Protein Cell 2013, 4(8): 569–572
DOI 10.1007/s13238-013-3903-2

Protein & Cell

NOTE

The sequence signature of an Ig-fold

Jia-Huai Wang✉

Dana-Farber Cancer Institute and Department of Pediatrics, Harvard Medical School, Boston, MA 02215, USA
College of Life Sciences, Peking University, Beijing 100871, China
✉ Correspondence: jwang@red.dfci.harvard.edu

Ig superfamily (IgSF) constitutes the largest superfamily in human genome. In particular, Ig-like domains are the most abundant structural module within cell surface receptors, functioning in nervous as well as immune system. Here I describe some key sequence signature of an I-set Ig-like domain from known structures of IgSF members. These signature residues define the I-set Ig-like domain, which should aid structural and functional studies of cell surface receptors.

Immunoglobulin (Ig) fold was first recognized in antibody, where the nomenclature came from (Bork et al., 1994). In fact, Ig-fold is an evolutionarily much ancient structural unit that can be found in *C.elegans* (Teichmann and Chothia, 2000). Ig superfamily (IgSF) constitutes the largest superfamily in human genome, due to its extensive usage in more recently developed immune system in vertebrates (Lander et al., 2001). Although Ig-like domains also exist in intracellular environment like muscle proteins titin and telokin (Holden et al., 1992; Politou et al., 1994), they are nonetheless the most abundant structural units within cell surface receptors, serving nervous as well as immune functions. Along with some other domains such as fibronectin type III domains, epidermal growth factor (EGF) domains, etc. they form modular structures of receptor molecules on the cell surface (Chothia and Jones, 1997; Wang and Springer, 1998).

An Ig-like domain is composed of roughly 100 residues, folding into two β sheets packing face-to-face, with strands from two sheets making an angle of 30°. Since an antibody has variable domains and constant domains, Ig domain has originally been classified into V-set and C-set. A V-set Ig domain has β strands A, B, E and D on one sheet and A′, G, F, C, C′ and C″ strands on the other, whereas a C-set Ig domain lacks A′ and C″ strand on either edges. The two sheets are linked together by a conserved disulfide bond between B strand and F strand. C-set has been further divided into C1 and C2 sets. The difference is that C2-set does not have a D strand, whereas the C1-set has a shorter C′ strand (Wang and Springer, 1998). More recently, Harpaz and Chothia have noticed a new class of Ig domain, the I-set (Harpaz and Chothia, 1994; Meijers et al., 2007). It can be described as a V-set truncated just on one side, missing the C″ strand, as shown in the Fig. 1. Similar to the division of C1 and C2 sets, the I-set also can be defined as I1 and I2 sets (Wang and Springer, 1998).

Modular receptors often have a V-set Ig-like domain at the N-terminus for ligand-binding, such as antigen-recognition. By contrast, I-set Ig-like domains usually function as one of the building blocks lined up in tandem to present the ligand-binding domain on the cell surface. This can be seen in CD2 (Jones et al., 1992), CD4 (Wu et al., 1997) and many other receptors. There is also a pool of receptors like immune receptor intercellular adhesion molecule-1

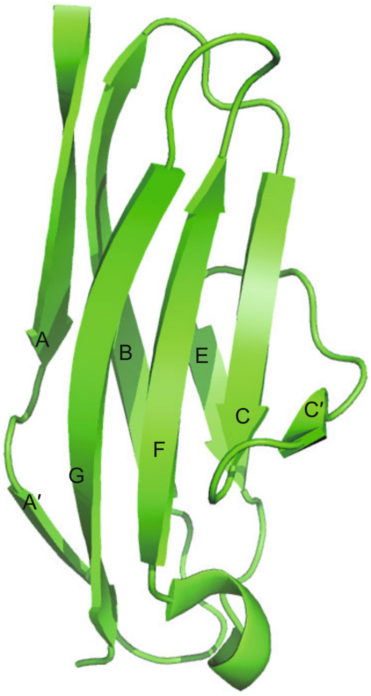


Figure 1. Overall topology of an I-set Ig-like domain exemplified by Dscam domain 4. The domain has A′GFCC′ β sheet on the front and ABE β sheet on the back. At the bottom, there is a short helix between E and F strands.

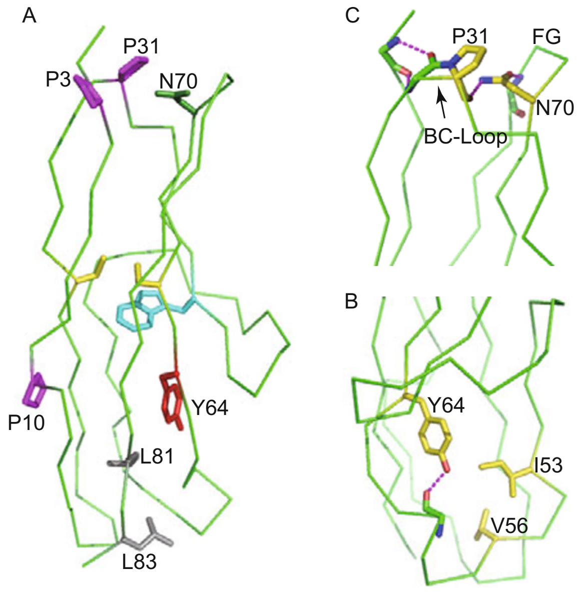


Figure 2. Signature residues of an I-set Ig-like domain, Dscam domain 4. (A) Sequentially they are Pro3, *cis*-Pro10, Cys25, *cis*-Pro31, Trp37, Tyr64, Cys66, Asn70, Leu81 and Leu83. There are two more signature residues, Ile53 and Val56, which are shown in Fig. 3. These residues are colored differently. (B) This is the bottom part of an I-set Ig-like domain, Dscam domain 4. Shown here is the Tyr64 packed with a β bulge formed by Ile53 and Val56. Note that Tyr64's hydroxyl group forms a hydrogen-bond to the main chain carbonyl oxygen 4-residue upstream the peptide chain at the corner of EF-turn, hence the name Tyr-corner. (C) This is the top part of an I-set Ig-like domain, Dscam domain 4. Asn70 bridges BC-loop and FG-loop with two hydrogen-bonds. On the other side, one residue N-terminal to the conserved *cis*-Pro31 forms two main chain hydrogen-bonds to the very N-terminal residue Ala2. This defines the beginning of the domain.

(ICAM-1) (Wang and Springer, 1998) and neuro-receptor Dscam (Meijers et al., 2007; Sawaya et al., 2008) that exclusively have I-set for all Ig-like domains. Therefore I-set is actually the most abundant Ig-fold, and plays a critical biological role in these cell surface receptors. In order to facilitate functional as well as structural studies of these receptors, it is important to know how to accurately define the I-set domain. Careful examination of the known structures of these receptors has allowed for the definition of a clear sequence signature that determines an I-set Ig-like domain.

Fig. 2A is a representative I-set Ig domain from the domain 4 of Dscam, a neuro-receptor of *Drosophila* (Meijers et al., 2007). The figure gives an overview of an I-set Ig domain with side chains painted in different color code as signature residues. The conserved residues marked play a critical role in maintaining an I-set Ig domain. Below is a detailed description of what part each one of the residues plays.

HYDROPHOBIC CORE AT THE CENTER

At the center of the domain is the well-known conserved disulfide bond between Cys25 on the B strand and Cys66 on the F strand. Underneath this disulfide bond is an equally conserved Trp37 on the C strand with its side chain indole ring packing against the disulfide bond, forming a hydrophobic core of the domain. In very few cases the Cys and/or Trp may be replaced by other hydrophobic residues. CD2 domain 1, for example, does not have the disulfide bond. Instead it has an Ile and a Val at the corresponding places on the B and F strands, respectively (Jones et al., 1992).

TYR-CORNER AND INTERACTING β-BULGE AT THE BOTTOM

Two-residue upstream from Cys66 is another conserved residue Tyr64. This Tyr forms a so-called Tyr-corner (Hemmingsen et al., 1994) with its hydroxyl group making hydrogen bond to the main chain carbonyl oxygen of a residue 4-residue upstream at the corner of the domain, as illustrated in Fig. 2B. In fact, the Tyr-X-Cys (X can be any other residue) motif on the F strand is the most conserved sequence signature of any Ig-like domain. Adjacent to the Tyr-corner at the bottom of the domain there are two hydrophobic residues, Ile53 and Val56, pointing inward the domain, contacting the side chain of this Tyr64. These two hydrophobic residues push two consecutive hydrophilic residues Glu54-Ser55 out, forming a β bulge (Fig. 2B), a conserved structural feature that marks the end of the E strand. Sequence-wise, the pattern is Pho-X-Pho-X-X-Pho, where the Pho represents one hydrophobic residue, and X can be anything (Fig. 2B). Very often, from the end of the E strand to the beginning of F strand is a short helix. All of these features offer a solid structure at an Ig-like domain's bottom.

THE MARKER FOR C-TERMINAL BOUNDARY

One hallmark of being an I-set, like V-set, is to have an A-A′ kink on one side of domain. But distinct from V-set, I-set does not have a C″ strand on the other side. In both V-set and I-set, the Ig-like domain begins with an A strand. Half way down the domain in the ABED sheet, the polypeptide chain crosses over the domain at a kink, and then becomes A′ strand, joining the A′GFCC′ sheet. In many cases, a *cis*-Pro resides at the kink. A′ strand runs parallel to the G strand, and shields the G strand from exposure. Sitting in the middle of a β sheet, the C-terminal of G strand will obey the usual pattern for any such typical β strand to have alternate distribution of hydrophobic-hydrophilic-hydrophobic residues. Fig. 2A clearly shows Leu81 and Leu83 (in grey color) pointing inward. A general pattern is Pho-X-Pho-X. Since this is the end of an Ig-like domain, this pattern becomes a very reliable sequence signature that defines the C-terminal boundary of an I-set (V-set as well) Ig domain.

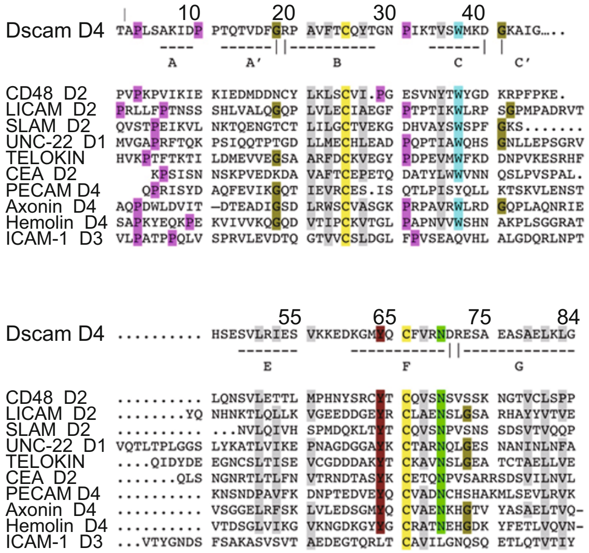
ASN-BRIDGE

Fig. 2C demonstrates the exquisite conformation of I-set Ig domain's top portion. On the right side, Asn70 acts like a bridge connecting BC loop and FG loop, using its amide group hydrogen-bonding to main chain atoms from the two loops, respectively. In antibodies and T cell receptors, their variable domain has so-called CDR loops at the top of the domain. Most variable BC loop (CDR1) and FG loop (CDR3) are widely open in order to "grab" antigen in a groove structure for immune function (Rudolph et al., 2006). This is even seen in CD8, one of T cell receptor's co-receptors. A CD8αα homo-dimer or CD8αβ hetero-dimer binds to class I major histocompatibility complex (MHC) also using these CDR-like loops of their V-set Ig-like domains (Gao et al., 1997; Kern et al., 1998; Wang et al., 2009). By contrast, as a building block, I-set Ig-like domain is required to have a more compact shape at the top of a domain such that many of these domains can be lined up on top of one another like beads on a string on the cell surface. A CDR-like loop with open conformation is not suitable. Instead, an "Asn-bridge" seals up the ligand-binding groove in I-set Ig-like domain.

A PAIR OF MAIN CHAIN HYDROGEN BONDS THAT DEFINES THE N-TERMINUS

Equally interesting structural feature at the top is the existence of another conserved *cis*-Pro31 situated at the middle of the BC loop. This is the residue with its carbonyl oxygen hydrogen-bonded to Asn70, described above (Fig. 2C). Immediately N-terminal to this Pro31 is a residue that forms a pair of main chain hydrogen bonds to Ala2, the very beginning of the I-set Ig-like domain. This very conserved feature has been observed in many known I-set structures (Wang and Springer, 1998). In overwhelming majority cases, the regular β strand A does not start till 3–5 residues from the N-terminus. It is this pair of main chain hydrogen bonds that clearly defines an I-set Ig-like domain's N-terminus. Proline is frequently seen in this 3–5-residue irregular portion of polypeptide chain. In the Dscam domain 4 shown in Fig. 2C, for instance, this irregular fragment includes Ala-Pro-Leu and the β strand A commences at Ser5. Since Pro cannot offer two hydrogen bonds, the N-terminus of the domain can thus be recognized one residue before this Pro, the Ala2 (Meijers et al., 2007). Pro being the second residue of an I-set Ig-like domain is a fairly common theme.

The Asn-bridge on the right and the pair of main chain hydrogen bonds to Ala2 on the left in Fig. 2C pull A strand, BC loop and FG loop altogether, consolidating the domain's top portion, and making it suitable as a compact building block.



```
             |       10         20         30          40
Dscam D4     TAPLSAKIDP PTQTVDFGRP AVFTCQYTGN PIKTVSWMKD GKAIG….
                ----    ------||----------      ------|  |
                  A       A'          B               C    C'

CD48  D2     PVPKPVIKIE KIEDMDDNCY LKLSCVI.PG ESVNYTWYGD KRPFPKE.
LICAM  D2    PRLLFPTNSS SHLVALQGQP LVLECIAEGF PTPTIKWLRP SGPMPADRVT
SLAM  D2     QVSTPEIKVL NKTQENGTCT LILGCTVEKG DHVAYSWSPF GKS.......
UNC-22  D1   MVGAPRFTQK PSIQQTPTGD LLMECHLEAD PQPTIAWQHS GNLLEPSGRV
TELOKIN      HVKPTFTKTI LDMEVVEGSA ARFDCKVEGY PDPEVMWFKD DNPVKESRHF
CEA  D2          KPSISN NSKPVEDKDA VAFTCEPETQ DATYLWWVNN QSLPVSPAL.
PECAM D4          QPRISYD AQFEVIKGQT IEVRCES.IS QTLPISYQLL KTSKVLENST
Axonin  D4   AQPDWLDVIT –DTEADIGSD LRWSCVASGK PRPAVRWLRD GQPLAQNRIE
Hemolin  D4  SAPKYEQKPE KVIVVKQGQD VTIPCKVTGL PAPNVVWSHN AKPLSGGRAT
ICAM-1  D3   VLPATPPQLV SPRVLEVDTQ GTVVCSLDGL FPVSEAQVHL ALGDQRLNPT

                                55         65         75         84
Dscam D4     .......... HSESVLRIES VKKEDKGMYQ CFVRNDRESA EASAELKLG
                          --------     -----------||-------------
                              E              F             G

CD48  D2     .......... LQNSVLETTL MPHNYSRCYT CQVSNSVSSK NGTVCLSPP
LICAM  D2    ........YQ NHNKTLQLLK VGEEDDGEYR CLAENSLGSA RHAYYVTVE
SLAM  D2     .......... ...NVLQIVH SPMDQKLTYT CQVSNPVSNS SDSVTVQQP
UNC-22  D1   VQTLTPLGGS LYKATLVIKE PNAGDGGAYK CTARNQLGES NANINLNFA
TELOKIN      ....QIDYDE EGNCSLTISE VCGDDDAKYT CKAVNSLGEA TCTAELLVE
CEA  D2      .......QLS NGNRTLTLFN VTRNDTASYK CETQNPVSARRSDSVILNVL
PECAM D4     .......... KNSNDPAVFK DNPTEDVEYQ CVADNCHSHAKMLSEVLRVK
Axonin  D4   .......... VSGGELRFSK LVLEDSGMYQ CVAENKHGTV YASAELTVQ-
Hemolin  D4  .......... VTDSGLVIKG VKNGDKGYYG CRATNEHGDK YFETLVQVN-
ICAM-1  D3   ...VTYGNDS FSAKASVSVT AEDEGTQRLT CAVILGNQSQ ETLQTVTIY
```

Figure 3. Aligned sequences with signature residues having the same color code as that in Fig. 1. Listed here are immune receptors human CD48 (GenBank: CAG33293.1), human L1 cell adhesion molecule (L1CAM) (Reid and Hemperly, 1992), human signaling lymphocytic activation molecule (SLAM) (GenBank: AAK77968.1), human ICAM-1 (Yang et al., 2004) and insect hemolin (Su et al., 1998), and other adhesion molecules human CEA (Tan et al., 2002), human PECAM-1 (Newman et al., 1990), neuro-receptors *Drosophila* Dscam (Meijers et al., 2007) and chicken axonin (Freigang et al., 2000), as well as muscle proteins UNC-22 from *C.elegans* (Benian et al., 1989) and turkey telokin (Holden et al., 1992). On the figure there are more grey-shaded residues that are buried inside the domain. ICAM-1 domain 3 does not seem to have many signature residues, but it does belong to the I-set Ig-like domain (Yang et al., 2004).

AN APPROACH TO DEFINE AN I-SET IG-LIKE DOMAIN USING THE SEQUENCE SIGNATURE

Fig. 3 collects a few typical I-set Ig-like domains, aligned to Dscam domain 4 based on the signature residues. Referring to this sequence alignment table a suggested approach to define an I-set Ig-like domain can be described as below.

(i) The most conserved YXC motif can be found at the F strand. This should be the first reliably recognized sequence signature for any Ig-like domain, as discussed above about Tyr-corner. It can be the first conserved sequence feature to be recognized for any Ig-like domain.

(ii) The second conserved Cys upstream along the sequence is at the B strand, which forms the disulfide bond with the Cys on the F strand mentioned above.

(iii) Six-residue N-terminal to the second C is an almost invariant G, a commonly seen amino acid at a β turn (Richardson, 1981). This Gly is located at the AB turn and hence initiates the B strand.

(iv) 11–12-residue C-terminal to the second C is the invariant W at the center of C strand, which forms part of the hydrophobic core discussed previously.

(v) Located between the second C and W should be the *cis*-P on BC loop, which structural role has been mentioned above.

(vi) Three-residue following the YXC motif is the N that acts as the bridge between BC loop and FG loop.

(vii) Roughly 10-residue downstream the sequence from this N is the Pho-X-Pho-X pattern at the G strand that marks the C-terminus of the domain.

(viii) N-terminal to the YXC motif a Pro-X-Pro-X-X-Pro sequence pattern defines the E strand and the β bulge.

(ix) About 23–26-residue N-terminal to the second C at the B strand, a Pro should be the second residue of the domain, which defines the N-terminus of the I-set Ig domain. There may be another conserved *cis*-P positioned 5–7-residue down the sequence at the A-A′ kink.

In conclusion, I-set Ig-like domain as a building block for cell surface receptors has a very conserved sequence signature. The signature residues allow for easy recognition and clear definition of the domain boundaries for an I-set Ig-like domain. The sequence signature is also largely shared with other types of Ig-like domains. This will help construct design and functional studies for IgSF proteins in general.

In addition, once each Ig-like module is defined within a multiple-domain cell surface receptor in tandem, it is possible to recognize whether the adjacent domains are immediately abutted with limited flexibility like the domain junction between ICAM-1's domains 1 and 2 (Wang and Springer, 1998) or there is a short linker in between to facilitate more relative movement between domains such as in CEACAM1a's domain junction (Tan et al., 2002). In the cases of hemolin (Su et al., 1998) and Dscam (Meijers et al., 2007), a 5–6-residue linker between domains 2 and 3 even allows for the domain 1–2 to bend over and touch domains 3–4 to form a unique horseshoe-like configuration at the N-terminus. All these different types of domain junction are likely to reflect their important functional requirement.

FOOTNOTES

This work has been supported by NIH grants and the support from the Ministry of Education of China.

This article does not contain any studies with human or animal subjects performed by the any of the authors.

REFERENCES

Benian, G.M., Kiff, J.E., Neckelmann, N., Moerman, D.G., and Waterston, R.H. (1989). Nature 342, 45–50.

Bork, P., Holm, L., and Sander, C. (1994). J Mol Biol 242, 309–320.

Chothia, C., and Jones, E.Y. (1997). Annu Rev Biochem 66, 823–862.

Freigang, J., Proba, K., Leder, L., Diederichs, K., Sonderegger, P., et al. (2000). Cell 101, 425–433.

Gao, G.F., Tormo, J., Gerth, U.C., Wyer, J.R., McMichael, A.J., et al. (1997). Nature 387, 630–634.

Harpaz, Y., and Chothia, C. (1994). J Mol Biol 238, 528–539.

Hemmingsen, J.M., Gernert, K.M., Richardson, J.S., and Richardson, D.C. (1994). Protein Sci 3, 1927–1937.

Holden, H.M., Ito, M., Hartshorne, D.J., and Rayment, I. (1992). J Mol Biol 227, 840–851.

Jones, E.Y., Davis, S.J., Williams, A.F., Harlos, K., and Stuart, D.I. (1992). Nature 360, 232–239.

Kern, P.S., Teng, M.-K., Smolyar, A., Liu, J.-H., Liu, J., et al. (1998). Immunity 9, 519–530.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., et al. (2001). Nature 409, 860–921.

Meijers, R., Puettmann-Holgado, R., Skiniotis, G., Liu, J.H., Walz, T.,et al. (2007). Nature 449, 487–491.

Newman, P.J., Berndt, M.C., Gorski, J., White, G.C., 2nd, Lyman, S.,et al. (1990). Science 247, 1219–1222.

Politou, A.S., Gautel, M., Pfuhl, M., Labeit, S., and Pastore, A. (1994). Biochemistry 33, 4730–4737.

Reid, R.A., and Hemperly, J.J. (1992). J Mol Neurosci 3, 127–135.

Richardson, J.S. (1981). Adv Protein Chem 34, 167–339.

Rudolph, M.G., Stanfield, R.L., and Wilson, I.A. (2006). Annu Rev Immunol 24, 419–466.

Sawaya, M.R., Wojtowicz, W.M., Andre, I., Qian, B., Wu, W.,et al. (2008). Cell 134, 1007–1018.

Su, X.D., Gastinel, L.N., Vaughn, D.E., Faye, I., Poon, P.,et al. (1998). Science 281, 991–995.

Tan, K., Zelus, B.D., Meijers, R., Liu, J.H., Bergelson, J.M., et al. (2002). EMBO J 21, 2076–2086.

Teichmann, S.A., and Chothia, C. (2000). J Mol Biol 296, 1367–1383.

Wang, J.-H., and Springer, T.A. (1998). Immunol Rev 163, 197–215.

Wang, R., Natarajan, K., and Margulies, D.H. (2009). J Immunol 183, 2554–2564.

Wu, H., Kwong, P.D., and Hendrickson, W.A. (1997). Nature 387, 527–530.

Yang, Y., Jun, C.-D., Liu, J.-H., Zhang, R., Joachimiak, A., et al. (2004). Mol Cell 14, 269–276.