombination of new incoming modules. According to the model, new modules would be directed to the right place in the genome and in the gene, and at the same time be in the ORF. They would also automatically be flanked by flexible spacers, which will improve the binding function of the protein. Taken together, this would increase the chance of a successful recombination leading to a functional hybrid protein by module shuffling without introns. If the hybrid protein is favorable it is then to be decided by natural selection.

Bacterial species are clonal and multiply by binary fission. Genetic diversity and evolution of a species has to be accomplished by random mutations or by uptake and incorporation of new DNA. Transfer of DNA within and between species has been described to take place in a number of ways (36). After transfer, the DNA has to be incorporated by recombination into the genome to be propagated and fixed in the population. That recombination has occurred in nature can best be shown by comparing related gene sequences. There are a number of examples of transfer of entire genes or gene clusters (37), as well as gene segments (38), but most of these examples do not seem to respect domain borders. However, evidence indicating that some eubacterial and bacteriophage genes have evolved by genetic transfer of discrete domains is now mounting. This has been shown by sequence compilation of gene families. In different soil bacteria transfer has been described of the fibronectin type III domains of glycohydrolases (39). The response regulators of bacterial sensory transduction systems consist of a receiver domain linked to an effector domain. In some examples of these genes, sequence data reveal that domains have evolved independently and have later been interchanged to generate new combinations of the two domains (40). In these studies, no modes of recombination have been suggested and structural data have not been included, whereas in the example of the tail fiber genes of different bacteriophages a site-specific recombination system governed by a DNA invertase has been proposed for the shuffling of host-range determining segments (41). Albeit, at the protein level these segments do not form separate folded structures, but are part of a long extended conformation. Therefore, the data presented here on bacterial module shuffling in a family of P. magnus genes are unique for several reasons. First, the sequences of two genes were found to differ only by the shuffled module. Second, three-dimensional structural data and binding properties show that the modules are distinct structural and functional units flanked by flexible linkers. Third, the linkers are composed of conserved SDRs, suggesting a role for these sequences in the evolution of multidomain bacterial proteins.

The evolution of proteins has been highly accelerated by the exchange of entire functional units—i.e., modules. This has led to the emergence of multifunctional proteins with distinct structural domains separated by structureless linkers like beads on a string. In eukaryotes the appearance of introns and exon shuffling has been suggested to have led to the Big Bang of animal evolution some 570 million years ago (42), when a multitude of multicellular animals evolved (43). The sudden increase in the speed of protein evolution that presumably lead to this tremendous radiation might have been triggered by a great environmental disaster (44). The emergence of the antibiotic era can be viewed as such an environmental challenge to the bacteria (45), possibly stimulating the rapid evolution of new proteins in the bacterial species that are most exposed, i.e., those belonging to the normal flora such as P. magnus.