Protein Science (1996), *5*:1939–1941. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Members of the immunoglobulin superfamily in bacteria

ALEX BATEMAN,[1] SEAN R. EDDY,[2] AND CYRUS CHOTHIA[1]

[1] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England
[2] Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110

(RECEIVED June 4, 1996; ACCEPTED June 26, 1996)

Abstract: We report a prediction that two prokaryotic proteins contain immunoglobulin superfamily domains. Immunoglobulin-like folds have been identified previously in prokaryotic proteins, but these share no recognizable sequence similarity with eukaryotic immunoglobulin superfamily (IgSF) folds, and may be the result of the physics and chemistry of proteins favoring certain common folds. In contrast, the prokaryotic proteins identified have sequences whose match to the immunoglobulin superfamily can be detected by hidden Markov modeling, BLASTP matches, key residue analysis, and secondary structure predictions. We propose that these prokaryotic immunoglobulin-like domains are almost certain to be related by divergence from a common ancestor to eukaryotic immunoglobulin superfamily domains.

Keywords: hidden Markov model; immunoglobulin domain; immunoglobulin superfamily; prokaryotic

Members of the immunoglobulin superfamily (IgSF) play a central role in the vertebrate immune system, where they include antibodies, T-cell receptors, major histocompatibility antigens, and in the cell adhesion molecules of most, if not all, metazoa (Williams, 1987; Williams & Barclay, 1988; Kuma et al., 1991; Bork et al., 1994; Harpaz & Chothia, 1994). Immunoglobulin superfamily domains are also found in intracellular proteins, such as in the giant muscle proteins twitchin and titin (Benian et al., 1989; Labeit & Kolmerer, 1995). Speculation as to when the immunoglobulin superfamily arose makes identification of prokaryotic examples interesting.

Here we report that there is strong evidence that two bacterial proteins, the IgA Fc receptor of *Streptococcus agalactiae* and endoglucanase C of *Cellumonas fimi*, contain IgSF domains. The evidence for this comes from a hidden Markov model (HMM) (Krogh et al., 1994; Eddy et al., 1995) of the IgSF protein fold, the BLASTP local sequence alignment tool (Altschul et al., 1990), key residue analysis (Chothia & Lesk, 1987; Harpaz & Chothia, 1994), and a prediction of their secondary structures by the PHD program (Rost & Sander, 1994)

Nearly all the members of the immunoglobulin superfamily can be assigned to one of a number of different sets: V, C1, C2, and I (Williams & Barclay, 1988; Harpaz & Chothia, 1994). Members of one set are more similar to one another in sequence and structure than to members of the other sets. From alignments of the sequences of a protein family, an HMM can be built to encode the probabilities of different residues occurring at particular sites. This model can be used to detect other very distant members of the family (Krogh et al., 1994; Eddy et al., 1995). We built an HMM from a multiple alignment of the IgSF domains belonging to the I set. Forty domains from neural cell adhesion molecules were aligned to the sequence of telokin. The HMM was used to search the SWISS-PROT release 32 database (Bairoch & Boeckmann, 1991).

Residues 434–534 of the IgA Fc receptor of *S. agalactiae* matched this HMM with a score of 35 bits. Residues 918–1006 of endoglucanase C of *C. fimi* matched the HMM with a score of 34 bits. We would expect that a score of 17 bits would be a significant match in searching a database of the size of SWISS-PROT (Krogh et al., 1994; Eddy et al., 1995). This theoretical calculation is supported empirically by the 347 other domains, which were detected by the HMM with scores greater than 18.6, that have been described previously as IgSF members on the basis of characteristics of their sequences.

Residues 918–1006 of endoglucanase C are 48% identical to residues 1008–1097. Although this second domain does not match the HMM with a significant score, this sequence identity clearly implies that endoglucanase C contains two tandem Ig I set domains.

HMM scores of 34 and 35 are highly reliable in our experience. We did, however, check our results by other methods: a database search with BLASTP, key residues analysis, and PHD secondary structure prediction.

BLASTP (Altschul et al., 1990) could not find a significant match between the sequence of the IgA Fc receptor of *S. agalactiae* and any other sequence in SWISS-PROT. The two domains of endoglucanase C of *C. fimi*, however, matched the IgSF protein perlecan with an *e*-value of 7.8×10^{-9} for residues 918–1006 and 2.4×10^{-9} for residues 1008–1097. These are very significant scores (Altschul et al., 1990). Indeed, Perlecan, a basement membrane protein, contains 15 tandem IgSF domains (Noonan et al., 1991) and several of these matched the two endoglucanase C domains with significant scores.

Reprint requests to: Alex Bateman, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England; e-mail: agb@mrc-lmb.cam.ac.uk.

Key residues are those that, through their packing, hydrogen bonds, or unusual torsion angles, play the major role in determining three-dimensional structure of a protein. These residues tend to be strongly conserved, in type if not in identity, over long evolutionary periods and can be used to detect distant evolutionary relationships. The key residues of the first I set structure, telokin, were determined by Harpaz and Chothia (1994), and in Figures 1 and 2 we show, as an example, the key residues that together mainly determine certain loop conformations. On the basis of the matches made to the key residues found in telokin, it was predicted that certain other members of the immunoglobulin superfamily belong to the I set with structures close to that of telokin (Harpaz & Chothia, 1994). So far, the predictions for two of these domains, the M5 domain of titin and domain 1 of VCAM, have been shown to be correct by their subsequent structure determination (Jones et al., 1995; Pfuhl & Pastore, 1995). Note that both of these proteins have a very low sequence identities with telokin: 23 of 90 in the first case and 10 of 89 in the second.

In Figure 1, we align the sequences of the three known I set structures with the three sequences from the IgA Fc receptor and endoglucanase C. The three bacterial sequences share only a small number of identical residues with the known structures; these are in the range 6–25. Inspection of the alignment shows, however, that to a very large extent the key residues in the core of the known structures are the same or conservatively substituted in the three new sequences and in several cases the conservation extends to the loop regions.

The conserved tryptophan in the C strand is conserved in endoglucanase C and conservatively substituted by phenylalanine in the IgA Fc receptor. The cysteines in the B and F strands of telokin are conservatively substituted by alanine, valine, and leucine in the bacterial proteins; these substitutions can be found in many of the members of the IgSF that occur in muscle proteins and whose structures are known (Improta et al., 1996), and in the second domain of CD4 (Wang et al., 1990; Garret, 1993). Notably the domains of endoglucanase C conserve the key residues in the BC and FG loops found in telokin, implying close structural similarity in this region (Figs. 1, 2).

The known I set structures contain two β-sheets formed by eight or nine strands: A, A′, B, C, (C′), D, E, F, and G; see Figure 1. The sequences of the three bacterial domains were submitted to the PHD secondary structure prediction server (Rost & Sander, 1994). PHD predicts for both proteins strand conformations in regions equivalent to the A, A′, B, E, F, and G strand regions in the known structures (Fig. 1). Predictions for strand regions equivalent to the other strands, C, C′, and D, are clearly made for one of the two bacterial proteins, but are weak or absent in the other. Given that the PHD program has an expected accuracy of above 70% (Rost & Sander, 1994), the predictions for the bacterial sequences agree very well with that observed in the proteins of known structure.

S. agalactiae is an important human pathogen. *S. agalactiae* is a B group *Streptococcus* that causes meningitis in newborns (Jerlstrom et al., 1991). The IgA Fc receptor is part of the c complex that may contribute to the virulence of this organism (Lancefield et al., 1975). Two specific IgA-binding regions have been defined; the second of these two regions contains the Ig-like domain (Jerlstrom et al., 1991). It is noteworthy that the human IgA Fc receptor contains two immunoglobulin-like domains. Endoglucanase C is a cellulase that catalyzes the endohydrolysis of 1,4-β-D-glucosidic linkages in cellulose. The function of the region containing the Ig-like domains is unknown (Coutinho et al., 1992). Endoglucanases B and D from *C. fimi* contain fibronectin type III domains, which have an immunoglobulin-like fold, but there is no evidence that these domains are members of the IgSF (Little et al., 1994).

```
Strand    A----A       A'--A'  B--------B    C-----C  C'-C'
Telokin   VKPYFTKTI--LDMDVVEGSAARFDCKVEGY-PDPEVMWFKD--DNPVKESR--
Titin M5  ---RILTKP--RSMTVYEGESARFSCDTDGE-PVPTVTWLRk--gqvlstsa--
VCAM D1   --fKIettpe-sRYLAQIGDSVSLTCSTtgc-eSPFFSWRTq---idspln---

FCR  434  QKIELTVSP--ENITVYEGEDVKFTVTAKsd-skTTLDFSDl--ltkynpsvsd
PHD        -----        ----   ----------     --++++  +--    ---
ENDC 918  TAPVVTRQP--VDATVALGADATFTAEASGV-PAPTVRWQVr-agrgwkdva--
ENDC 1008 AAPVVTQHP--ADVRARVGTRAVFRAAADGY-PTPCVVWQVrwgggswrPIP--
PHD          ----      ----------------       -----      -----


 D-----D     E-----E       F------F     G----------G    Strand
HFQIDYDE----EGNCSLTISEVCGDDDAKYTCKAVNS-LGEA-TCTAELLVE   Telokin
RHQVTTTK-----YKSTFEISSVQASDEGNYSVVVENS-EGKQ-EAEFTLTIQ   Titin M5
-GKVTNEG-----TTSTLTMNPVSFGNEHSYLCTATce-srKL-EKGIQVEIY   VCAM D1

RISTNYKtntdnhKIAEITIKNLKLNESQTVTLKAKDD-SGNVvekTFTITVQ 534  FCR
-- -----      --------       -------      ----------        phd
-GATGT----------TLTVRATARTDGTRYRAVFTNA-AGSVesAVVRLTVE 1006 ENDC
-WATST----------TLSVPVTVLAaGTEYRAVFTNA-VGTAatepAELAVQ 1097 ENDC
                ---------    -------     -----------        phd
```

Fig. 1. Regions in the bacterial sequences that are expected to share the same structure as telokin are shown in upper case. Regions expected to differ in conformation from telokin are shown in lower case. The secondary structure (Strand) of telokin is shown, as is the PHD secondary structure prediction (PHD) of the single IgA Fc receptor domain (FCR) and the pair of endoglucanase C (ENDC) domains: residues predicted to be in an extended conformation are marked by − signs and those predicted to be in an helical conformation by +. Key residues important for the structure of the Ig fold of telokin are shown in bold letters (Harpaz & Chothia, 1994). The chemical natures of these side chains are largely conserved, implying structural similarity of the bacterial domains to telokin. The SWISS-PROT identifiers are BAG_STRAG and GUNC_CELF1 for the Fe receptor and endoglucanase, respectively.

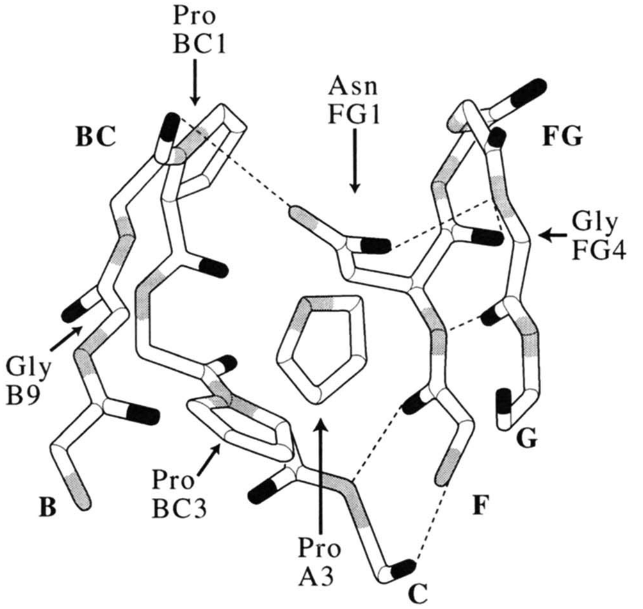

Fig. 2. Packing and key residues of the BC and FG turns in telokin and the M5 domain of titin. The BC and FG turns pack together with the beginning of the A strand. Residues important for this structure are the hydrophobic or neutral residues at A3, B9, and BC3; a Pro with a *cis* peptide conformation at BC1; the Asn with its side-chain hydrogen bonds at FG1 and the Gly with its $+\phi, -\psi$ main-chain conformation at FG4. All these key residues are buried except for that at BC1. For the conformation of the BC and FG to be conserved, we expect these key residues to be conserved. If the loops change in size or lose these key residues, we would expect different conformations as occurs in domain 1 of VCAM (Jones et al., 1995).

We conclude that, taken together, the results reported here clearly show that the three bacterial domains are related to the eukaryotic immunoglobulin superfamily by divergent evolution from a common ancestral sequence. Previously proposed prokaryotic examples of the IgSF have been recognized on the basis of similar folding topologies, but their sequence similarity to eukaryotic IgSF members has been so low as to raise questions of possible evolution from two independent ancestors (Hofmann et al., 1989; Bork et al., 1994; Gaskell et al., 1995). We have detected these domains based solely on sequence similarities, and thus feel confident drawing a conclusion of divergent evolution. However, we cannot dismiss, at present, the possibility that these Ig domains were acquired horizontally by prokaryotes from eukaryotes after their divergence [as has been proposed for the bacterial examples of the fibronectin type III domain (Bork & Doolittle, 1992; Little et al., 1994)] rather than their being the descendants of a true bacterial ancestor of the Ig superfamily.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Bairoch A, Boeckmann B. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res 19*:2247–2249.

Benian GM, Kiff JE, Neckelmann N, Moerman DG, Waterston RH. 1989. Sequence of an unusually large protein implicated in regulation of myosin activity in *C. elegans*. *Nature 342*:45–50.

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

Bork P, Holm L, Sander C. 1994. The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol 242*:309–320.

Chothia C, Lesk AM. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol 196*:901–917.

Coutinho JB, Gilkes NR, Warren RA, Kilburn DG, Miller RC. 1992. The binding of *Cellumonas fimi* endoglucanase C (CenC) to cellulose and Sephadex is mediated by the N-terminal repeats. *Mol Microbiol 6*:1243–1252.

Eddy SR, Mitchison G, Durbin R. 1995. Maximum discrimination hidden Markov models of sequence consensus. *J Comput Biol 2*:9–23.

Garret TPJ. 1993. Refinement and analysis of the structure of the first two domains of human CD4. *J Mol Biol 234*:763–778.

Gaskell A, Crennell S, Taylor G. 1995. The three domains of a bacterial sialidase: A β-propeller, an immunoglobulin module and a galactose-binding jelly-roll. *Structure 3*:1197–1205.

Harpaz Y, Chothia C. 1994. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J Mol Biol 238*:528–539.

Hofmann BE, Bender H, Schulz GE. 1989. Three dimensional structure of cyclodextrin glycosyltransferase from *Bacillus circulans* at 3.4 Å resolution. *J Mol Biol 209*:793–800.

Improta S, Politou AS, Pastore A. 1996. Immunoglobulin-like modules from titin I-band: Extensible components of muscle elasticity. *Structure 4*:323–337.

Jerlstrom PG, Chhatwal GS, Timmis KN. 1991. The IgA-binding β antigen of the c protein complex of Group B streptococci: Sequence determination of its gene and detection of two binding regions. *Mol Microbiol 5*:843-849.

Jones EY, Harlos K, Bottomley MJ, Robinson RC, Driscoll PC, Edwards RM, Clements JM, Dudgeon TJ, Stuart DI. 1995. Crystal structure of an integrin-binding fragment of vascular cell adhesion molecule-1 at 1.8 angstroms resolution. *Nature 373*:539-544.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. *J Mol Biol 235*:1501–1531.

Kuma K, Iwabe N, Miyata T. 1991. The immunoglobulin family. *Curr Biol 1*:384-393.

Labeit S, Kolmerer B. 1995. Titins: Giant proteins in charge of muscle ultrastructure and elasticity. *Science 270*:293–296.

Lancefield RC, McCarty M, Everly WN. 1975. Multiple mouse-protective antibodies directed against group B streptococci. Special reference to antibodies effective against protein antigens. *J Exp Med 142*:165–179.

Little E, Bork P, Doolittle RF. 1994. Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *J Mol Evol 39*:631–643.

Noonan DM, Fulle A, Valente P, Cai S, Horigan E, Sasaki M, Yamada Y, Hassell JR. 1991. The complete sequence of Perlecan, a basement membrane heparan sulfate proteoglycan, reveals extensive similarity with laminin A chain, low density lipoprotein-receptor, and the neural cell adhesion molecule. *J Biol Chem 266*:22939–22947.

Pfuhl M, Pastore A. 1995. Tertiary structure of an immunoglobulin-like domain from the giant muscle protein titin: A new member of the I set. *Curr Biol 3*:391–401.

Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins Struct Funct Genet 19*:55–72.

Wang J, Yan Y, Garret TPJ, Liu J, Rodgers DW, Garlick RL, Tarr GE, Husain Y, Reinharz EL, Harrison SC. 1990. Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature 348*:411–418.

Williams AF. 1987. A year in the life of the immunoglobulin superfamily. *Immunology Today 8*:298–303.

Williams AF, Barclay AN. 1988. The immunoglobulin superfamily-domains for cell surface recognition. *Annu Rev Immunol 6*:381–405.