Article No. mb982115 J. Mol. Biol. (1998) 283, 837–862

JMB

An Extracellular β-Propeller Module Predicted in Lipoprotein and Scavenger Receptors, Tyrosine Kinases, Epidermal Growth Factor Precursor, and Extracellular Matrix Components

Timothy A. Springer

Center for Blood Research and Harvard Medical School Department of Pathology 200 Longwood Avenue, Boston MA 02115 USA

An abundant, widely dispersed, extracellular sequence repeat that contains a consensus YWTD motif is shown here to occur in groups of six contiguous repeats. Thirteen lines of evidence, including experimental and computational data, predict with $p < 3 \times 10^{-9}$ that the repeats do not form tandem domains, but rather each group of six repeats folds into a compact β-propeller structure. The six β-sheets are arranged about a 6-fold pseudosymmetry axis, and each repeat contributes loops to the faces surrounding the pseudosymmetry axis. Seven different endocytic receptors that contain from one to eight YWTD β-propeller domains act as lipoprotein, vitellogenin, and scavenger receptors. In the low density lipoprotein receptor (LDLR), the many mutations in familial hypercholesterolaemia that map to the YWTD domain can now be interpreted. In the extracellular matrix component nidogen, the YWTD domain functions to bind laminin. Three YWTD domains and interspersed fibronectin type III (FN3) domains constitute almost the entire extracellular domain of the *sevenless* and c-*ros* receptor tyrosine kinases. YWTD domains often are bounded by epidermal growth factor (EGF) modules, including in the EGF precursor itself. YWTD β-propellers have a circular folding pattern that brings neighboring modules into close proximity, and may have important consequences for the architecture of multi-domain proteins.

Keywords: structure prediction; β-propeller; low density lipoprotein receptor; epidermal growth factor precursor; nidogen

Introduction

The vast majority of proteins on cell surfaces and in the extracellular matrix are modular. Modular or mosaic proteins contain a number of different domains or modules arranged in tandem on a single polypeptide chain (Bork *et al.*, 1996). Since each module may have a different enzymatic, signaling, ligand-binding, regulatory, or structural function, the modular architecture allows the evolution of proteins with complex and highly specialized functions. Approximately 60 types of extracellular modules have been identified based on sequence homology, and for about half of these the three-dimensional structure has been determined for at least one representative member. One of the most abundant of the 60 known extracellular repeats is 43 residues long and contains a Tyr-Trp-Thr-Asp sequence; i.e. YWTD in the one letter code (Bork *et al.*, 1996). YWTD repeats have important functions, and are found in functionally diverse surface and extracellular matrix proteins. The abundance of YWTD repeats is exceeded only by immunoglobulin, epidermal growth factor (EGF)-like, fibronectin type III (FN3), complement control, C-type lectin domains, and leucine-rich repeats (Bork *et al.*, 1996).

Abbreviations used: EGF, epidermal growth factor; FN3, fibronectin type III; IgSF, immunoglobulin superfamily; LDL, low density lipoprotein; LDLR, low density lipoprotein receptor; LDVR, very low density lipoprotein receptor; LRP, LDLR-related proteins; VLDL, very low density lipoprotein; S.D., standard deviation; pdb, Protein Data Bank.

E-mail address of the corresponding author: springer@sprsgi.med.harvard.edu

0022–2836/98/440837–26 $30.00/0

The epidermal growth factor (EGF) precursor and the low density lipoprotein receptor (LDLR) were among the first modular proteins to be described and also the first to reveal the YWTD repeat (Sudhof *et al*., 1985a,b; Doolittle, 1995). Both proteins contain multiple epidermal growth factor-like (EGF) repeats and clusters of YWTD repeats (Figure 1). The second cluster of YWTD repeats and its two flanking EGF repeats in the EGF precursor are 33% identical with the corresponding unit in the LDLR, a remarkable extent of identity for proteins with such divergent functions. The EGF hormone itself is derived from the most membrane-proximal EGF module by proteolysis of the membrane-bound EGF precursor (Parries *et al*., 1995). A third type of module present in the LDLR but not in the EGF precursor is known as the LDLR class A module (Figure 1). A very similar modular architecture is found in the very low density lipoprotein receptor (LDVR; Figure 1), and specificity for different classes of lipoproteins maps to the class A repeats (Brown & Goldstein, 1986; Brown *et al*., 1997). The crystal structure of one such class A module has recently been determined (Fass *et al*., 1997). This module contains acidic residues that are important for ligand binding; however, they coordinate a Ca^{2+} and thus appear to have a structural role rather than a direct role in ligand recognition (Brown *et al*., 1997; Fass *et al*., 1997). The three-dimensional structure of the EGF module is well known from several proteins (Bork *et al*., 1996).

By contrast to the LDLR class A and EGF repeats, nothing is known about the structure of YWTD repeats (Norton *et al*., 1990; Krieger & Herz, 1994; Brown *et al*., 1997). Even at the sequence level the repeats are not well studied; for example, the beginning and end of the repeat unit have not been defined. It is widely reported that there are five repeats per cluster, but a closer analysis reported here shows that there are six. The repeats have become known as ''YWTD spacers'' and are shown in current reviews as squiggly lines that serve to space apart other domains; however, as shown here the repeats assume a compact rather than extended conformation.

YWTD repeats are present in a diverse array of functionally important proteins, and where studied in detail have been found to be functionally important. The LDLR is critical for the cellular uptake of lipoproteins from plasma, and in the metabolism of cholesterol (Table 1) (Brown & Goldstein, 1986; Lestavel & Fruchart, 1994). Mutations in the LDLR cause familial hypercholesterolaemia, and accelerated atherosclerosis and coronary heart disease (Hobbs *et al*., 1992). The importance of the YWTD repeats is underscored by the observation that almost as many point mutations that cause familial hypercholesterolaemia map to this region as to the class A repeats (Soutar, 1992; Hobbs *et al*., 1990, 1992). Receptors that lack the YWTD repeats and surrounding EGF domains bind β-VLDL but not LDL, are defective in recycling, and fail to release β-VLDL after acidification (Davis *et al*., 1987).

Six other endocytic receptors have structures and functions related to the LDL receptor (Figure 1 and Table 1). These include two giant receptors termed LDLR-related proteins (LRP) 1 and 2. Besides lipoproteins, LRP1 and LRP2 scavenge many types of secreted proteins and proteinase:inhibitor complexes (Krieger & Herz, 1994; Farquhar *et al*., 1995). Remarkably, deletion of either LRP1 or LRP2 is lethal (Herz *et al*., 1992; Willnow *et al*., 1996).

Nidogen and osteonidogen are extracellular matrix components (Timpl & Aumailley, 1993; Figure 1 and Table 1). Nidogen binds to and is present in stoichiometric amounts with collagen IV and laminin, and organizes them into basement membranes. Nidogen contains three globular regions (Fox *et al*., 1991; Paulsson *et al*., 1986; Mann *et al*., 1988; Figure 1). A recombinant fragment containing globules 1 and 2 binds collagen IV. Globules 2 and 3 are connected by a stiff rod ~10.2 nm long that appears to correspond to the intervening four EGF-like modules (Figure 1). A globule 3 fragment containing only the YWTD repeats and the C-terminal EGF repeat is 4.8 nm in diameter and binds laminin with a K_D value of ~1 nM. Thus, the YWTD repeats in nidogen appear to have an important role in ligand binding, and assume a compact, globular structure (Fox *et al*., 1991). Circular dichroism of nidogen and its fragments suggests an absence of any significant α-helix and the presence of an all-β structure (Paulsson *et al*., 1986; Fox *et al*., 1991).

sevenless is a receptor tyrosine kinase important in photoreceptor development in *Drosophila* (Hafen *et al*., 1987; Krämer *et al*., 1991; Cagan *et al*., 1992; Table 1). The extracellular domain of *sevenless* is large, with 2100 residues. Previously, it has been found to contain seven fibronectin type III repeats, and only two, widely separated and hence unclustered YWTD repeats (Norton *et al*., 1990; outlined in black in Figure 1). The c-*ros* receptor tyrosine kinase is a vertebrate homologue of *sevenless* that gives rise to transforming genes isolated from human tumors and an avian sarcoma virus (Matsushime & Shibuya, 1990; Birchmeier *et al*., 1990; Riethmacher *et al*., 1994).

The sheer number of YWTD repeats emphasizes the importance that evolution has assigned to them. Not counting species homologues, 165 YWTD repeats are described above. This confirms an estimate in 1996 that placed the number at 140 and showed that the YWTD repeat is the seventh most abundant extracellular repeat (Bork *et al*., 1996).

Here, I show that YWTD repeats are always present as groups of six contiguous repeats, and predict that they fold into a compact structure known as a six-bladed β-propeller domain. β-Propellers are large, toroidal domains that contain six, seven or eight β-sheets arranged radially about a pseudo-symmetry axis (Murzin, 1992). Domains with four

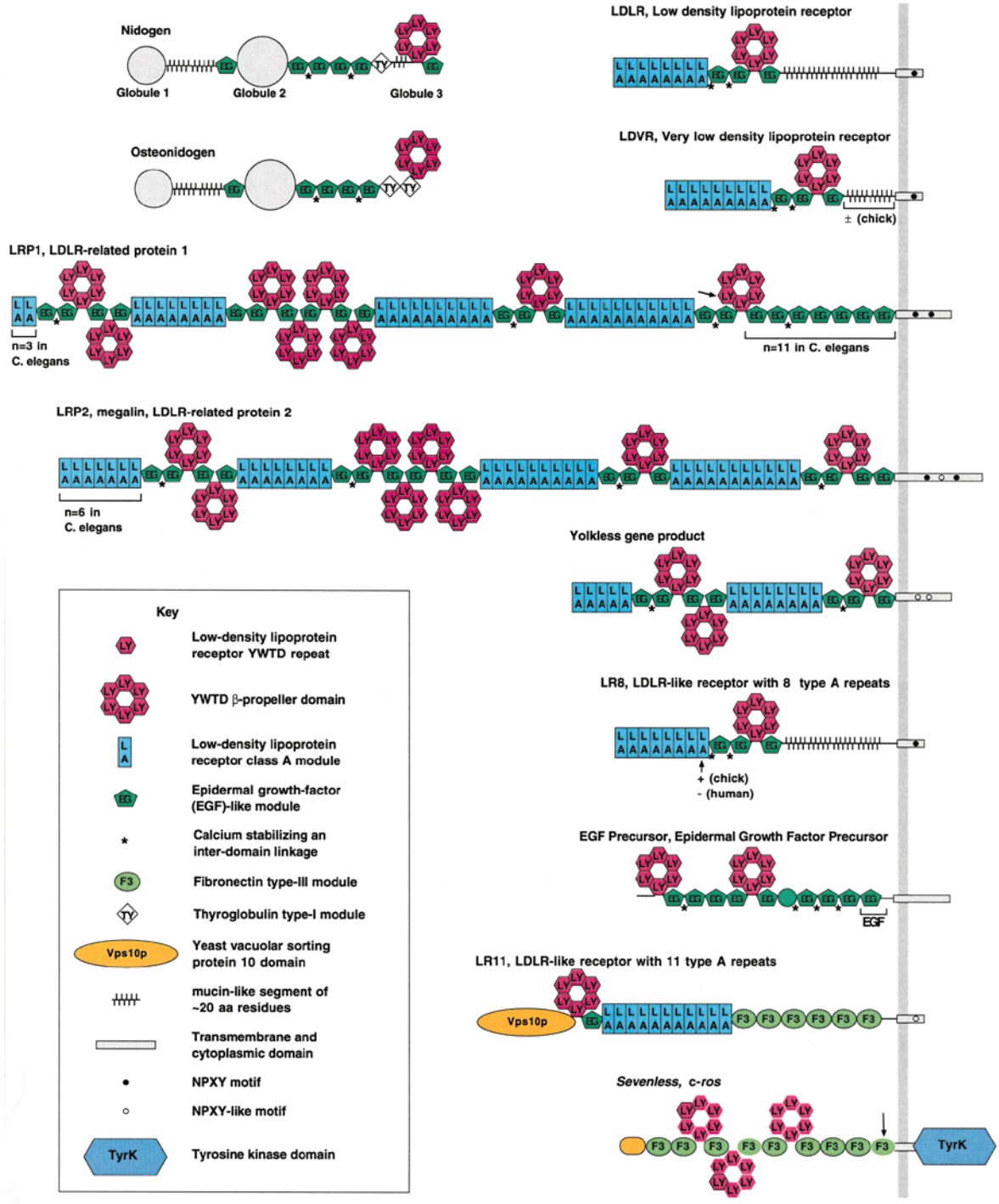

Figure 1. The modular organization of proteins containing YWTD repeats. This schematic representation is one-dimensional, with all proteins oriented with N termini to the left and C termini to the right. The plasma membrane is represented as a vertical bar, with the extracellular space to the left. The number (*n*) of groups of tandem modules is as shown for all known species, except where indicated otherwise. Module assignments were made based on analysis of database sequences, related annotations and publications (Bork *et al.*, 1996; Krieger & Herz, 1994; Farquhar *et al.*, 1995). The abbreviations and symbols for the modules are adapted from Bork & Bairoch (1995) and Bork *et al.* (1996) and are based on recommendations from an international workshop (Bork & Bairoch, 1995), except for the organization of the YWTD repeats into β-propeller-like modules. Arrows mark proteolytic processing sites in LRP1 and *sevenless*. FN3 domains or YWTD repeats in *sevenless* and c-*ros* identified previously (Norton *et al.*, 1990) and here are shown with and without black outlines, respectively.

β-sheets have also been termed β-propellers; however, this is a misnomer because an α-helix is present between each sheet, and these should be termed βα-propellers. β-Propeller domains have diverse enzymatic, ligand-binding, and regulatory functions. Interest in this fold has recently heigh-

Table 1. Proteins with YWTD domains

Name	Ligands	Major expression	Species
LDLR, Low density lipoprotein receptor	Lipoproteins containing ApoB-100 and ApoE	Liver, oocyte, cholesterol-regulated	Mammalia, Chondricthyes, Amphibia
LDVR, very low density lipoprotein receptor, vitellogenin receptor of chicken and *Xenopus*	Lipoproteins containing ApoE, vitellogenin	Muscle, heart, white fat, oocyte	Mammalia, Aves, Amphibia, *C. elegans*
LRP1, LDLR-related protein 1	Remnant lipoproteins enriched in ApoE, proteinase: α2- macroglobulin complexes, urokinase and plasminogen activator: plasminogen activator inhibitor 1 complexes, lipoprotein lipase, lactoferrin	Liver, brain, lung, ovary, placenta	Mammalia, Aves, *C. elegans*
LRP2, LDLR-related protein 2, GP330, Megalin	Lipoproteins enriched in ApoE, plasminogen activator, lipoprotein lipase, lactoferrin	Kidney	Mammalia, *C. elegans*
LR11, LDL-like receptor containing 11 type A repeats	Lipoproteins containing ApoE	Brain	Mammalia, Aves
LR8, LDL-like receptor containing 7 or 8 type A repeats	Lipoproteins containing ApoE	Brain	Mammalia, Aves
YL, *Yolkless*, vitellogenin receptor	Vitellogenin	Oocyte	*Drosophila*
Nidogen, entactin	Laminin, collagen IV, perlecan	Basement membranes	Mammalia, Ascidia
Nidogen-2, osteonidogen		Bone matrix	Mammalia
EGF, epidermal growth factor precursor	EGF receptor	Kidney, submaxillary gland	Mammalia
Sevenless	*Bride of sevenless*, *BOSS*	Photoreceptor R7 precursor cell	*Drosophilia*
c-*ros*, Proto-oncogene related to v-*ros* of avian sarcoma virus		Epithelia	Mammalia, Aves

tened with the discovery that the WD40 repeats of G protein β-subunits fold into seven-bladed propellers (Sondek *et al*., 1996; Lambright *et al*., 1996; Wall *et al*., 1995). Diverse types of signaling proteins contain WD40 repeats, and some are predicted to fold into six-bladed β-propellers (Neer *et al*., 1994; Saxena *et al*., 1996). β-Propellers have been predicted for the kelch protein in *Drosophila*, and the related actin-binding protein, scruin (Bork, 1994; Sun *et al*., 1997). The above proteins are located in the cytoplasm, but β-propeller enzymes, including viral and bacterial neuraminidases, are extracellular. Furthermore, a β-propeller domain has been predicted in the extracellular domain of integrin α-subunits (Springer, 1997). The prediction here that YWTD repeats assume a β-propeller fold is strongly supported by independent structural and genetic criteria, by computational molecular biology, and by experimental evidence from circular dichroism, electron microscopy, and disulfide bond topology. The prediction allows extensive data on molecules containing YWTD repeats and on mutations within YWTD repeats to be interpreted. β-Propellers fold with a circular topology; rather than acting as spacers, YWTD repeats bring neighboring modules into close proximity to one another (Figure 1). YWTD repeats may, therefore, not only have important functions on their own, but may also have important orienting and architectural functions in the modular proteins in which they are found.

Results

Distinguishing autonomous and interdependent sequence repeats

Sequence repeats can be divided into two groups. In autonomous repeats, there is one repeat per domain. The domains are largely structurally independent, and are arranged in tandem in the three-dimensional structure like beads on a string. The repeat corresponds to the modular unit. The Ig, FN3, and EGF domains are examples of this group. In interdependent repeats, multiple repeats fold into a single domain. The repeats cannot fold individually, because large hydrophobic interfaces are present between the repeating units. Therefore, multiple repeats are required to construct a modular unit. Examples are β-propellers and leucine-rich repeats.

In extracellular proteins, autonomous and interdependent repeats differ in two characteristics that are apparent from their sequences: cysteine content and hydrophobicity. Furthermore, certain types of interdependent repeats, such as those of β-propellers, are constrained in the number of repeats per module and hence in the number of contiguous sequence repeats. These characteristics were used previously to predict a β-propeller domain in integrins (Springer, 1997), and were tested here for their generality and for identification of further β-propeller domains. The 60 families of extracellular repeats that were identified in an excellent review by Bork and colleagues (Bork *et al*., 1996) served as the test set. The three-dimensional structures of about half of these modules are known, and this set can be used to deduce a ''cysteine-to-size rule'' that can be applied to proteins of unknown structure. Among the known structures, every autonomous repeat of 70 residues or less contains at least one disulfide bond. As domain size decreases below 70, there is a strong tendency for more disulfide bonds to be present. Larger protein domains are stabilized by their hydrophobic cores; as the size of this core decreases, covalent cross-links through disulfide bonds become increasingly important for stability. By contrast, repeats of similar length that fold interdependently can lack cysteine residues, because their domains have large hydrophobic cores.

The compilation of Bork *et al*. (1996) reveals only three repeats of 70 residues or less that contain, on average, less than two cysteine residues per repeat, and that therefore may be expected to fold cooperatively into larger domains. Two are structurally characterized. The hemopexin-like repeats of 60 residues fold into the special four-bladed βα-propellers and form domains of ~240 residues. The leucine-rich repeats of 20 residues contain one α-helix and one β-strand per repeat, and fold cooperatively into a horseshoe-shaped domain of ~300 residues. The cysteine-to-size rule is thus supported by 14 of 14 autonomous repeats and two of two interdependent repeats of known structure (Bork *et al*., 1996). The YWTD repeat stands out as the only structurally uncharacterized repeat of $\leqslant$70 residues with a paucity of cysteine residues. The 534 YWTD repeats examined here contain, on average, 46 residues and only 0.3 cysteine per repeat. By contrast, all autonomous repeats of 40 to 50 residues contain four to six cysteine residues (Bork *et al*., 1996). This supports the hypothesis that YWTD repeats fold interdependently into a domain that contains multiple YWTD repeats.

YWTD repeats occur in groups of six

Previously, the boundaries of YWTD repeats have not been defined, and the repeats have been reported to be present in groups of five per cluster. The boundaries of the YWTD repeats can be rigorously determined with reference to the flanking EGF domains, since three-dimensional structures are known for tandem EGF modules (Brandstetter *et al*., 1995; Downing *et al*., 1996). Thus, the boundaries of the EGF modules are two or three residues before their first cysteine, and two or three residues after their last cysteine (Figure 2B). Therefore, the YWTD repeats are defined, e.g. in the LDLR as beginning with residue 396 and ending with residue 664 (Figure 2A). Aligning YWTD repeats 2 to 6 shows that the YWTD motif is at the beginning of each repeat (Figure 2A). This reveals an additional repeat that follows the N-terminal EGF domain and precedes the five repeats with clear

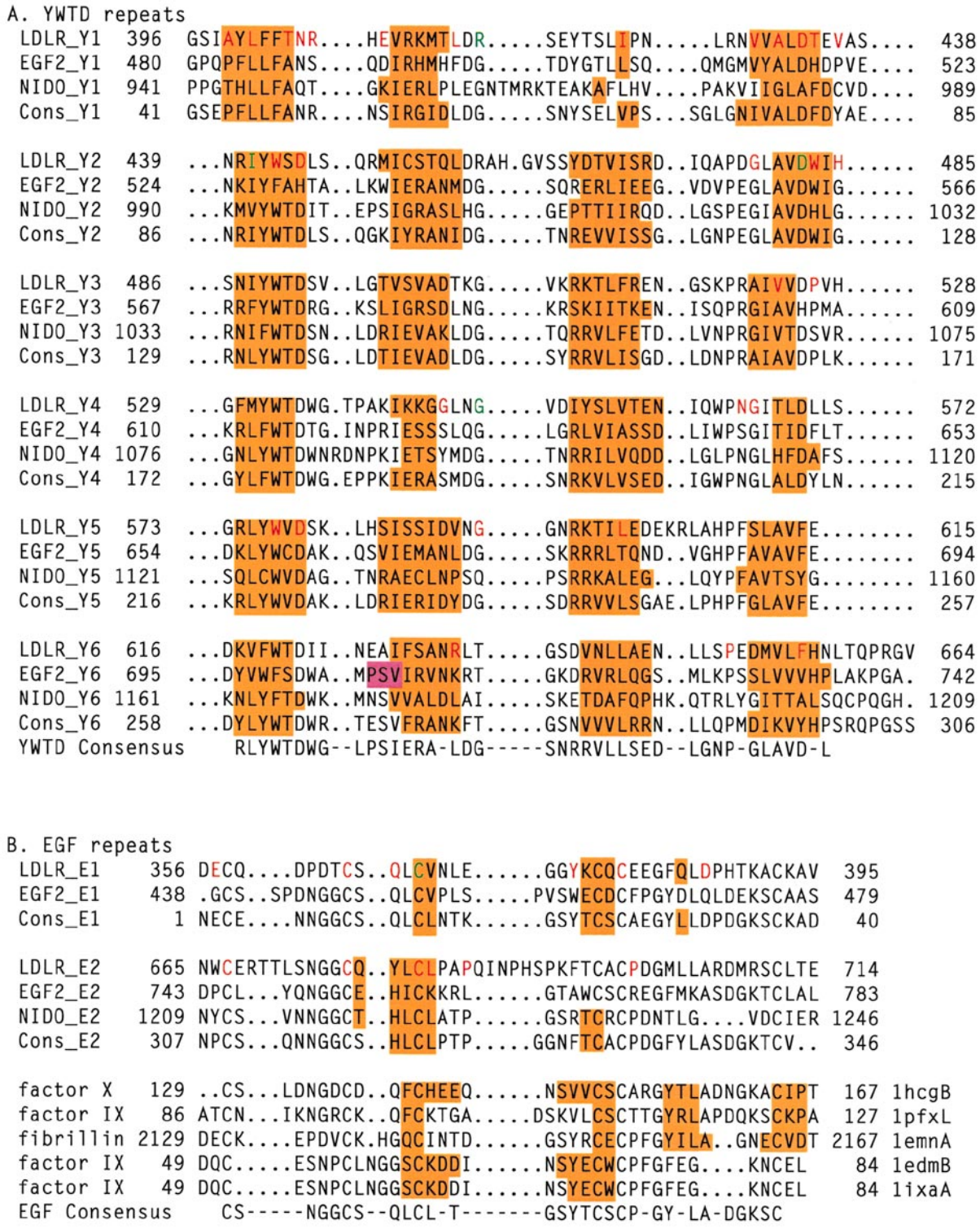

Figure 2. The YWTD repeats, flanking EGF domains, and predicted secondary structures. The secondary structure of YWTD domains (A) and their flanking EGF domains (B) was predicted using PHD (Methods). Sequences are from the LDLR, the second YWTD domain in EGF precursor (EGF2), nidogen (NIDO), and a consensus of 89 YWTD domains (cons), and are for YWTD repeats 1 to 6 (Y1 to Y6) and N-terminal (E1) and C-terminal (E2) flanking EGF domains. LDLR residues with one or two distinct missense mutations in familial hypercholesterolaemia are shown in red or green, respectively (see the legend to Figure 8). Residues predicted to be in β-strand or α-helix with PHD (Rost, 1996) are highlighted in gold and magenta, respectively. Prediction of β-strand 4 of the second EGF domain was impaired by truncation of the sequence alignments one residue after the last cysteine, as shown for Cons_E2 (Figure 4B). EGF domains of known structure with the greatest sequence homology to Cons_E1 or Cons_E2 found with a BLAST search and SCOP (Murzin *et al.*, 1995) are aligned at the bottom of B. The domains are ordered from highest homology at top ($p = 2 \times 10^{-10}$) to lowest at bottom ($p = 5 \times 10^{-6}$). β-Strands in the known structures as determined with DSSP (Kabsch & Sander, 1983) are highlighted in gold. PDB and chain identifiers are shown for the EGF domain structures.

YWTD motifs. This repeat, repeat 1 (Figure 2A) is similar in length to the other repeats, and except for the YWTD motif itself, is as homologous to the other repeats as they are to themselves. In place of YWTD (Tyr-Trp-Thr-Asp) are found similar residues, i.e. LFAN (Leu-Phe-Ala-Asn) as the consensus, and FFTN (Phe-Phe-Thr-Asn) for human LDLR (Figure 2A). Furthermore, a hydrophobic residue precedes the YWTD-like motif both in repeat 1 and in repeats 2 to 6 (Figure 2A). Because of the homologies throughout the lengths of all six repeats, and since even for repeats 2-6 the YWTD motif is not absolutely conserved, it is appropriate to refer to all six repeats as YWTD repeats. A contiguous group of six YWTD repeats will be referred to here as a YWTD domain.

For repeats that fold autonomously, i.e. with a single repeat per domain, the number of sequence repeats found in tandem is not constrained. For example, among the proteins reviewed here the number of contiguous, tandem LDLR class A, EGF, and FN3 repeats is highly variable (Figure 1). By contrast, YWTD repeats are always found in contiguous groups of six (Figure 1). This pattern has been maintained, since the divergence of nematodes and chordates, in at least ten different proteins with varied functions, and in 27 instances of contiguous repeats in these proteins not counting species homologues (Figure 1). The stringent maintenance of six contiguous repeats argues strongly that these repeats are not autonomous, but are structurally constrained to fold into a single domain.

Exon structure suggests that six YWTD repeats form a modular unit

In modular proteins, exon boundaries and phases are highly correlated to the modular units (Patthy, 1987; Barclay *et al.*, 1993; Bork *et al.*, 1996; Doolittle, 1995; Long *et al.*, 1995). Phase refers to the position of intron insertion within the codon; e.g. after nucleotide 1 for phase 1 introns. The major mechanisms of exon shuffling, i.e. exon insertion and exon duplication, ''involve modules that have introns of the same phase class at both their 5′- and 3′ ends'' (Patthy, 1987). This allows modular units corresponding to autonomously folding domains to be inserted or deleted without changing the reading frame. The most frequently occurring modules within extracellular proteins, i.e. the immunoglobulin, EGF-like, fibronectin type-III, complement control, and C-type lectin domains, as well as the LDL-receptor class A modules, all have phase 1 introns. This allows exon shuffling to be used for the construction of mosaic proteins containing various mixtures of these modules. Comparison between the LDL receptor and EGF precursor provided one of the earliest examples of exon shuffling (Sudhof *et al.*, 1985a,b), and exon shuffling has no doubt been important in the evolution and proliferation of proteins containing YWTD repeats (Figures 1 and 3). Indeed, the

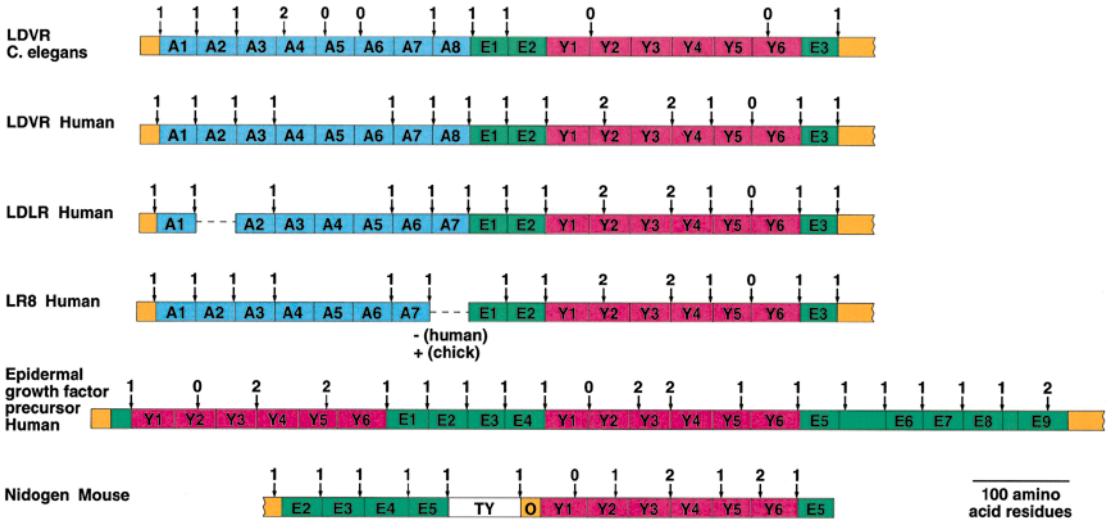

Figure 3. Exon structure of proteins containing YWTD domains. Intron positions are marked with arrows with the phase of the intron above, e.g. phase 1 introns split the codon after nucleotide 1 (Patthy, 1987). Broken lines show the positions of class A exons, that based on sequence homology between LDVR, LDLR, and LR8, appear to have been shuffled out during evolution. Protein domains or repeats and intron positions are shown to scale. N terminus is to left and C terminus to right. Regions are A, LDL receptor class A repeats; E, EGF repeats; O, *O*-glycosylated; TY, thyroglobulin repeat; and Y, YWTD repeats. Boundaries of class A repeats were taken as halfway between the last cysteine of one and the first cysteine of the next; boundaries of other repeats are as described for Figure 2. The signal sequence is shown in yellow to the left; C-terminal mucin-like, transmembrane, and cytoplasmic domains, and the N-terminal portion of nidogen are omitted, as symbolized by yellow segments with a jagged end. Exon-intron boundaries are for human LDLR (Sudhof *et al.*, 1985a), human LDVR (Sakai *et al.*, 1994), human LR8/7 (Kim *et al.*, 1997), *C. elegans* LDVR beginning from nucleotide 16,738 of cosmid T13C2 (Methods); human EGF precursor (Bell *et al.*, 1986); and mouse nidogen (Durkin *et al.*, 1995).

class A modules of the LDVR, LDLR, and LR8 genes provide clear examples of this (broken lines represent ''deleted'' exons in Figure 3). The great majority of LDL class A, EGF, and FN3 modules are flanked near their domain boundaries by phase 1 introns, and thus can readily be shuffled. By contrast, not one of the 42 different YWTD repeats illustrated in Figure 3 or any other known YWTD repeat is flanked by introns of compatible phases. Thus, shuffling of individual YWTD repeats cannot occur. However, in vertebrates phase 1 exons flank each group of six YWTD repeats (Figure 3). In *Caenorhabditis elegans*, exons are often more condensed. In the LDVR homologue in *C. elegans*, the six YWTD repeats plus their bordering EGF domains are flanked by phase 1 exons (Figure 3). Thus, the group of six YWTD repeats, or the group of six YWTD repeats plus the two flanking repeats, is the modular unit that can readily be shuffled. A concordance of exon structure with domain structure and not sequence repeats is found for domains composed of multiple sequence repeats. Thus, leucine-rich repeats fold cooperatively into a single domain, and sequence repeat boundaries do not correlate with intron boundaries and introns differ in phase (Barclay *et al*., 1993). Similarly, in seven-bladed β-propellers known in the G-protein β-subunit (van der Voorn *et al*., 1990) or predicted in integrin α-subunits (Corbi *et al*., 1990; Springer, 1997), seven readily discernible sequence repeats are present but do not correlate with exon boundaries. The clear prediction from exon structure is that the group of six YWTD repeats, not an individual YWTD repeat, is the unit that can be shuffled in evolution. This provides further support for the hypothesis that the six repeats fold up into a single domain.

The six-bladed β-propeller domain hypothesis

The findings that YWTD repeats are always found in groups of six, that compatible exon phases are only found for the group of six repeats, and that the six YWTD repeats appear as a globule rather than an extended structure in electron microscopy (Fox *et al*., 1991), all suggest that they fold into a single domain. The sequence homology between repeats 1 to 6 strongly suggests that each repeat has a similar structure. Therefore, the repeats should fold into a domain with six similar structural units. Several folds are known that have 2- or 3-fold pseudosymmetry, e.g. γ-crystallin, β-trefoil, and β-prism I and II folds (Murzin *et al*., 1995). These might exist in tandem in groups of three or two to give six similar units; however, this would not be compatible with the universal finding of six contiguous repeats and the exon structures. Only one fold is known in which a single polypeptide chain folds into a single domain with six structurally similar units; this is the six-bladed β-propeller (Murzin *et al*., 1995). β-Propellers are large domains that contain six, seven or eight antiparallel β-sheets (Murzin, 1992) (Figure 4). Each β-sheet of a β-propeller has an almost identical tertiary structure, and in many β-propellers known at atomic resolution, this is reflected in amino acid sequence repeats. These sequence repeats range from 40 to 60 residues in length. The sheets in β-propellers are arranged radially about a pseudosymmetry axis, yielding a compact domain that is cylindrical or toroidal in shape (Figure 4). Each sheet has four β-strands. The β-strand closest to the pseudosymmetry axis, strand 1, is almost parallel to this axis, but the inherent twist of β-sheets results in strand 4 being almost perpendicular to the pseudosymmetry axis, giving each sheet the appearance of a propeller blade.

Secondary structure prediction

Secondary structure was predicted to test the β-propeller domain hypothesis. All β-propellers contain four antiparallel strands per β-sheet. Multiple alignments were submitted to the PHD server (Rost, 1996) to predict the secondary structure of segments of 340 residues including the YWTD repeats and flanking EGF domains for four representative sequences (Figure 2). Automatic sequence alignment with PRRP (Gotoh, 1996) revealed four sequence blocks per YWTD repeat (Figure 2A). Remarkably, a β-strand was predicted near the middle of each sequence block. The β-propeller hypothesis predicts four β-strands in each of six repeats; this was in perfect agreement with the predictions, with 24 out of 24 expected β-strands predicted in each of four sequences (Figure 2A). Predictions for the EGF domains were in good agreement with the structures determined for the EGF domains most homologous in sequence to those that flank YWTD domains (Figure 2B).

The secondary structure and solvent accessibility predicted by PHD for the four YWTD domain sequences (Figure 2A) were aligned with the secondary structure and solvent accessibility of a database of structures with TOPITS, to predict protein fold (Rost, 1995). This method is distinct from potential-based threading (see below). The top hit with TOPITS was the G protein β-subunit β-propeller, and seven of the eight top hits were β-propeller domains (Table 2). Thus, the secondary structure and solvent accessibility predictions for YWTD domains suggest a β-propeller fold.

The *sevenless* and c-*ros* tyrosine kinases

Previously, only two YWTD repeats have been reported in the *sevenless* tyrosine kinase in *Drosophila*, and these are 400 residues apart in the amino acid sequence (Norton *et al*., 1990). By contrast, the hypothesis that YWTD repeats fold into six-bladed β-propellers predicts that YWTD repeats should be contiguous, and present in groups of six. Therefore, *sevenless* and its vertebrate homologue, the c-*ros* tyrosine kinase, were tested with PairWise (Birney *et al*., 1996) for regions of homology with a profile of the 89 YWTD domains described above.

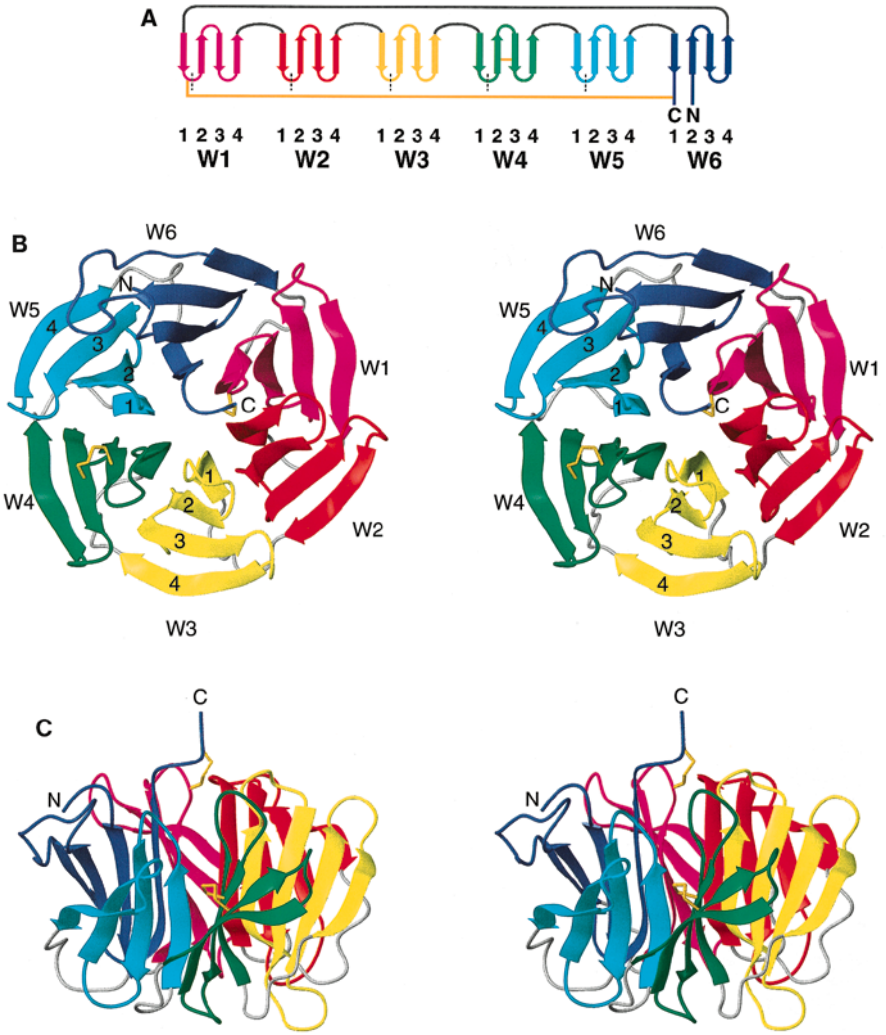

Figure 4. Topology and ribbon diagrams of the six-bladed β-propeller domain predicted for the YWTD domain of nidogen. A, Topology diagram, with each β-sheet given a different color. Sequence repeats are separated by vertical broken lines. Note the offset between sequence repeats and β-sheets. Each sheet is termed a W, with each β-strand representing one leg of the W. β-Strands are represented by arrows. The disulfide bonds in nidogen between strands 2 and 3 of W4 and between the 1-2 loop of W1 and the C-terminal segment in W6 are shown as gold lines. B, Stereo view ribbon representation, with each β-sheet or blade of the nidogen β-propeller model given the same color code as in A. The 4-1 loops that connect each sheet are grey. The view is up the 6-fold pseudosymmetry axis, with the "bottom" of the propeller containing the strand 1 to 2 loops, the strand 3 to 4 loops, and the N and C termini of the propeller domain in the foreground. C, Side view, with the 1-2, 3-4, and N and C termini upward. The side-chain bonds for the cysteine residues in the two disulfide bonds of the nidogen β-propeller domain are shown in gold. The β-strands shown as ribbons are as defined by DSSP (Kabsch & Sander, 1983) from the nido model (see Figure 7 and Table 6). Prepared with MOLMOL (Koradi *et al.*, 1996).

Remarkably, three different regions each containing six contiguous YWTD repeats were identified (Figure 5A). The same three regions were identified with the consensus sequence of the 89 YWTD domains in the second iteration of a PSI-BLAST search with expectation values ranging from 10^{-35} to 10^{-4} (Altschul *et al.*, 1997). The individual YWTD repeats in *sevenless* and c-*ros* are more divergent from one another but nonetheless are readily identifiable (Figure 5A). As an independent means of determining module boundaries, the receptor sequences were examined for fibronectin type III (FN3) repeats, since seven FN3 repeats have been reported to be present in *sevenless* (Norton *et al.*, 1990). Two additional FN3 repeats were found (Figure 5B). The newly identified repeats, repeats 4 and 9, are significantly related to the other repeats ($p < 10^{-7}$ and $<2.1 \times 10^{5}$, respectively), as shown using sequence motifs (Bailey, 1995) constructed from repeats 1 to 3 and 5 to 8.

The FN3 repeat boundaries confirm the YWTD module boundaries in *sevenless* and c-*ros*, since each YWTD module is flanked by FN3 modules (Figure 1). Indeed, the three YWTD domains and the nine FN3 domains together define a contiguous segment of ~1800 residues in c-*ros* and *sevenless* (Figure 5A and B). The arrangement of the multiple YWTD modules in c-*ros* and *sevenless* is

Table 2. Fold prediction for YWTD domains with TOPITS

Rank	Z-score	Structure	
1	2.84	1got	7-blade β-propeller, G protein β-subunit
2	2.83	1nir	8-blade β-propeller, cd1 nitrite reductase
3	2.73	1fnf	FN3 domains 7-10, fibronectin
4	2.68	2bbk	7-blade β-propeller, methylamine dehydrogenase
5	2.51	1gof	7-blade β-propeller, galactose oxidase
6	2.46	2sil	6-blade β-propeller, *Salmonella typhimurium* sialidase
7	2.41	1nsc	6-blade β-propeller, Influenza virus Beijing neuraminidase
8	2.38	1nnc	6-blade β-propeller, avian Influenza virus neuraminidase

The PHD predictions for the YWTD domains for the four sequences shown in Figure 2A were submitted to TOPITS (Rost, 1995). The top 20 hits were received for each prediction. Redundant hits for structures with the same sequence were removed. The *Z*-scores for structures common to all four predictions were averaged.

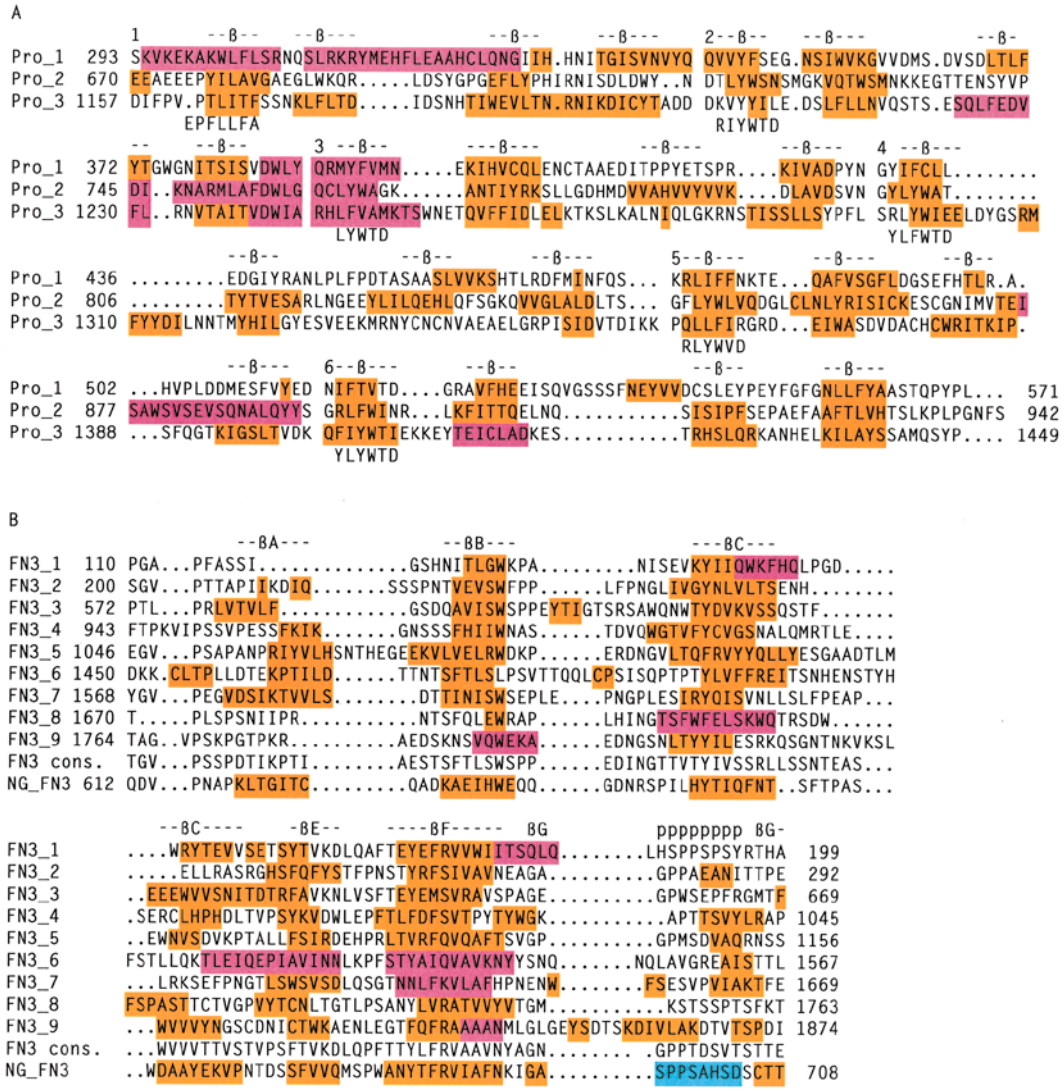

Figure 5. The YWTD and FN3 domains of *sevenless* and c-*ros*. A and B, Alignment and secondary structure prediction of the YWTD domains (A) and FN3 domains (B) of chicken c-*ros*. Residues predicted to be β-strand or α-helix are highlighted in gold and magenta, respectively. Residues in FN3 domain 1 of neuroglian (NG_FN3) (Huber *et al.*, 1994) known to be β-strand or polyproline II helix are highlighted in magenta and cyan, respectively. The boundaries of FN3 repeats 1 to 3 and 5 to 8 are by homology to the repeats defined for *D. melanogaster sevenless* (Norton *et al.*, 1990). The boundaries of FN3 repeats 4 and 9, and the YWTD domains, were defined as described in the text. The entire extracellular region of chicken c-*ros* was submitted at the top of an alignment with other vertebrate c-*ros* sequences to PHD (Rost, 1996) for secondary structure prediction. Above the alignment in A are putative β-strands and the YWTD repeat number; below the alignment are the consensus sequences for YWTD motifs 1 to 6 in the LDLR-related proteins (Figure 2A). Above the alignment in B are β-strand and polyproline II helix (p) assignments for FN3 domain 1 of neuroglian.

remarkably like that of the other proteins shown in Figure 1, except that FN3 modules take the place of EGF modules. Furthermore, the number of adjacent YWTD repeats is always six, whereas the number of adjacent EGF or FN3 modules is variable.

Secondary structure was predicted for c-*ros* using PHD. The information content of the c-*ros* and *sevenless* sequence alignments is low, because the percentage identity in extracellular domains between vertebrate c-*ros* sequences is >50%, and between *D. melanogaster* and *D. virilis sevenless* is 60% (Michael *et al.*, 1990). The ~20% identity between vertebrate and insect sequences, or between different YWTD domains in the same protein (Figure 5A) is too low for inclusion in a common alignment for purposes of secondary structure prediction. Therefore, secondary structure was predicted independently for each of the three YWTD modules in chicken c-*ros* (Figure 5A), and the three modules were then aligned together for comparison of the predictions. Out of 24 positions expected in a six-bladed β-propeller, β-strand was predicted for at least two of the three modules in 21 instances, supporting the β-propeller hypothesis. The secondary structure of the nine FN3 domains was independently predicted at the same time (Figure 5B). The consensus of the nine chicken c-*ros* FN3 repeats is 27% identical in sequence ($E < 0.002$) with FN3 domain 1 from neuroglian, a *Drosophila* neural cell adhesion molecule (Huber *et al.*, 1994). The sequence alignment with this structure shows that the consensus for prediction of β-strands by PHD in the c-*ros* FN3 repeats is in excellent agreement with the known position of β-strands in neuroglian.

During biosynthesis of *sevenless*, proteolytic cleavage occurs at a sequence of nine contiguous arginine residues, resulting in two chains of 220 kDa and 58 kDa that remain non-covalently associated (arrow, Figure 1; Simon *et al.*, 1989). This site is predicted to be not between domains, but within the highly extended C-C′ loop in FN3 domain 9 (not shown), in agreement with non-covalent association of the chains.

Fold prediction by threading

"Threading" is a method for predicting the tertiary structure or fold of a protein. A sequence is aligned with or "threaded through" each structure in a database. The sequence-structure alignments are completely analogous to sequence-sequence alignments, including provision for gaps or insertions, but what is calculated is the energy or statistical potential of the test sequence in a given three-dimensional structure. The lowest energy alignment that can be found for the test sequence is calculated for each structure in the database. Each of the 89 YWTD domains identified here was threaded through a non-redundant database containing over 1900 structures with THREADER 2.1 (Jones *et al.*, 1995). Because homologous sequences adopt the same fold, averaging of threading results for sequence homologues improves prediction accuracy (Edwards & Perkins, 1996). I caution that with smaller domains such as IgSF domains, averaging is even more important than is evident with YWTD domains (data not shown). An evolutionary tree prepared with PHYLP (Gotoh, 1996) clustered the 89 YWTD domains into 14 subfamilies and one left-over group of the 12 most divergent sequences (Table 3). For each of these 15 groups of YWTD domains, the highest threading score was with the β-propeller domain of the G protein β-subunit, 1gotB1 (Table 3). The average *Z*-score of 1gotB1 for the 89 YWTD domains was 3.97, and was far above the *Z*-scores for all other structures; the next highest score was 2.11 (Table 3 and Figure 6A). The transducin β-subunit is a seven-bladed β-propeller domain. A six-bladed β-propeller modeling template made from the transducin β-subunit (see below) was added to the threading database and gave an even higher average *Z*-score of 4.07. With this threading program and number of structures in the database, a *Z*-score > 3.5, even for a single sequence is considered to be "very significant – probably a correct prediction." The large number of YWTD sequences examined here gives even greater weight to this conclusion.

YWTD domains 1, 2, and 3 from the two insect *sevenless* and four vertebrate c-*ros* sequences yielded *Z*-scores of 3.71, 3.38, and 4.27, respectively, with the transducin β-propeller domain, 1gotB1 (Table 3). The average Z-score of 3.79 for all 18 sequences was far higher than the scores for all other structures (Table 3 and Figure 6B). The sequence identity between the lipoprotein receptor-related YWTD domains and those in *sevenless* is low. Therefore, threading results (1) identify a β-propeller fold for each of the three YWTD domains in *sevenless* and c-*ros* with a high level of confidence, and (2) the agreement with the largely independent results on the 89 lipoprotein receptor-related YWTD domains provides an even higher level of confidence in the β-propeller fold.

Averaging *Z*-scores for each structure for groups of sequences tended to lower the *Z*-score for the second highest hit and decreased the S.D. of the distribution of average *Z*-scores. By definition, *Z*-scores should be normalized to the S.D. The normalized average *Z*-scores with the G protein β-propeller domain for the groups of 89 and 18 sequences were above 4 (Table 4), and the expectation value for obtaining such a good hit among a database of ~2000 structures in the absence of a structural relationship was 1.6×10^{-3} and 7.2×10^{-3}, respectively. Furthermore, the two-tailed *t* test was used to determine whether the distribution of *Z*-scores for 1gotB1 with all 107 YWTD sequences was significantly different from that for the next most ten high-scoring structures (the top and second hit are compared in Figure 6C). The *p* values ranged from 2×10^{-62} to 4×10^{-51}. For a database of 2000 structures, the expectation value for finding such a difference by chance alone is

Table 3. Threading scores for subfamilies of YWTD domains

Subfamily[a]	*n*	Top hit Code[b]	Top hit Z-score	2nd hit Code[b]	2nd hit Z-score
lrp_7[c]	5	1gotB1	4.00	1pkyA2	2.72
lrp_3[d]	7	1gotB1	4.31	2aaiB0	2.57
ldlr,ldvr,lr8[e]	16	1gotB1	3.60	1scs00	2.19
egf_2[f]	3	1gotB1	3.95	1pkyA2	2.89
lrp_8[g]	5	1gotB1	4.49	1edt00	2.91
lr11,yl_1[h]	4	1gotB1	3.18	1pkyA2	2.98
lrp_ce7[i]	3	1gotB1	3.78	2dri00	2.90
lrp_2[j]	5	1gotB1	3.43	1abrB0	2.55
lrp_5[k]	5	1gotB1	3.92	1mat00	2.39
lrp_4[l]	5	1gotB1	4.20	1scs00	2.40
nido[m]	4	1gotB1	3.81	1treA0	2.74
lrp_1[n]	6	1gotB1	4.26	1abrB0	2.53
lrp_6[o]	6	1gotB1	4.17	1scuA0	2.76
egf_1[p]	3	1gotB1	4.39	1abrB0	2.83
divergent[q]	12	1gotB1	4.20	1plq00	1.98
89 ywtd	89	1gotB1	3.97	2aaiB0	2.11
sev, ros_1[r]	6	1gotB1	3.71	1mdr02	2.34
sev, ros_2[r]	6	1gotB1	3.38	2dri00	2.58
sev, ros_3[r]	6	1gotB1	4.27	1pkyA2	2.30
18 ywtd	18	1gotB1	3.79	2dri00	2.27
All YWTD	107	1gotB1	3.94	2aaiB0	1.94

Z-scores with THREADER 2.1 are for combined pairwise and solvation energies.

"Experience has shown the following interpretation to be useful:" (THREADER 2 User guide; Jones *et al.*, 1995). $Z > 3.5$, very significant, probably a correct prediction; $Z > 2.9$, significant, good chance of being correct; $2.7 < Z < 2.9$, borderline significant, possibly correct; $2.0 < Z < 2.7$, poor score, could be right, but needs other confirmation; $Z < 2.0$, very poor score, probably there are no suitable folds in the library.

[a] Subfamilies are listed in order of increasing divergence from the consensus of 89 YWTD sequences. Sequences are defined in footnotes c to r and sequence names in Methods.

[b] THREADER database structures are coded with the FOUR character pdb code followed by one character for the chain identifier (0 if null) and one number for the domain (0 if entire chain). 1gotB1, transducin β-subunit β-propeller domain, residues 45 to 340; 2aaiB0, ricin β-trefoil domain; 1pkyA2, pyruvate kinase TIM-barrel domain; 1scs00, concanavalin A β-sandwich domain; 1edt00, endo-β-*N*-acetylglucosaminidase TIM-barrel domain; 2dri00, D-ribose-binding protein, α/β periplasmic protein-like domain; 1abrB0, abrin β-trefoil domain; 1mat00, methionine aminopeptidase α + β fold; 1treA0, triosephosphate isomerase TIM-barrel domain; 1scuA0, succinyl-Coa synthetase α/β NAD-binding and flavodoxin-like domains; 1plq00, proliferating cell nuclear antigen, two α + β DNA clamp domains; 1mdr02, mandelate racemase TIM-barrel domain.

[c] lrp_7: LRP1_CH7, LRP1_HU7, LRP1_MO7, LRP2_HU7, LRP2_RA7.

[d] lrp_3: LRP1_CE3 LRP1_CH3, LRP1_HU3, LRP1_MO3, LRP2_HU3, LRP2_RA3, LRP_CAE3.

[e] ldlr, ldvr, lr8: LDL1_XE1, LDL2-XE1 LDLR-CP1, LDLR_CR1, LDLR_MO1, LDLR_HU1, LDLR_RB1, LR8_CH1, LR8_HU1, LDVR_CH1, LDVR_HU1, LDVR_MO1, LDVR_RA1, LDVR_RB1, LDVR_XE1, LDVR_CE1.

[f] egf2: EGF_HU2, EGF_MO2, EGF_RA2.

[g] lrp_8: LRP1_CH8, LRP1_HU8, LRP1_MO8, LRP2_HU8, LRP2_RA8.

[h] lr11, yl_1: YL_DM1, LR11_CH1, LR11_HU1, LR11_RB1.

[i] lrp_ce7: LRP_CAE7, LDR_CE1, LRP1_CE7.

[j] lrp_2: LRP1_CH2, LRP1_HU2, LRP1_MO2, LRP2_HU2, LRP2_RA2.

[k] lrp_5: LRP1_CH5, LRP1_HU5, LRP1_MO5, LRP2_HU5, LRP2_RA5.

[l] lrp_4: LRP1_CH4, LRP1_HU4, LRP1_MO4, LRP2_HU4, LRP2_RA4.

[m] nido: NIDO_HR1, NIDO_HU1, NIDO_MO1, ONID_HU1.

[n] lrp_1: LRP1_CH1, LRP1_HU1, LRP1_MO1, LRP2_HU1, LRP2_RA1, LRP_CAE1.

[o] lrp_6: LRP1_CE6, LRP1_CH6, LRP1_HU6, LRP1_MO6, LRP2_HU6, LRP2_RA6.

[p] egf_1: EGF_HU1, EGF_MO1, EGF_RA1.

[q] Divergent: LRP1_CE2, LRP_CAE5, LRP_CAE6, LRP_CAE8, LRP_CAE4, LRP_CAE2, LRP1_CE8, YL_DM3, YL_DM2, LRP1_CE5, LRP1_CE4, LRP1_CE1.

[r] sev, ros: ywtd domains 1, 2, or 3 are from SEV_DM, SEV_DV, SEV_CH, SEV_HU, SEV_RA, SEV_MO.

$<10^{-47}$. By contrast, the *p* values for comparison among hits 2 to 11 ranged from 0.025 to 0.994, giving expectation values of 50 to 2,000.

Above, nine FN3 repeats were predicted by sequence homology to be present in *sevenless* and c-*ros*. The two newly predicted repeats, FN3 repeats 4 and 9, were subjected to threading (Table 5). In all cases, the top hit was a FN3 domain, and for FN3_4 of *sevenless* the second hit was also a FN3 domain. The *Z*-scores were well above the score of 3.5 considered to be highly significant, and were far above scores for non-FN3 domains (Table 5).

Three-dimensional models

Models were built to further test the prediction of the six-bladed β-propeller domain, and to make predictions about YWTD domain structure and function. Better results were obtained using the G protein β-subunit than bacterial sialidases as template. The orientation between neighboring β-sheets, although not their number, is conserved in β-propellers (Murzin, 1992); this served as the basis for a novel method for constructing a template from the G protein β-subunit that possessed

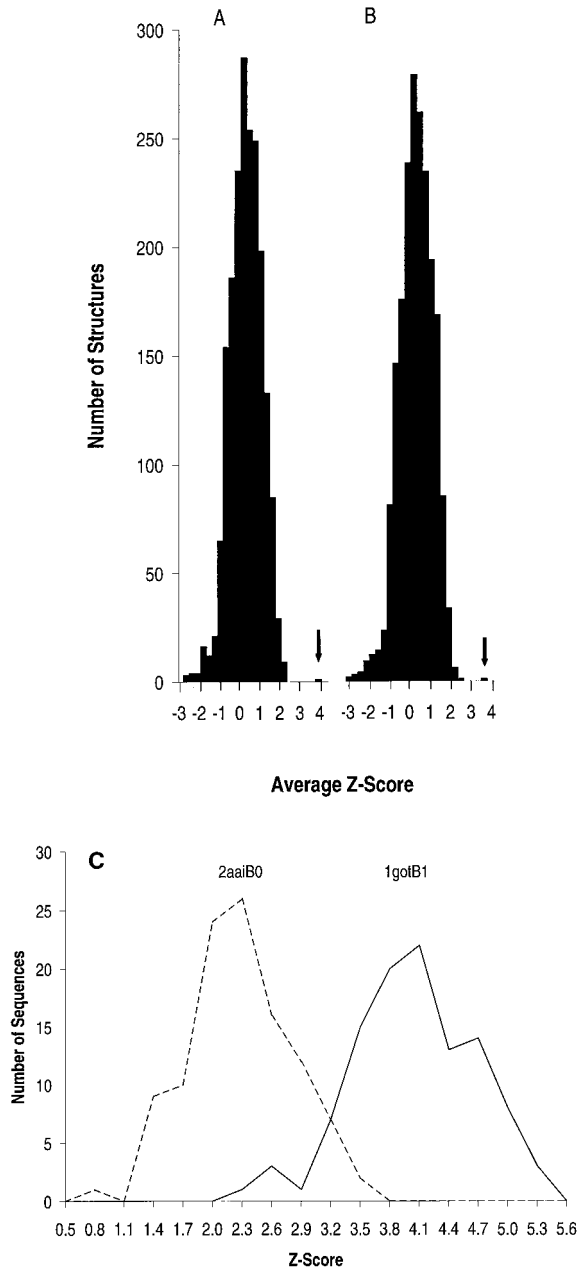

Figure 6. Threading scores for YWTD domains. A and B, Average *Z*-scores for 89 YWTD domains related to LDLR (A) and 18 YWTD domains from *sevenless* and c-*ros* (B). *Z*-scores for each structure were averaged for the 89 or 18 sequences using a program written by Kemin Tan, and plotted with the histogram tool of Microsoft Excel. The arrows mark 1gotB1. C, The distribution of *Z*-scores for all 107 YWTD domain sequences for the second highest hit, 2aaiB0, ricin β-trefoil domain; and the top hit, 1gotB1, G protein β-propeller domain.

6-fold instead of 7-fold pseudosymmetry (see Methods).

The sequence repeats in β-propellers are usually offset relative to the β-sheets. This is possible because the amino acid sequence threads through the propeller in a circular fashion; the N and

Table 4. Normalized average threading scores for the 1gotB1 β-propeller domain

	Z-score[a]	*P*[b]	*E*[c]
89 YWTD[d]	4.52	8.5×10^{-7}	1.6×10^{-3}
18 YWTD[e]	4.20	3.7×10^{-6}	7.2×10^{-3}

[a] The distribution of average *Z*-scores for 1945 structures in the threading library was fit to the normal distribution, and the *Z*-score ((average *Z*-score for 1gotB1 − mean average *Z*-score)/S.D.) was determined.
[b] Probability in a normal distribution of obtaining a *Z*-score as high or higher.
[c] Expectation value for finding a score as high or higher in a database with 1945 structures ($P \times 1945$).
[d] The 89 LDLR-related YWTD domains.
[e] The 18 *sevenless* and c-*ros* YWTD domains.

C-terminal β-strands are adjacent in the structure, and both contribute to the last W (Figure 4). Using a common convention, each sheet of the β-propeller will be called a W, with each leg of the W corresponding to one β-strand (Figure 4A). In the conventional view from the side with the Ws in the propeller upright, the strand 1 to 2 loops and the strand 3 to 4 loops are on the bottom. The strand 2 to 3 loops and the connections between strand 4 of one W and strand 1 of the next are on the top. β-Strand 1 is innermost and runs nearly parallel to the central pseudosymmetry axis; β-strand 4 is outermost. β-Strand 1 in β-propellers has an unusual close-packing mode, in which the side-chains of neighboring β-strands 1 interdigitate; β-strand 1 is more close-packed in β-propellers with six blades than those with seven or eight blades (Murzin, 1992). Residues in β-strand 1 with short side-chains with no more than γ-atoms are preferred, i.e. Gly, Ala, Ser, Cys, Thr, and Val; residues with no more than δ atoms may also be allowed, i.e. Asp, Asn, Ile, and Leu. The only YWTD domain β-strand with appropriately small residues is the fourth β-strand of each YWTD sequence repeat, which contains a GLAVD consensus sequence and therefore can be identified with β-strand 1 of each sheet (Figures 2A, 7A). Thus, β-strand 2 is the first β-strand of each repeat and contains the YWTD motif. The residue in ladder position 2 in β-strand 1 of G protein β is particularly close-packed, and this position can be identified with the G of the GLAVD sequence (Figure 7).

The sequence-structure alignment between YWTD domains and the G protein β-propeller template is surprisingly straightforward. The alignment shown in Figure 7A readily results when (1) the sheets in G protein β are aligned with each other structurally, (2) the predicted sheets in YWTD domains are aligned with one another by sequence as in Figure 2, (3) the overlap between the β-strands in G protein β and predicted β-strands in YWTD domains is maximized, and (4) the number of gaps and insertions is minimized. It is remarkable that insertions or deletions were present on average in only 49% of the 23 loops shared between the template and the three YWTD domain sequences that were modeled, and

Table 5. Threading scores for FN3 repeats

Repeat	Top hit		2nd hit	
	Code	Z-score	Code	Z-score
FN3_4 c-*ros*	1fna00[a]	5.22	1fvcA0[b]	2.37
FN3_4 *sevenless*	3hhrB2[a]	3.68	1fnf02[a]	3.28
FN3_9 c-*ros*	1cfb01[a]	4.51	2tprA3[b]	2.50
FN3_9 *sevenless*	1fnf02[a]	3.66	1gca01[b]	2.57

See footnote[a] to Table 3. FN3 repeat sequences defined as in Figure 5 for chicken, human, rat and mouse c-*ros* and *D. melanogaster* and *D. virilis sevenless* were subjected to THREADER 2.1. The gap penalty was the default of 0.5 for c-*ros* and 0.4 for *sevenless*. The *Z*-scores for combined pairwise and solvation energy were averaged for the four species of c-*ros* and two species of *sevenless*. Structure classifications are from SCOP (Murzin *et al.*, 1995).

[a] FN3 domains are 1fna00, fibronectin FN3 domain 10; 3hhrB2, growth hormone receptor FN3 domain 2; 1cfb01, neuroglian FN3 domain 1; 1fnf02, fibronectin FN3 domain 8.

[b] Non-FN3 domains are 1fvcA0, antibody Fv fragment immunoglobulin V domain; 2tprA3, trypanothione reductase α + β domain; 1gca01, galactose-binding protein α/β domain.

these averaged only 1.6 residues per insertion or deletion. Furthermore, sequence similarity exists between YWTD domains and G protein β that is sufficient to guide alignment. Among β-propellers with known X-ray structures, G protein β and cd1 nitrate reductase appear most closely related in sequence to YWTD domains. The cd1 nitrite reductase β-propeller (Baker *et al.*, 1997) contains eight blades and the sequence begins with β-strand 2, as predicted for YWTD domains. The G protein β-subunit β-propeller contains seven blades and the sequence begins with β-strand 4. The sequence of these structures, the consensus sequence for 21 G protein β-subunits, the consensus sequence for 17 proteins related to cd1 nitrite reductase, and the consensus sequence for 89 YWTD repeats were included in a common alignment and aligned by sequence with PRRP (Gotoh, 1996; unpublished results). Disparities in the length and offset of these sequences are present that are challenging to multiple sequence alignment algorithms. Therefore, it was striking that every β-strand of G protein β was aligned with the correct β-strand of cd nitrite reductase; i.e. each β-strand 1 of one protein was aligned with β-strand 1 of the other, and so on. Furthermore, the sequence-based alignment agreed with a structure-based alignment of each sheet for 21 out of 28 β-strands as determined with the residue in ladder position 3 of Figure 7A. In this same automatic multiple sequence alignment, the alignment between the G protein β-subunit and YWTD repeats agrees with that shown in Figure 7A in 24 of 24 instances for alignment of β-strands 1, 2, 3 and 4 with β-strands 1, 2, 3, and 4, respectively, and in 12 of 24 instances for alignment of the residues in ladder position 3. Thus, similarities in amino acid sequence are present that help confirm (1) the offset between the sequence repeats and the β-sheets, and (2) the specific structure-sequence alignment used for creation of YWTD domain models. When the bacterial sialidase Ws are structurally aligned with those of G protein β and added to the alignment, a LLFTN sequence in W5 aligns with the similar LYWTD consensus sequence of YWTD domains (Figure 7B). Although YWSS and YESS sequences in *Trypanosoma cruzi* neuraminidase have been proposed as YWTD motifs (Pereira *et al.*, 1991), subsequent bacterial neuraminidase structures (Crennell *et al.*, 1993) show that these sequences are not structurally homologous to one another, and are not in β-strand 2 as predicted here for the YWTD motif.

Representative models were made with SEGMOD of LOOK for the YWTD domains of human LDLR, human nidogen (Figure 4B and C), and YWTD domain 7 of chicken LRP1, which has the greatest sequence identity to the YWTD domain consensus sequence (Table 6). The models were evaluated with programs that have been developed to check the quality of X-ray and NMR structures as well as models (Wilson *et al.*, 1998; Sippl, 1993; Vriend, 1990). These methods evaluate structural features that differ from those used in refinement, and thus provide an independent check on quality. Models of yeast Sec-13, a predicted six-bladed β-propeller domain that contains WD40 repeats with low but significant sequence homology to the G-protein β-subunit β-propeller domain (Saxena *et al.*, 1996), served as controls for comparable homology models. As a control for an incorrect model, an ''optimal'' alignment was obtained with THREADER for the 2aai structure (the second highest hit, Table 3), and used to create a model (lrp1-mis). The ldlr, lrp1, and nido models received good scores with all three evaluation methods (Table 6). Prosa II did not ''flunk'' the lrp1-mis misthread model, probably because Prosa II scores with a threading potential that is similar to that used by THREADER to ''optimize'' the alignment with 2aai; however, lrp1, ldlr, and nido received good scores without having alignments optimized in this way. The first generation quality evaluation tool QUACHK of WHATIF and WHATCHECK gave the ldlr, lrp1, nido, and the Sec-13 model made here remarkably good scores that are in the same range as X-ray and NMR structures, and gave the misthread model a markedly lower score. The second generation quality evaluation tool NQACHK flunked the misthread model with a score <-5.0, that indicates it ''is certain to be incorrect;'' moreover, it passed the ldlr, lrp1, and nido models with scores in the -4.00

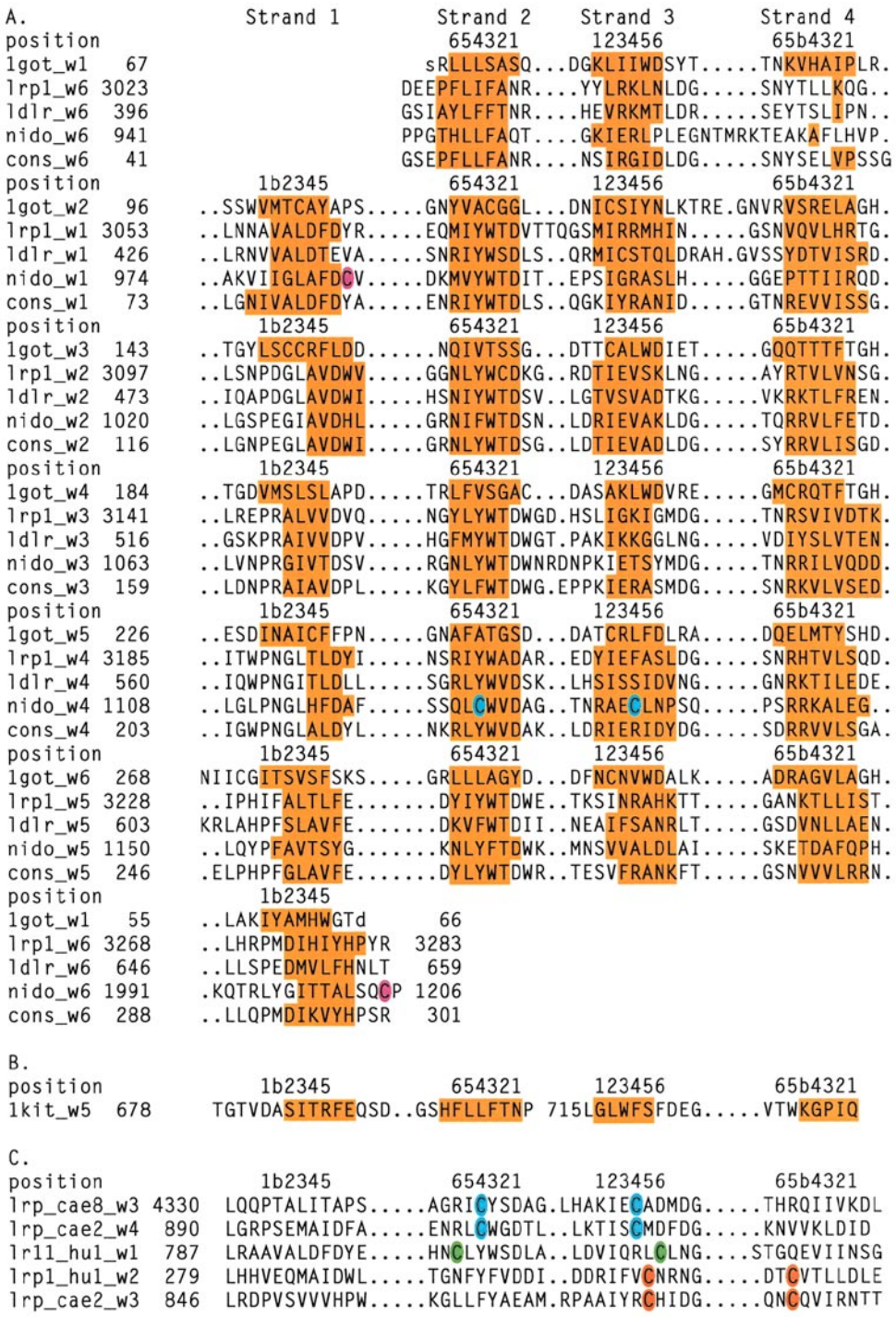

Figure 7. Sequence-structure alignments for models, and position of cysteine residues known or predicted to form disulfide bonds. A, Alignment of the six-bladed β-propeller template made from the transducin β-subunit (1got) as described in Methods with YWTD domain 7 of chicken LRP1 (lrp1), the YWTD domain of the human LDLR (ldlr), the YWTD domain of human nidogen (nido), and the consensus sequence of 89 YWTD domains (cons). The sequence numbers for 1got are for the native protein. Regions of β-strand in 1got as defined by DSSP (Kabsch & Sander, 1983), and in the sequences as predicted by PHD (Rost, 1996) are in gold. β-Sheet ladder positions are marked for Gβ; b is for bulge and varies in position in strand 4 and may differ between Gβ and YWTD domains. Cysteine residues known in nidogen to be disulfide-linked are coded with the same color. B, Structural alignment of W5 of *Vibrio cholerae* neuraminidase, 1kit. β-Strands as defined by DSSP are in gold. The W were cut out of 1kit and 1got and structurally aligned with 3DMALIGN of Modeller. The LLFTN sequence in strand 2 aligns with the LYWTD sequence in YWTD repeats. The long 2-3 loop of 1kit_W5, residues 703 to 714, is deleted. C, YWTD repeats with cysteine residues that are in a position predicted to permit disulfide bond formation between neighboring β-strands. Cysteine residues predicted to be disulfide-bonded are color coded; cysteine residues with equivalent β-sheet ladder positions are given the same color. lr11_hu1_w1 is representative of W1 of LR11_HU1, LR11_CH1, and LR11_RB1. lrp1-hu1_w2 is representative of W2 of LRP1_HU1, LRP1_CH1, LRP1_MO1, YL_DM1, and LRP_CAE7. lrp_cae2_w4 is representative of W4 of LRP_CAE2, LRP1_CE5, NIDO_HR1, NIDO_HU1, NIDO_MO1, and ONID_HU1.

Table 6. Model evaluation

Structure or model	Residues	Prosa II[a] Z-score	QUACHK[b] score	NQACHK[c] Z-score
G protein β-propeller domain[d]	296	−9.67	−0.270	−1.45
Six-bladed β-propeller template[e]	257	−9.74	−0.208	−2.31
ldlr, model of LDLR_HU1[f]	264	−5.34	−0.569	−3.65
lrp1, model of LRPI_CH7[f]	260	−6.70	−0.651	−3.65
nido, model of NIDO_HU1[f]	266	−6.22	−0.549	−3.48
Sec-13, 6-bladed WD40 domain modeled here[g]	282	−8.89	−0.876	−4.07
Sec-13, 6-bladed WD40 domain published model[h]	282	−7.43	−1.360	−5.04
lrp1-mis, control misthreaded model[i]	260	−5.73	−1.793	−5.75

[a] Prosa II combined C^{β} pairwise and surface potential Z-scores relative to the pII3.0.short.ply polyprotein (Sippl, 1993). Lower scores are better. The potentials or pseudo energies used are analogous to those used in threading programs. All models passed the Prosa II check, i.e. ranked 1 relative to all decoys.

[b] Structural average packing environment quality score with the quality check (QUACHK) option of WHATIF. Higher (less negative) values are better. Scores receive the following messages: <−2.7, error, certain to be wrong; −2.7 to −2.0, error, quality is very low; −2.0 to −1.4, warning, quality is a bit low; > −1.4, note, quality is within normal ranges.

[c] New or second generation average structural packing environment Z-score with the NQACHK option of WHATIF. Higher (less negative) values are better. The average Z-score for properly refined X-ray structures is 0.0 ± 1.0. Scores receive the following messages: <−5.0, error, the structure is certain to be incorrect; −5.0 to −4.0, error, abnormal score, quality is very low; −4.0 to −3.0, warning, quality is a bit low, the protein is probably threaded correctly; > −3.0, note, quality is within normal ranges.

[d] Residues 45 to 340 of 1tbgB.

[e] The mo6_s2 model template (Methods).

[f] ldlr, lrp1, and nido models made with the alignment of Figure 7 (Methods).

[g] Made with the mo6_s4 template (Methods) analagously to ldlr, lrp1, and nido.

[h] From Saxena *et al.* (1996), kindly provided by C. Gaitatzes, Boston University, Boston, MA.

[i] Of the three YWTD domains modeled, LRP1_CH7 had the best THREADER Z-score with 2aaiB0 of 2.47. The alignment was optimized by setting the number of alignment paths considered by THREADER to 1000 and used to make a model with SegMod as for other models.

to −3.00 range that are ''a bit low,'' but suggest that ''the protein is probably threaded correctly.'' Notably, the ldlr, lrp1, and nido models scored comparably to or better than two models of yeast Sec-13 that are certain to be correctly threaded based on sequence homology. These results strongly suggest that the six-bladed β-propeller fold is correct, and that the sequence-structure alignment is largely correct.

Agreement with experimental data

Model predictions were tested against biochemical data. Circular dichroism of nidogen, proteolytic fragments of nidogen, and a recombinant fragment of nidogen comprising the YWTD domain and the C-terminal EGF domain shows that they are all-β-structures, i.e. they contain β-sheets and little or no α-helix (Paulsson *et al.*, 1986; Fox *et al.*, 1991). This is completely consistent with the β-propeller domain. Circular dichroism of the intact LDLR shows a ratio of β-sheet to α-helix of 2.2 (Saxena & Shipley, 1997). Using the proportion of β and α structure defined by DSSP (Kabsch & Sander, 1983) in an LDLR class A module (Fass *et al.*, 1997) and in EGF-modules (Downing *et al.*, 1996) to estimate the total in these modules, the total β and α in the ldlr model to estimate the total in the YWTD domain, and assuming a 22 residue transmembrane α-helix, the ratio of β to α is 2.3, in excellent agreement.

Nidogen and its recombinant fragments have been visualized by electron microscopy (Fox *et al.*, 1991). The C-terminal portion of nidogen, corresponding to the YWTD domain is a globule. The YWTD β-propeller domain is also globular (Figure 4B and C).

Studies on proteolytic fragments of nidogen show two disulfide bonds in the C-terminal globule: a long-range disulfide bond connecting C987 to C1205, and a short-range bond connecting C1124 to C1135 (Mann *et al.*, 1988; Fox *et al.*, 1991). These disulfide bonds are confirmed by more recent studies that demonstrate the disulfide pattern in EGF domains (Bork *et al.*, 1996), and by the sequence of osteonidogen, which conserves the four disulfide-linked cysteines in the YWTD domain but lacks the C-terminal EGF module (Figure 1). The structure-sequence alignment places C1124 and C1135 adjacent in ladder position 4 of β-strands 2 and 3 of W4 (Figures 4 and 7A). Disulfide bonds between adjacent β-strands can only form between residues in the same ladder position. C987 in the model is in the 1-2 loop of W1, on the bottom of the β-propeller domain with its side-chain pointing out. C1205 is adjacent in W6, in the segment following the C-terminal β-strand 1 (Figures 4C, 7A). L1202, the last template-defined residue in the model, is 5 Å from C987. There are many possible orientations of the L1202-S1203-Q1204-C1205 loop that permit formation of the C987–C1205 disulfide. The probability of residues a given number of positions apart being close enough to directly form a disulfide bond (C1124–C1135) or close enough (C987 and L1202) to permit a three-residue loop to form the C987–C1205 disulfide bond can readily be determined with the protein database (Table 7). The sequence-structure alignment was independent of the cysteines, and the alignments of the two pairs of cysteines are independent of one another; they are in separate

Table 7. Frequency of sequence-structure distances permissive for disulfide bond formation

Disulfide	Residues	Sequence separation	Structure distance (Å)	Frequency[a]
C987–C1205[b]	C987, L1202	215	$C^{\alpha} < 15$	0.033
C1124–C1135[c]	C1124, C1135	11	$C^{\beta} < 6$	0.017

[a] Distance distributions were calculated by Azat Badretdinov and Andrej Šali, Rockefeller U., with a non-redundant set of all chains in the protein database that are <30% identical in sequence and have ⩾200 residues. These 456 chains are structurally diverse and contain representatives of all structure classes.

[b] Residue i is defined by the template; residue $i + 218$ occurs three residues after the C-terminal template-defined position, $i + 215$. The loop from $i + 215$ to $i + 218$ can span some distance to form the disulfide. Residues i and $i + 215$ must be surface exposed in the absence of the loop, so the loop can form unhindered by the template-defined residues, and must be close enough to allow the loop to form with appropriate stereochemistry. With the nido model as template and using LOOK 3.0 for sequence and alignment editting and Modeller 4.0 for modelling, the sequence was altered around a number of loops varying in distance from C987. A SQCP sequence was added after varying template-defined positions, a chain break was added, loop residues on the other side of the chain-break that might impede disulfide formation were removed, and the SQCP sequence was left undefined by the template. The cysteine of this sequence was patched to C987 and models were built. C^{α} distances from C987 to the residue preceding the SQCP sequence of ⩽13.4 Å permitted disulfide bond formation, but distances of 15.4 and 15.6 Å resulted in major cysteine residue and disulfide bond restraint violations. Therefore, 15.0 Å was taken as the maximum distance for disulfide bond formation. C^{α} atom distance distributions were calculated for all residues that have more than 10% surface exposure, and are 210 to 230 sequence positions apart.

[c] Disulfide-bonded cysteine C^{β} atoms range from 3.0 to 4.6 Å apart. The SegMod modeling program will form disulfide bonds with template C^{β} atoms somewhat farther apart than this. Since C^{β} atoms in equivalent β-sheet ladder positions in the G protein β template were <6 Å distant, 6 Å was taken as the cut off. C^{β} atom distance distributions were calculated for all residues 11 positions apart. Results were similar with a small set of all β structures.

sequence blocks (Figure 7A). Therefore, the null hypothesis of obtaining by chance alone a model that fits the experimental disulfide data can be rejected with $p = 0.033 \times 0.017 = 5.6 \times 10^{-4}$ (Table 7).

Discussion

The evidence that each cluster of six YWTD repeats folds into a six-bladed β-propeller domain, and that a largely correct sequence-structure alignment has been deduced, is summarized in Table 8. The evidence includes (1) wet bench experimental data; (2) deductions based on generalizations from the protein structure and sequence databases on protein fold families, sequence repeats, and exon structures; and (3) results from computational molecular biology. The types of conclusions that can be drawn and their degree of certainty deserve comment. Lines of evidence (1) to (5), taken together in line (6) (Table 8) show that YWTD repeats fold into a compact globular domain with six structurally similar units. It is estimated that most but not all protein folds are currently known (Murzin, 1996). Only one out of the 393 currently known protein folds arranges a single polypeptide into a single domain with six structurally similar units, the six-bladed β-propeller (Murzin *et al.*, 1995). Assuming that a similar proportion of known, and yet to be discovered, folds will have six structurally similar units, lines of evidence (1) to (5) assign the YWTD domain to a fraction of approximately 0.003 of all folds.

The normalized averaged THREADER *Z*-scores show a structural relationship of the 89 YWTD domains to the G protein β-propeller domain, with an expectation that this could occur by chance alone of 1.6×10^{-3} (line (11), Table 8). It is difficult to detect a sequence relationship of the YWTD domains related to the LDLR with the YWTD domains of *sevenless* and c-*ros*; it should be noted that a relationship has been reported with only two isolated YWTD repeats in *sevenless* (Norton *et al.*, 1990), whereas three groups of six repeats each are identified here. Therefore, lines of evidence (11) and (12) are at least partially independent.

Lines of evidence (13) and (14) in Table 8 demonstrate not only identification of a six-bladed β-propeller fold, but also a largely correct structure-sequence alignment. The most dramatic confirmation of the fold of YWTD domains is the C987 to C1205 disulfide bond that links the first and the last β-sheets of the β-propeller. The C1124–C1135 disulfide bridge between β-strands 2 and 3 of W5 of nidogen cannot form if the relative alignment of these strands is shifted by only one residue. The structure alignment with other sequences is consistent with formation of further disulfide bonds between β-strands 2 and 3, and between β-strands 3 and 4, that are predicted by pairs of conserved cysteine residues (Figure 7C). The probability that the two nidogen disulfide bonds would form by chance alone in the case of assignment of an incorrect fold to the YWTD domain is 5.6×10^{-4}.

The structure quality evaluation programs used here readily ''flunk'' deposited X-ray structures that have one residue structure to sequence frameshifts, or switches in loops. These programs also readily detect incorrect structure-sequence alignments and incorrect fold assignments in models, at least when the models are constructed independently of the criteria used for evaluation. It is particularly impressive that QUACHK places the quality of YWTD domain models within the normal range for X-ray and NMR structures, and that they score as well or better than comparable

Table 8. Lines of evidence for the YWTD β-propeller domain

Evidence	Conclusion	Probability, expectation, frequency, or strength of conclusion
(1) Cysteine-to-size rule	YWTD repeats are not separate domains	High
(2) Always six YWTD repeats	Six YWTD repeats form a single domain	Moderate
(3) Exons of same phase at domain boundaries, not at repeat boundaries	Modular unit is group of six repeats, not individual repeats	High
(4) Nidogen electron microscopy	YWTD repeats form a single globular domain	Very high
(5) YWTD repeats have sequence homology	Repeats have similar structure	Very high
(6) (1, 2, 3, 4) + 5: Single, globular domain with 6 structurally similar units	Consistent with only one of 393 known folds: 6-bladed β-propeller domain	$F = 0.003$
(7) Prediction of groups of six YWTD repeats in *sevenless* /c-*ros*	Hypothesis has heuristic value	High
(8) YWTD domains have 24 predicted β-strands, with 4/repeat	Agrees with 6-bladed β-propeller domain	High
(9) Fold prediction with TOPITS: 7 of 8 top hits are β-propeller	β-Propeller domain	High
(10) Circular dichroism of nidogen shows all-β-structure	Agrees with β-propeller domain	Moderate
(11) THREADER normalized *Z*-score for LDLR-related YWTD domains of 4.52	β-Propeller domain	$E = 1.6 \times 10^{-3}$
(12) THREADER normalized *Z*-score for *sevenless* /c-*ros* YWTD domains of 4.20	β-Propeller domain	$E = 7.2 \times 10^{-3}$
(13) Model of nidogen agrees with experimentally determined disulfides	Six-bladed β-propeller domain with largely correct sequence alignment	$P = 5.6 \times 10^{-4}$
(14) Representative models have good structural quality	Six-bladed β-propeller domain with largely correct sequence alignment	Extremely high

homology models. In the author's experience, this can only happen with a correct fold assignment and a largely correct sequence alignment; even small differences between models and the correct fold, such as a shift in an edge β-strand from one sheet to another in Ig domains, are readily detected by poor scores. Shifting individual or several β-strands in the YWTD domain sequence relative to the alignment with Gβ shown in Figure 7 gave markedly poorer scores. Use of quality control programs for statistical evaluation of models is under development (R. Sánchez & A. Šali, personal communication), and suggests that the probability that at least 30% of the residues in the three YWTD models will be within 3.5 Å of the actual structures is 92 to 98%.

Lines of evidence (11) and (13) in Table 8 are independent of one another and are associated with estimates of their statistical reliability. Line of evidence (6) is independent of (11) and (13) and narrows the choice of folds to a fraction of about 0.003 of all folds; the probability that these three lines of evidence would all yield the same fold by chance alone is 3×10^{-9}. Line of evidence (12) is at least partially independent of line of evidence (11). Therefore, the chance of the β-propeller fold being incorrect is between 3×10^{-9} and 2×10^{-11}. Furthermore, this does not take into account lines of evidence (7) to (10) and (14), which are more difficult to convert to a probability, are independent of each other, (6) and (11) to (13), and provide support that ranges from moderate to extremely high. Therefore, it can be firmly concluded that YWTD domains have a six-bladed β-propeller fold.

A number of recent fold predictions have been proven correct by subsequent structure determinations (Russell & Sternberg, 1995; Edwards & Perkins, 1996; Brissett & Perkins, 1996). Here, I take the prediction process further than in the past, by quantifying the amount of uncertainty in the prediction and showing that it is below the level required for acceptance of a finding as a fact. This may not be possible, in general, for fold predictions. The information content available for the current prediction was extraordinarily rich. It included clearly defined domain boundaries, 107 sequences with an almost ideal amount of diversity, experimental data including circular dichroism, electron microscopy, and disulfide bonds, and the presence of structurally similar units within the fold. It should be cautioned that although prediction algorithms can, at least in some cases, be extraordinarily powerful, expertise is required for their optimal use and interpretation, just as for wet bench experimental techniques.

Any description of the atomic features of the YWTD domain models is speculative; however, some discussion appears justified by the presence of interesting contrasts with the WD40 repeat. Although the Trp residue of the YWTD motif is in β-strand 2, whereas that of the WD40 motif is in β-strand 3, the YWTD Trp occupies a similar position in the models, parallel to and in between adjacent β-sheets, and pointing out toward the perimeter of the β-propeller. In G protein β-subunit WD40 repeats, four highly conserved side-chains form hydrogen bonds that help knit together each sheet and orient adjacent sheets (Sondek *et al.*, 1996). All four residues in the YWTD motif have side-chains that are capable of forming hydrogen bonds, and are likely to serve a similar function. The YWTD motif appears in the upper portion of β-strand 2. The Thr and Asp side-chains are particularly well situated to form hydrogen bonds to residues in strand 1 and the loop preceding strand 1, in the same and the following β-sheet. The much poorer conservation of the YWTD motif in *sevenless* and c-*ros* is consistent with a less uniform orientation between neighboring β-sheets.

The YWTD domain meets all the criteria for protein modules (Bork *et al.*, 1996). It can neighbor different types of modules, including EGF, FN3, thyroglobulin-like, and Vps10p modules. It appears in proteins with widely varying functions. The phase 1 introns that flank YWTD modules make them compatible with the most widely dispersed extracellular modules.

The architecture of the YWTD domain may have interesting consequences for the mosaic proteins in which it is present. The N and C termini of the YWTD domain are only 5 Å apart on the broad, 40 Å wide lower surface of the β-propeller. β-Propeller enzymes, the G protein β-propeller domain, and the predicted integrin β-propeller domain have their active or ligand binding sites on the upper surface, which is richly endowed with neighboring loops that run in opposite directions. In the YWTD domain this face is unhindered by the connections to neighboring domains on the lower surface. EGF and FN3 domains have their C and N termini at opposite ends of the module. Tandem arrays of these domains generally have a linear, extended architecture. Rather than acting as spacers as previously thought, the YWTD domains bring their N-terminal and C-terminal module neighbors into close proximity. It is interesting that two YWTD domains are never adjacent; EGF or FN3 modules always intervene. It remains to be determined whether there is a preferred orientation between YWTD domains and their module neighbors; however, it is not unlikely that a change or even reversal in direction of tandem arrays of neighboring EGF or FN3 domains would result. By contrast to the impression given in Figure 1 which is a one-dimensional schematic, YWTD domains are likely to alter the linear, extended architecture generally seen for EGF and FN3 domains, and result in a much more compact, three-dimensionally elaborate structure. It is intriguing that in LRP1, LRP2, and *yolkless*, there are groups of two or four YWTD domains in which only a single EGF domain intervenes between each YWTD domain (Figure 1). These YWTD domains must be quite close to one another; and might act together, for example in binding a single, large ligand.

Studies on nidogen are consistent with the connections to the neighboring domains and the binding site for laminin being on opposite faces of the YWTD domain. A fragment of nidogen containing the YWTD domain, the C-terminal neighboring EGF domain, but lacking the N-terminal neighboring thyroglobulin-like domain, binds laminin with the same nM K_d value as intact nidogen (Fox *et al.*, 1991). The lack of effect of removal of the thyroglobulin-like domain suggests that the binding site is not adjacent to its connection on the lower face of the β-propeller domain. Electron micrographs of a laminin fragment bound to nidogen (Fox *et al.*, 1991) are consistent with laminin binding to the top of the β-propeller domain, since binding appears to be on the side of globule 3 opposite to the rod that connects to globule 2. Interestingly, the integrin α3β1 appears to bind laminin through loops on the top of its predicted β-propeller domain (Zhang *et al.*, 1998; Springer, 1997), although the binding site in laminin may differ.

A homologue of LRP1 in *C. elegans* is reported here (Figure 1 and Table 1; see Methods). Thus, the pair of large, similar LRP1 and LRP2 genes have been conserved, since the divergence of nematoda and chordata. Their importance is further underscored by the lethality of deletion of either gene in mice (Herz *et al.*, 1992; Willnow *et al.*, 1996). Certain of the YWTD modules are unusually well conserved in the LRPs particularly for extracellular proteins. In YWTD domain 3 of LRP1, the human sequence is 95 and 46% identical to chicken and *C. elegans*, respectively. Because of this conservation and the ligand-binding function of the YWTD domain of nidogen, it is important to keep open the hypothesis that YWTD domains may function in ligand binding in endocytic receptors, especially in the LRP proteins which specifically bind such a wide range of ligands (Table 1).

Homologues for vertebrate hormones and cytokines are often absent from lower organisms; no EGF precursor has yet been identified in invertebrates. Given the wide radiation of endocytic receptors bearing YWTD and EGF modules both in *C. elegans* and vertebrates, it is tempting to speculate that the EGF precursor evolved from an endocytic receptor that lost its LDLR class A repeats.

The importance of the YWTD domain in the LDLR is emphasized in familial hypercholesterolaemia by the large proportion of missense mutations that map to it, particularly to YWTD repeats 1 and 2 (Hobbs *et al.*, 1992; Krawczak & Cooper, 1997; Figure 8A). Deletion experiments suggest that the YWTD and EGF domains are important for proper exposure of LDL but not β-VLDL binding sites in the class A repeats, for resistance of the receptor to degradation under basal conditions or during endocytosis of β-VLDL, and for acid-induced dissociation of the receptor-β-VLDL complex (Davis *et al.*, 1987; Van der Westhuyzen *et al.*, 1991). However, many of the same results are obtained by deletion of EGF domain 1, EGF domains 1 and 2, or mutation of the Ca^{2+} binding site between class A module 7 and EGF domain 1 (Esser *et al.*, 1988; Van der Westhuyzen *et al.*, 1991; Figure 1). There is little effect of deletion of EGF domain 2 or mutation of the Ca^{2+} binding site between EGF domains 1 and 2 (Esser *et al.*, 1988); the effect of deleting the YWTD domain alone is unknown.

The positions of 39 independent missense mutations were mapped within the model of the LDLR β-propeller domain (Figures 2A, 8B and C). Mutations are concentrated in W1 and W6 (49% of all mutations). In all W, mutations in the inner β-strands 1 and 2 are far more common (51%) than in strands 3 and 4 (13%). Interestingly, mutations also cluster on the bottom of the β-propeller domain, the side that is connected to the EGF domains. It is attractive to think that the focus of mutations in W1 and W6 and near the bottom of the YWTD domain may reflect the functional importance of connections to the EGF domains and possible conformational changes accompanying acid-induced ligand dissociation (Davis *et al.*, 1987); however, these regions may also be critical structurally. The deleterious effect of conservative Asp445 → Glu and Asp579 → Asn mutations in the YWTD motif in repeats 2 and 5 (Figure 2) supports the idea that the Asp side-chain of the YWTD accepts important hydrogen bonds.

sevenless is one of the most fascinating receptor tyrosine kinases known (Hafen *et al.*, 1987; Krämer *et al.*, 1991; Cagan *et al.*, 1992). The location within its large extracellular domain of the ligand recognition site for *bride of sevenless*, or *boss*, is unknown. This study has advanced knowledge on the modular organization of this extracellular segment, by identifying three YWTD β-propeller domains and two further FN3 domains, and defining the organization of a contiguous segment of 1800 residues. *boss* is a seven transmembrane receptor with a long N-terminal extracellular segment, and *sevenless* acts as an endocytic receptor in internalization of *boss* in a process that appears important for signaling development into R7 photoreceptor cells. Possible functions for YWTD domains in ligand recognition or endocytosis by *sevenless* can now be investigated.

The six-bladed β-propeller fold for YWTD domains defines an interesting new class of extracellular module, and has substantial implications for a diverse and biologically important group of molecules. Although lacking in atomic precision, the models described here provide important structural information on the threading, approximate location, and relative proximities of amino acid residues within each YWTD domain. Moreover, it is important for future structure-function and atomic resolution studies that domain boundaries are now identified, and that the YWTD repeats previously envisioned as five separate modules can now be seen to be six structurally interdependent units that fold in a specific manner into a single domain.

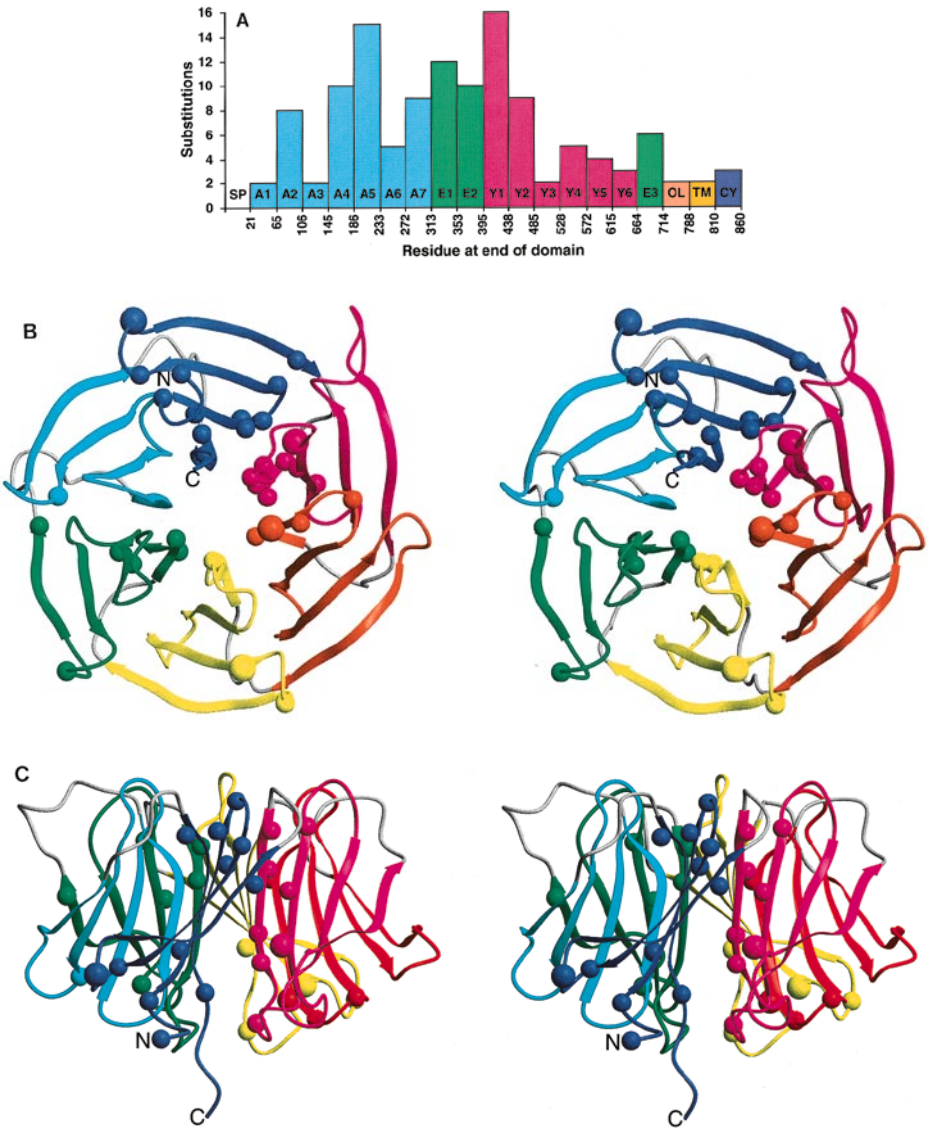

Figure 8. Mutations in the LDLR in familial hypercholesterolaemia. A, Point mutations by domain. All missense mutations, i.e. point mutations that caused a change in amino acid sequence, except in the iniation codon, were obtained by querying the Human Gene Mutation Database, Cardiff UK (http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html?) (Krawczak & Cooper, 1997). Positions with mutations to two different residues were counted twice. B, Ribbon stereo view of the LDLR YWTD domain model. Each W is a different color, with 4-1 loops in grey. Spheres mark the C^{α} position of each residue with a missense mutation in familial hypercholesterolaemia; the spheres are larger for residues with two different substitutions. The view is approximately parallel to the 6-fold pseudosymmetry axis, with the ''bottom'' of the propeller containing the N and C termini, 1-2, and 3-4 loops in the foreground. Figure prepared with MOLMOL (Koradi *et al.*, 1996). C, Same as B, with a view of the LDLR β-propeller domain from the side. W6 and W1 are in the foreground; the ''top'' of the β-propeller with the 4-1 and 2-3 loops uppermost.

Methods

Initial searches

Initially, the segment of LDLR_HU containing the YWTD repeats was used to search SWISS-PROT with BLAST. A consensus sequence containing 50 homologous sequences was prepared with PairWise (Birney *et al.*, 1996) and used for BLAST2 searches (Altschul *et al.*, 1997) of non-redundant protein databases using the http://www.bork.embl-heidelberg.de:8080/BLAST2/server and the http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html server with WU-BLASTP + BEAUTY.

Proteins

The abbreviations for the protein sequences obtained directly from databases, with synonyms and SWISS-PROT/NCBI, and/or EMBL accessions are EGF_HU, EGF_HUMAN; EGF_MO, EGF_MOUSE, EGF_RA, EGF_RAT; LDL1_XE, LDL1_XENLA; LDL2_XE, LDL2_XENLA; LDLR_CP, L36118, CPLDLREC_1 *Chiloscyllium plagiosum* LDLR; LDLR_CR, LDLR_CRIGR; LDLR_HU, LDLR_HUMAN; LDLR_MO, LDLR_MOUSE; LDLR_RB, LDLR_RABIT; LDVR_CH, LDVR_CHICK (vitellogenin receptor); LDVR_HU, LDVR_HUMAN; LDVR_MO, LDVR_MOUSE; LDVR_RA, LDVR_RAT; LDVR_RB, LDVR_RABIT; LDVR_XE, 2118675,

vitellogenin receptor of *X. laevis*; LR11_CH, 1552268, chick LR11; LR11_HU, 1552324, human LR11; LR11_RB, 1665753, rabbit LR11; LR8_CH, X97001 chick LR8B; LR8_HU, 1483143 human LR8, apolipoprotein E receptor 2; LRP1_CH, LRP1_CHICK; LRP1_HU, LRP1_HUMAN; LRP1_MO, 49942; LRP2_RA, LRP2_RAT; LRP_CAE, LRP_CAEEL; NIDO_HR, HRENTI_1, *Ascidian Halocynthia roretzi* nidogen; NIDO_HU, NIDO_HUMAN; NIDO_MO, NIDO_MOUSE; ONID_HU, 1449167 human osteonidogen; YL_DM, YL_DROME, *yolkless* gene product of *Drosophila melanogaster*, vitellogenin receptor; SEV_DM, 7LES_DROME, *D. melanogaster sevenless*; SEV_DV, 7LES_DROVI, *Drosophila virilis sevenless*; SEV_CH, TVCHSR, chicken c-*ros*; SEV_HV, TVHURS, human c-*ros*; SEV_MO, 2117869, mouse c-*ros*; SEV_RA, 2117871, rat c-*ros*.

Caenorhabditis elegans genes

Searches yielded several genes from the *C. elegans* genome sequencing project, the splicing of which has been predicted with the program Gene Finder and by correspondence to cDNA expressed sequence tags (Wilson *et al.*, 1994). The predicted genes were submitted to BLAST2 to identify the closest homologues in other species. With PairWise (Birney *et al.*, 1996) a ''negative sequence profile'' was constructed from the sequences of 79 YWTD domains together with their flanking EGF domains, weighting sequences by PHYLIP tree branch length and using the BLOSUM62 amino acid substitution matrix, and used to search for multiple YWTD domains within a single protein. Once these were identified, PairWise searches were done with the corresponding segments of the spliced genes and the unspliced cosmid clones; the automatic scaling function was used to set frame shift and extension penalties for cDNA (spliced) and genomic sequences, respectively.

A homologue of LRP1, designated here LRP1_CE, is predicted to be encoded by the cosmids F47B3 (U97017) and T21E3 (AF003133), which are adjacent on chromosome I. Although separate genes were predicted in each cosmid, it is clear that a single gene spans them. The F47B3.8 gene contains the signal sequence and YWTD domains 1 to 6; the T21E3.3 gene contains YWTD domains 7 and 8, a transmembrane domain, and a cytoplasmic domain containing NPXY-like endocytosis motifs. Splicing from F47B3 to T21E3 is predicted by PairWise at a position upstream from and in frame with the previously predicted initiation codon of T21E3.3. There may be an error in the predicted splicing of T21E3.3 in the third of 11 EGF-like domains that follow the last YWTD domain and precede the transmembrane segment; this third ''domain'' appears EGF-like only in its C-terminal half. BLASTP2 searches showed the highest homology to mammalian LRP1, and less to mammalian LRP2 and LRP_CAEEL, the *C. elegans* LRP2-like protein. Phylogenetic trees of 89 different YWTD domains with their two flanking EGF domains confirmed this; YWTD domain 3 of LRP2_HU, LRP_RA, and LRP_CAE formed a single cluster, while domain 3 of LRP1_CE, LRP1_CH, LRP1_HU, and LRP1_MO formed a single, separate cluster.

A homologue of LDVR, designated LDVR_CE here, is encoded by cosmid T13C2 (U40030). The four top BLAST2P hits were LDVR_MOUSE, LDVR_RAT, LDVR_RABIT, and LDVR_HUMAN. The predicted T13C2.4 gene appears to contain an unrelated gene fused by the splicing prediction N-terminal to the LDVR homology region, which begins with the exon starting at nucleotide 16,738. This region contains the same number of all types of repeats as LDVR, and like the major LDVR_CH transcript (Bujo *et al.*, 1995) lacks an *O*-glycosylated region. It contains a transmembrane domain, and a cytoplasmic domain with an NPIY endocytosis motif. PairWise analysis with the unspliced gene revealed a frameshift within the threonine codon in one of the YWTD motifs. This was repaired by insertion of an A after nucleotide 19,760 of cosmid T13C2, and the exon was extended by 138 nucleotides to position 19,927. The revised splice donor sequence, CAAGgtatgt, matches the consensus better. The predicted amino acid sequence contains ten revised residues and 56 additional residues; these 56 residues are 46% identical to the corresponding region from the YWTD domain consensus.

The cosmid F14B4 (Z75535) contains a predicted gene, F14B4.1, with similar homology to widely different proteins containing YWTD domains, and therefore may currently lack a vertebrate orthologue. It contains an N-terminal signal sequence, a single YWTD domain with flanking EGF domains, and a basic C-terminal region but apparently no transmembrane domain. It was designated LDR_CE here.

Sequence alignments

Sequences are designated according to the protein and the position of the YWTD β-propeller domain (defined as in the key to Figure 1); i.e. LRP1_HU3 for YWTD domain 3 of LRP1_HU. YWTD sequences flanked by EGF or thyroglobulin repeats, i.e. excluding *sevenless* and c-*ros*, were placed in one alignment that contained 89 different YWTD domains. The sequences included one EGF domain preceding and following each YWTD domain, except for LR11 or nidogen which have another type of N-terminal domain, and osteonidogen, which lacks a C-terminal domain. EGF domain boundaries were identified from SWISS-PROT annotations, and by alignment followed by inspection of sequences in the following groups: nidogen and osteonidogen; LDLR and LDVR; LRP1; LRP2; LR8; LR11; and EGF precursor. In LRP1, LRP2, and *yolkless*, certain EGF domains are C-terminal to one YWTD domain and N-terminal to the next. Thus, in certain cases the identical EGF sequence was present on opposite ends of the alignment, presenting a challenge to alignment programs. Alignment was with PRRP (Gotoh, 1996), using Gonnet amino acid substitution matrices. Initially, the prealignment before the iterative refinement was by the rough length method and the penalty for terminal gaps was set identically to that for internal gaps (=1). The resulting alignment was then used as the input for a further iterative refinement. The conversational mode was used to set terminal gap penalties = 0.001, and five series were done. Gap extension and opening penalties were tested; those of three and eight, respectively, gave sequence blocks with the least number of gaps in positions of preliminarily predicted β-strands.

Secondary structure predictions

A phylogenetic tree of the 89 sequences with YWTD domains and flanking EGF domains was constructed with the PHYLP program in the PRRP package (Gotoh, 1996). The eight phylogenetically most distant sequences had <21% identity to the three target sequences for structure prediction; these sequences, LRP1_CE1, LRP1_CE4,

LRP1_CE5, LRP1_CE8, YL_DM2, YL_DM3, LRP_CAE2, and LRP_CAE4 were discarded from the multiple alignment of 89 sequences. The consensus of all 89 sequences, from the .prf profile file made by PairWise using branch length weights, was added to the alignment. The resulting 82 sequences were realigned with the same PRRP settings as above. The alignment contained 575 positions, and 39(±1)% of the positions were gaps for the target sequences. The gaps, many resulting from just one or two sequences, most often the LRPs, greatly degraded secondary structure prediction because PHD does not remove gaps present in the target sequence from submitted multiple sequence alignments. Therefore, all blocks of sequence where no residue was present in the target sequence were removed by editing. This resulted in 307 to 359 alignment positions for the different targets, and sequences with <23% identity to the targets were omitted from individual alignments, as was the consensus sequence. The number of sequences aligned to each target, and their mean and S.D. sequence identity were EGF_HU2, $n = 79$, $\bar{x} = 31(\pm 9)\%$; LDLR_HU1, $n = 79$, $\bar{x} = 36(\pm 14)\%$; NIDO_HU1, $n = 73$, $\bar{x} = 29(\pm 8)\%$; consensus, $n = 81$, $\bar{x} = 44(\pm 6)\%$. These alignments were submitted to PHD (Rost, 1996) (http://www.embl-heidelberg.de/predictprotein/predictprotein.html).

For secondary structure prediction of c-*ros*, the entire extracellular and transmembrane domains of SEV_CH, SEV_HU, SEV_RA and SEV_MO were aligned with PRRP. Gaps in the SEV_CH target sequence were removed, and MSF alignments were submitted to PHD. Thus, secondary structures of individual YWTD domains were predicted independently of one another; sequence identity was too low to yield an accurate multiple alignment that included YWTD domains 1 to 3.

Motif searches with *sevenless*

An alignment of the previously identified (Norton *et al.*, 1990) FN3 repeats 1 to 3 and 5 to 8 from SEV_CH, SEV_HU, SEV_RA, SEV_MO, SEV_DM, and SEV_DV was divided into four sequence blocks. Motifs were identified in each block with MEME (Bailey, 1995) (http://www.sdsc.edu/MEME). Motif profiles were submitted to MAST (Bailey, 1995) (http://www.sdsc.edu/MEME/meme/website/mast.html) and run against the non-redundant database to determine the significance of motif matches to FN3 repeats 4 and 9 in *sevenless* proteins.

Threading predictions

Threading was with THREADER v2.1a (Jones *et al.*, 1995) (http://globin/bio.warwick.ac.uk/ ~ jones/threader.html) and used default options unless specified otherwise. The database of 1908 representative chains and domains was updated by including chains and domains from pdb structures 1hex, 1tad, 1tbg, 2bat, 2sil, 4aah, 1eur, 1got, 1ido, 1jlm, 1kit, 1nir, 1nnc, 1zxq, 1tit, 1wit, a 1.28 Å structure of cytochrome *cd*1 from *Thiosphaera pantotropha* (Baker *et al.*, 1997) and structures of ICAM-1 (Casasnovas *et al.*, 1998) and MAdCAM-1 (Tan *et al.*, 1998). Files were converted to THREADER database format using STRSUM.

Templates and models

The seven-bladed G protein β-propeller domain (chain B of 1tbg.pdb) was transformed with the pseudosymmetry axis on the z axis, and permuted to place residues 45 to 54 after residues 55 to 340, so that all of sheet 7 was together at the end of the pdb file (1tbg_per_sym.pdb). A model seven-bladed propeller was made with SegMod (Look 3.0; Molecular Applications Group, Palo Alto, CA; Levitt, 1992) in which the sequence of sheet 6 was incorporated into sheet 7. The sequence from I270 to H311 was modeled on equivalent residues in sheet 7 (Springer, 1997), except that for residues D298 to D303, which took the place of a loop from the C terminus to the N terminus of the β-propeller domain, the template was the corresponding residues from sheet 6 superimposed on sheet 7 using framework residues (Figure 2 of Springer, 1997). The new sheet 7 with sheet 6 sequence was then pasted into 1tbg_per_sym in place of sheet 7, to make file 1tbg_per_sym_166. Since the contacts between neighboring sheets are the key determinants of β-propeller domain structure (Murzin, 1992), distance restraints between each pair of sheets were used to prepare a six-bladed propeller with Modeller4 (Šali & Blundell, 1993) (http://guitar.rockefeller.edu/modeller/modeller.html). Two alignments were used to generate restraint files:

```
A B - - - - -        - B C - - - -
- - C D - - -        - - - D E - -
- - - - E F -        A - - - - - f
1 2 3 4 5 6 -  and   1 2 3 4 5 - 6,
```

in which A, B, C, D, E, and F symbolize templates for sheets 1 to 6, f is the model sheet with sheet 6 sequence in place of sheet 7, and 1 to 6 symbolize the aligned sequence of the sheets for the six-bladed model. Each alignment instructed Modeller to make restraint (.rsr) files that contained distance restraints only within individual sheets, and between each pair of neighboring sheets. Sheet 6 was affected both by the restraints between sheets 5 and 6 (E and F) and sheets 7 and 1 (f and A). The two restraint files were combined, and the starting position for the model (.ini file) was taken from that of sheets 1 to 6 in a seven-bladed propeller, yielding 1tbg_per_sym_166_mo6.

To ensure a native-like sheet structure, the above six-bladed models were used as templates to prepare the final six-bladed templates. Sheets 1 to 6 were individually cut from the original 1tbg.pdb file, and superimposed individually on the corresponding sheets of the Modeller (mo6) model, using framework residues that align with the KIYAMHW, LLLS, KLIIWD, and KVHAI sequences in sheet 1. The coordinates for the superimposed, native sheets were written to a pdb file. The files were permuted so that the N terminus was placed at the beginning of strand 2 (_s2), or strand 4 (_s4) in sheet 1 to create e.g. file w1-6_on_1tbg_per_sym_166_mo6_s2. These files were then used as templates with SegMod in Look 3.0 to model the same sequence, except that H62 and R150 were mutated to alanine. Both residues protrude into the central cavity of G beta, and the central cavity will be smaller in a six-bladed propeller. The final template files were mo6_s2 and mo6_s4.

Final models were made using the sequence-structure alignment shown in Figure 7, with either w1-6_on_1tbg_per_sym_166_mo6_s2 or mo6_s2 as templates. A model of yeast Sec-13 used mo6_s4 as template, except that the template pdb file was permuted to begin with residue E130 to obtain greater sequence homology. The alignment was similar to a previous Sec-13 model (Saxena *et al.*, 1996) except that some insertions were moved out of β-strands into loops.

Data deposition

The coordinates for the theoretical models have been deposited with the Protein Data Bank: LDLR_HU, 1brx; LRP1_CH7, 1lpx; NIDO_HU1, 1ndx.

Acknowledgments

I am deeply grateful to Azat Badretdinov and Andrej Šali for calculation of distance distributions and support of Modeller, to Michael Levitt for construction of an alternative six-bladed β-propeller template, to Burkhard Rost for support of PHD and TOPITS, and to Osamu Gotoh for support of PRRP. I appreciatively thank Kemin Tan for writing a program, Roberto Sánchez and Andrej Šali for sharing an unpublished paper, and T. Huynh for installing and compiling programs. Supported by NIH grant HL-48675.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, **(17)**, 3389–3402.

Bailey, T. L. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning,* **21**, 51–80.

Baker, S. C., Saunders, N. F. W., Willis, A. C., Ferguson, S. J., Hajdu, J. & Fulop, V. (1997). Cytochrome cd_1 structure: unusual haem environments in a nitrite reductase and analysis of factors contributing to β-propeller folds. *J. Mol. Biol.* **269**, 440–455.

Barclay, A. N., Birkeland, M. L., Brown, M. H., Beyers, A. D., Davis, S. J., Somoza, C. & Williams, A. F. (1993). *The Leucocyte Antigen Facts Book,* Academic Press, London.

Bell, G. I., Fong, N. M., Stempien, M. M., Wormsted, M. A., Caput, D., Ku, L. L., Urdea, M. S., Rall, L. B. & Sanchez-Pescador, R. (1986). Human epidermal growth factor precursor: cDNA sequence, expression *in vitro* and gene organization. *Nucl. Acids Res.* **14(21)**, 8427–8446.

Birchmeier, C., O'Neill, K., Riggs, M. & Wigler, M. (1990). Characterization of ROS1 cDNA from a human glioblastoma cell line. *Proc. Natl Acad. Sci. USA,* **87**, 4799–4803.

Birney, E., Thompson, J. D. & Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl. Acids Res.* **24**, 2730–2739.

Bork, P. (1994). *Drosphila kelch* motif is derived from a common enzyme fold. *J. Mol. Biol.* **236**, 1277–1282.

Bork, P. & Bairoch, A. (1995). Extracellular protein modules: a proposed nomenclature. *Trends Biochem. Sci.* **2**, (Suppl.).

Bork, P., Downing, A. K., Kieffer, B. & Campbell, I. D. (1996). Structure and distribution of modules in extracellular proteins. *Quart. Rev. Biophys.* **29**, 119–167.

Brandstetter, H., Bauer, M., Huber, R., Lollar, P. & Bode, W. (1995). X-ray structure of clotting factor IXa: active site and module structure related to Xase activity and hemophilia B. *Proc. Natl Acad. Sci. USA,* **92**, 9796–9800.

Brissett, N. C. & Perkins, S. J. (1996). The protein fold of the hyaluronate-binding proteoglycan tandem repeat domain of link protein, aggrecan and CD44 is similar to that of the C-type lectin superfamily. *FEBS Letters,* **388**, 211–216.

Brown, M. S. & Goldstein, J. L. (1986). A receptor-mediated pathway for cholesterol homeostasis. *Science,* **232**, 34–47.

Brown, M. S., Herz, J. & Goldstein, J. L. (1997). Calcium cages, acid baths and recycling receptors. *Nature,* **388**, 629.

Bujo, H., Yamamoto, T., Hayashi, K., Hermann, M., Nimpf, J. & Schneider, W. J. (1995). Mutant oocytic low density lipoprotein receptor gene family member causes atherosclerosis and female sterility. *Proc. Natl Acad. Sci. USA,* **92**, 9905–9909.

Cagan, R. L., Krämer, H., Hart, A. C. & Zipursky, S. L. (1992). The bride of sevenless and sevenless interaction: internalization of a transmembrane ligand. *Cell,* **69**, 393–399.

Casasnovas, J. M., Stehle, T., Liu, J.-h. , Wang, J.-h. & Springer, T. A. (1998). A dimeric crystal structure for the N-terminal two domains of ICAM-1. *Proc. Natl Acad. Sci. USA,* **95**, 4134–4139.

Corbi, A. L., Garcia-Aguilar, J. & Springer, T. A. (1990). Genomic structure of an integrin alpha subunit, the leukocyte p150,95 molecule. *J. Biol. Chem.* **265**, 2782–2788.

Crennell, S. J., Garman, E. F., Laver, W. G., Vimr, E. R. & Taylor, G. L. (1993). Crystal structure of a bacterial sialidase (from *Salmonella typhimurium* LT2) shows the same fold as an influenza virus neuraminidase. *Proc. Natl Acad. Sci. USA,* **90**, 9852–9856.

Davis, C. G., Goldstein, J. L., Sudhof, T. C., Anderson, R. G. W., Russell, D. W. & Brown, M. S. (1987). Acid-dependent ligand dissociation and recycling of LDL receptor mediated by growth factor homology region. *Nature,* **326**, 760–765.

Doolittle, R. F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314.

Downing, A. K., Knott, V., Werner, J. M., Cardy, C. M., Campbell, I. D. & Handford, P. A. (1996). Solution structure of a pair of calcium-binding epidermal growth factor-like domains: implications for the Marfan syndrome and other genetic disorders. *Cell,* **85**, 597–605.

Durkin, M. E., Wewer, U. M. & Chung, A. E. (1995). Exon organization of the mouse entactin gene corresponds to the structural domains of the polypeptide and has regional homology to the low-density lipoprotein receptor gene. *Genomics,* **26**, 219–228.

Edwards, Y. J. K. & Perkins, S. J. (1996). Assessment of protein fold predictions from sequence information: the predicted α/β doubly wound fold of the von Willebrand factor type A domain is similar to its crystal structure. *J. Mol. Biol.* **260**, 277–285.

Esser, V., Limbird, L. E., Brown, M. S., Goldstein, J. L. & Russell, D. W. (1988). Mutational analysis of the ligand binding domain of the low density lipoprotein receptor. *J. Biol. Chem.* **263**, 13282–13290.

Farquhar, M. G., Saito, A., Kerjaschki, D. & Orlando, R. A. (1995). The Heymann nephritis antigenic complex: Megalin (gp330) and RAP. *J. Am. Soc. Nephrol.* **6**, 35–47.

Fass, D., Blacklow, S., Kim, P. S. & Berger, J. M. (1997). Molecular basis of familial hypercholesterolaemia

from structure of LDL receptor module. *Nature,* **388**, 691–693.
Fox, J. W., Mayer, U., Nischt, R., Aumailley, M., Reinhardt, D., Wiedemann, H., Mann, K., Timpl, R., Krieg, T., Engel, J. & Chu, M.-L. (1991). Recombinant nidogen consists of three globular domains and mediates binding of laminin to collagen type IV. *EMBO J.* **10**, 3137–3146.
Gotoh, O. (1996). Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264**, 823–838.
Hafen, E., Basler, K., Edstroem, J.-E. & Rubin, G. M. (1987). *Sevenless*, a cell-specific homeotic gene of *Drosophila*, encodes a putative transmembrane receptor with a tyrosine kinase domain. *Science,* **236**, 55–63.
Herz, J., Clouthier, D. E. & Hammer, R. E. (1992). LDL receptor-related protein internalizes and degrades uPA-PAI-1 complexes and is essential for embryo implantation. *Cell,* **71**, 411–421.
Hobbs, H. H., Russell, D. W., Brown, M. S. & Goldstein, J. L. (1990). The LDL receptor locus in familial hypercholesterolemia: mutational analysis of a membrane protein. *Annu. Rev. Genet.* **24**, 133–170.
Hobbs, H. H., Brown, M. S. & Goldstein, J. L. (1992). Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. *Human Mutat.* **1**, 445–466.
Huber, A. H., Wang, Y. E., Bieber, A. J. & Bjorkman, P. J. (1994). Crystal structure of tandem type III fibronectin domains from *Drosophila neuroglian* at 2.0 Å. *Neuron,* **12**, 717–731.
Jones, D. T., Miller, R. T. & Thornton, J. M. (1995). Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins: Struct. Funct. Genet.* **23**, 387–397.
Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.
Kim, D.-H., Magoori, K., Inoue, T. R., Mao, C. C., Kim, H.-J., Suzuki, H., Fujita, T., Endo, Y., Saeki, S. & Yamamoto, T. T. (1997). Exon/Intron organization, chromosome localization, alternative splicing, and transcription units of the human apolipoprotein E receptor 2 gene. *J. Biol. Chem.* **272**, **(13)**, 8498–8504.
Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
Krämer, H., Cagan, R. L. & Zipursky, S. L. (1991). Interaction of *bride of sevenless* membrane-bound ligand and the *sevenless* tyrosine-kinase receptor. *Nature,* **352**, 207–212.
Krawczak, M. & Cooper, D. N. (1997). The human gene mutation database. *Trends Genet.* **13**, 121–122.
Krieger, M. & Herz, J. (1994). Structures and functions of multiligand lipoprotein receptors: macrophage scavenger receptors and LDL receptor-related protein (LRP). *Annu. Rev. Biochem.* **63**, 601–637.
Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E. & Sigler, P. B. (1996). The 2.0 Å crystal structure of a heterotrimeric G protein. *Nature,* **379**, 311–319.
Lestavel, S. & Fruchart, J. C. (1994). Lipoprotein receptors. *Cell. Mol. Biol.* **40**, **(4)**, 461–481.
Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507–533.
Long, M., de Souza, S. J. & Gilbert, W. (1995). Evolution of the intron-exon structure of eukaryotic genes. *Curr. Opin. Genet. Dev.* **5**, 774–778.
Mann, K., Deutzmann, R. & Timpl, R. (1988). Characterization of proteolytic fragments of the laminin-nidogen complex and their activity in ligand-binding assays. *Eur. J. Biochem.* **178**, 71–80.
Matsushime, H. & Shibuya, M. (1990). Tissue-specific expression of Rat c-*ros*-1 gene and partial structural similarity of its predicted products with *sev* protein of *Drosophila melanogaster*. *J. Virol.* **64**, **(5)**, 2117–2125.
Michael, W. M., Bowtell, D. D. L. & Rubin, G. M. (1990). Comparison of the sevenless genes of *Drosophila virili*s and *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA,* **87**, 5351–5353.
Murzin, A. G. (1992). Structural principles for the propeller assembly of β-sheets: the preference for 7-fold symmetry. *Proteins: Struct. Funct. Genet.* **14**, 191–201.
Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
Neer, E. J., Schmidt, C. J., Nambudripad, R. & Smith, T. F. (1994). The ancient regulatory-protein family of WD-repeat proteins. *Nature,* **371**, 297–300.
Norton, P. A., Hynes, R. O. & Rees, D. J. G. (1990). Sevenless: seven found? *Cell,* **61**, 15–16.
Parries, G., Chen, K., Misono, K. S. & Cohen, S. (1995). The human urinary epidermal growth factor (EGF) precursor. *J. Biol. Chem.* **270**, **(46)**, 27954–27960.
Patthy, L. (1987). Intron-dependent evolution: preferred types of exons and introns. *FEBS Letters,* **214**, 1–7.
Paulsson, M., Deutzmann, R., Dziadek, M., Nowack, H., Timpl, R., Weber, S. & Engel, J. (1986). Purification and structural characterization of intact and fragmented nidogen obtained from a tumor basement membrane. *Eur. J. Biochem.* **156**, 467–478.
Pereira, M. E. A., Mejia, J. S., Ortega-Barria, E., Matzilevich, D. & Prioli, R. P. (1991). The *Trypanosoma cruzi* neuraminidase contains sequences similar to bacterial neuraminidases, YWTD repeats of the low density liproprotein receptor, and Type III modules of fibronectin. *J. Exp. Med.* **174**, 179–191.
Riethmacher, D., Langholz, O., Godecke, S., Sachs, M. & Birchmeier, C. (1994). Biochemical and functional characterization of the murine *ros* protooncogene. *Oncogene,* **9**, 3617–3626.
Rost, B. (1995). TOPITS: threading one-dimensional predictions into three-dimensional structures. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology: Cambridge, UK* (Rawlings, C., ed.), pp. 314–321, AAAI Press, Menlo Park.
Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **266**, 525–539.
Russell, R. B. & Sternberg, M. J. E. (1995). How good are we? *Curr. Biol.* **5**, 488–490.
Sakai, J., Hoshino, A., Takahashi, S., Miura, Y., Ishii, H., Suzuki, H., Kawarabayasi, Y. & Yamamoto, T. (1994). Structure, chromosome location, and expression of the human very low density lipoprotein receptor gene. *J. Biol. Chem.* **269**, **(3)**, 2173–2182.

Šali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.

Saxena, K. & Shipley, G. G. (1997). Structural studies of detergent-solubilized and vesicle-reconstituted low-density lipoprotein (LDL) receptor. *Biochemistry,* **36**, 15940–15948.

Saxena, K., Gaitatzes, C., Walsh, M. T., Eck, M., Neer, E. J. & Smith, T. F. (1996). Analysis of the physical properties and molecular modeling of Sec13– A WD repeat protein involved in vesicular traffic. *Biochem. J.* **35**, 15215–15221.

Simon, M. A., Bowtell, D. D. L. & Rubin, G. M. (1989). Structure and activity of the sevenless protein: a protein tyrosine kinase receptor required for photoreceptor development in *Drosophila*. *Proc. Natl Acad. Sci. USA,* **86**, 8333–8337.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct. Funct. Genet*. **17**, 355–362.

Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E. & Sigler, P. B. (1996). Crystal structure of a G_A protein βγ dimer at 2.1 Å resolution. *Nature,* **379**, 369–374.

Soutar, A. K. (1992). Familial hypercholesterolaemia and LDL receptor mutations. *J. Int. Med.* **231**, 633–641.

Springer, T. A. (1997). Folding of the N-terminal, ligand-binding region of integrin α-subunits into a β-propeller domain. *Proc. Natl Acad. Sci. USA,* **94**, 65–72.

Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985a). The LDL receptor gene: a mosaic of exons shared with different proteins. *Science,* **228**, 815–822.

Sudhof, T. C., Russell, D. W., Goldstein, J. L. & Brown, M. S. (1985b). Cassette of eight exons shared by genes for LDL receptor and EGF precursor. *Science,* **228**, 893–895.

Sun, S., Footer, M. & Matsudaira, P. (1997). Modification of Cys-837 identifies an actin-binding site in the β-propeller protein scruin. *Mol. Biol. Cell,* **8**, 421–430.

Tan, K., Casasnovas, J. M., Liu, J.-h., Briskin, M. J., Springer, T. A. & Wang, J.-h. (1998). The structure of immunoglobulin superfamily domains 1 and 2 of MAdCAM-1 reveals novel features important for integrin recognition. *Structure,* **6**, 793–801.

Timpl, R. & Aumailley, M. (1993). Other basement membrane proteins and their calcium-binding potential. In *Molecular and Cellular Aspects of Basement Membranes* (Rohrbach, D. H. & Jimpl, R., eds), pp. 211–235, Academic Press, Orlando FL.

van der Voorn, L., Gebbink, M., Plasterk, R. H. A. & Ploegh, H. L. (1990). Characterization of a G-protein β-subunit gene from the nematode *Caenorhabditis elegans*. *J. Mol. Biol.* **213**, 17–26.

Van der Westhuyzen, D. R., Stein, M. L., Henderson, H. E., Marais, A. D., Fourie, A. M. & Coetzee, G. A. (1991). Deletion of two growth-factor repeats from the low-density-lipoprotein receptor accelerates its degradation. *Biochem. J.* **278**, 677–682.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–56.

Wall, M. A., Coleman, D. E., Lee, E., Iniguez-Lluhi, J. A., Posner, B. A., Gilman, A. G. & Sprang, S. R. (1995). The structure of the G protein heterotrimer $G_{i\alpha1\beta1\gamma2}$. *Cell,* **83**, 1047–1058.

Willnow, T. E., Hilpert, J., Armstrong, S. A., Rohlmann, A., Hammer, R. E., Burns, D. K. & Herz, J. (1996). Defective forebrain development in mice lacking gp330/megalin. *Proc. Natl Acad. Sci. USA,* **93**, 8460–8464.

Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R. W. W., Vriend, G., Thornton, J. M., Laskowski, R. A. & Macarthur, M. W. *et al*. (1998). Who checks the checkers: four validation tools applied to eight atomic-resolution structures. *J. Mol. Biol.* **276**, 417–436.

Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J., Coulson, A., Craxton, M., Dear, S., Du, Z. & Durbin, R. *et al*. (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature,* **368**, 32–38.

Zhang, X.-P., Puzon-McLaughlin, W., Irie, A., Kovach, N., Prokopishyn, N. L., Laferte, S., Takeuchi, K.-i., Tsuji, T. & Takada, Y. (1998). Integrin α3β1/laminin-5 interaction: critical role of a predicted loop $CNSNTDYLETGMC^{153\text{-}165}$ in the α3 third N-terminal repeat. *J. Biol. Chem.* In the press.

Edited by P. E. Wright

(Received 11 June 1998; received in revised form 27 July 1998; accepted 29 July 1998)