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

Timothy A. Springer

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 \( \beta \)-propeller structure. The six \( \beta \)-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 \( \beta \)-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 \( \beta \)-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; \( \beta \)-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).
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 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 (Matsushima & 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
Beta-sheets have also been termed beta-propellers; however, this is a misnomer because an alpha-helix is present between each sheet, and these should be termed beta-propellers. Beta-Propeller domains have diverse enzymatic, ligand-binding, and regulatory functions. Interest in this fold has recently heigh-
<table>
<thead>
<tr>
<th>Name</th>
<th>Ligands</th>
<th>Major expression</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDLR, Low density lipoprotein receptor</td>
<td>Lipoproteins containing ApoB-100 and ApoE</td>
<td>Liver, oocyte, cholesterol-regulated</td>
<td>Mammalia, Chondrichthyes, Amphibia</td>
</tr>
<tr>
<td>LDVR, very low density lipoprotein receptor, vitellogenin receptor of chicken and Xenopus</td>
<td>Lipoproteins containing ApoE, vitellogenin</td>
<td>Muscle, heart, white fat, oocyte</td>
<td>Mammalia, Aves, Amphibia, C. elegans</td>
</tr>
<tr>
<td>LRP1, LDLR-related protein 1</td>
<td>Remnant lipoproteins enriched in ApoE, proteinase: α2- macroglobulin complexes, urokinase and plasminogen activator: plasminogen activator inhibitor 1 complexes, lipoprotein lipase, lactoferrin</td>
<td>Liver, brain, lung, ovary, placenta</td>
<td>Mammalia, Aves, C. elegans</td>
</tr>
<tr>
<td>LRP2, LDLR-related protein 2, GP330, Megalin</td>
<td>Lipoproteins enriched in ApoE, plasminogen activator, lipoprotein lipase, lactoferrin</td>
<td>Kidney</td>
<td>Mammalia, C. elegans</td>
</tr>
<tr>
<td>LR11, LDL-like receptor containing 11 type A repeats</td>
<td>Lipoproteins containing ApoE</td>
<td>Brain</td>
<td>Mammalia, Aves</td>
</tr>
<tr>
<td>LR8, LDL-like receptor containing 7 or 8 type A repeats</td>
<td>Lipoproteins containing ApoE</td>
<td>Brain</td>
<td>Mammalia, Aves</td>
</tr>
<tr>
<td>YL, Yolkless, vitellogenin receptor</td>
<td>Vitellogenin</td>
<td>Oocyte</td>
<td>Drosophila</td>
</tr>
<tr>
<td>Nidogen, entactin</td>
<td>Laminin, collagen IV, perlecan</td>
<td>Basement membranes</td>
<td>Mammalia, Ascidia, Mammalia</td>
</tr>
<tr>
<td>Nidogen-2, osteonidogen</td>
<td></td>
<td>Bone matrix</td>
<td>Mammalia</td>
</tr>
<tr>
<td>EGF, epidermal growth factor precursor</td>
<td>EGF receptor</td>
<td>Kidney, submaxillary gland</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Sevenless</td>
<td>Bride of sevenless, BOSS</td>
<td>Photoreceptor R7 precursor cell</td>
<td>Drosophila</td>
</tr>
<tr>
<td>c-ros, Proto-oncogene related to v-ros of avian sarcoma virus</td>
<td></td>
<td>Epithelia</td>
<td>Mammalia, Aves</td>
</tr>
</tbody>
</table>
tended 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, scrin (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 β2-propellers and form domains of ~240 residues. The leucine-rich repeats of 20 residues contain one z-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 ≤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 \( \beta \)-strand or \( \alpha \)-helix with PHD (Rost, 1996) are highlighted in gold and magenta, respectively. Prediction of \( \beta \)-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 (\( p \geq 2 \times 10^{-10} \)) to lowest at bottom (\( p = 5 \times 10^{-4} \)). \( \beta \)-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 consen-
sus, 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 con-
tiguous 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 LDLR (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 exon boundaries. 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).

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 TOPTITS, to predict protein fold (Rost, 1995). This method is distinct from potential-based threading (see below). The top hit with TOPTITS 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.
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

![Figure 4](attachment:image.png)

*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 gray. 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).
Table 2. Fold prediction for YWTD domains with TOPITS

<table>
<thead>
<tr>
<th>Rank</th>
<th>Z-score</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.84</td>
<td>1got 7-blade β-propeller, G protein β-subunit</td>
</tr>
<tr>
<td>2</td>
<td>2.83</td>
<td>1nr 8-blade β-propeller, cd1 nitrite reductase</td>
</tr>
<tr>
<td>3</td>
<td>2.73</td>
<td>1nf FN3 domains 7-10, fibroactinin</td>
</tr>
<tr>
<td>4</td>
<td>2.68</td>
<td>2bbk 7-blade β-propeller, methylamine dehydrogenase</td>
</tr>
<tr>
<td>5</td>
<td>2.51</td>
<td>1gof 7-blade β-propeller, galactose oxidase</td>
</tr>
<tr>
<td>6</td>
<td>2.46</td>
<td>2sil 6-blade β-propeller, <em>Salmonella typhimurium</em> sialidase</td>
</tr>
<tr>
<td>7</td>
<td>2.41</td>
<td>1rcs 6-blade β-propeller, Influenza virus Beijing neuraminidase</td>
</tr>
<tr>
<td>8</td>
<td>2.38</td>
<td>2ncs 6-blade β-propeller, avian Influenza virus neuraminidase</td>
</tr>
</tbody>
</table>

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) 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, prolyleucine 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 THREERAD 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 PHYLIP (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
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

<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
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 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 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 β-strand 1 of each sheet (Figures 2A, 7A).

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

![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 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)/SD) 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<1945). d The 89 LDLR-related YWTD domains. e The 18 sevenless and c-ros YWTD domains.

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

<table>
<thead>
<tr>
<th></th>
<th>Z-score</th>
<th>Pb</th>
<th>Ex</th>
</tr>
</thead>
<tbody>
<tr>
<td>89 YWTD</td>
<td>4.52</td>
<td>8.5×10⁻⁷</td>
<td>1.6×10⁻³</td>
</tr>
<tr>
<td>18 YWTD</td>
<td>4.20</td>
<td>3.7×10⁻⁶</td>
<td>7.2×10⁻³</td>
</tr>
</tbody>
</table>

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)/SD) was determined.
YWTD Repeats Fold into a Six-bladed $\beta$-Propeller

Table 5. Threading scores for FN3 repeats

<table>
<thead>
<tr>
<th>Repeat</th>
<th>Top hit</th>
<th>Z-score</th>
<th>Code</th>
<th>Z-score</th>
<th>2nd hit</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN3, 4 c-ros</td>
<td>1fna00*</td>
<td>5.22</td>
<td>1fvcA0*</td>
<td>2.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN3, 4 sevenless</td>
<td>3hhrB2*</td>
<td>3.68</td>
<td>1fn02*</td>
<td>3.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN3, 9 c-ros</td>
<td>1cb01*</td>
<td>4.51</td>
<td>2prA3*</td>
<td>2.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN3, 9 sevenless</td>
<td>1fn02*</td>
<td>3.66</td>
<td>1gca01*</td>
<td>2.57</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See footnote to Table 3. FN3 repeat sequences defined as in Figure 5 for chicken, human, rat and mouse c-ros and D. melanogaster and D. viridis 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).

* FN3 domains are 1fna00, fibronectin FN3 domain 10; 3hhrB2, growth hormone receptor FN3 domain 2; 1cb01, neuroglian FN3 domain 1; 1fn02, fibronectin FN3 domain 8.

† Non-FN3 domains are 1fvcA0, antibody Fv fragment immunoglobulin V domain; 2prA3, trypanothione reductase $\alpha$ + $\beta$ domain; 1gca01, galactose-binding protein $\alpha$/$\beta$ domain.

neuraminidase have been proposed as YWTD motifs (Pereira et al., 1991), subsequent bacterial neuraminidase structures (Cremnell et al., 1993) show that these sequences are not structurally homologous to one another, and are not in $\beta$-strand 2 as predicted here for the YWTD motif.

Representative models were made with SEG-MOD of LOOK for the YWTD domains of human LDLR, human nidogen (Figure 4B and C), and YWTD domain 7 of chicken LRPI, 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 $\beta$-propeller domain that contains WD40 repeats with low but significant sequence homology to the G-protein $\beta$-subunit $\beta$-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 ldr, 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, ldr, 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 ldr, 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 ldr, 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 \(\beta\)-propeller template made from the transducin \(\beta\)-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 \(\beta\)-strand in 1got as defined by DSSP (Kabsch & Sander, 1983), and in the sequences as predicted by PHD (Rost, 1996) are in gold. \(\beta\)-Sheet ladder positions are marked for Gb; \(b\) is for bulge and varies in position in strand 4 and may differ between Gb 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. \(\beta\)-Strands as defined by DSSP are in gold. The W were cut out of 1kit and structurally aligned with 3DALIGN 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 \(\beta\)-strands. Cysteine residues predicted to be disulfide-bonded are color coded; cysteine residues with equivalent \(\beta\)-sheet ladder positions are given the same color. lr11_hu_w1 is representative of W1 of LR11_HU1, LR11_CH1, and LR11_RB1. lr11_hu_w2 is representative of W2 of LR11_HU1, LR11_CH1, LR11_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

<table>
<thead>
<tr>
<th>Structure or model</th>
<th>Residues</th>
<th>Prosa II Z-score</th>
<th>QUACHK score</th>
<th>NQACHK score</th>
</tr>
</thead>
<tbody>
<tr>
<td>G protein β-propeller domain†</td>
<td>296</td>
<td>−9.67</td>
<td>−0.270</td>
<td>−1.45</td>
</tr>
<tr>
<td>Six-bladed β-propeller template†</td>
<td>257</td>
<td>−9.74</td>
<td>−0.208</td>
<td>−2.31</td>
</tr>
<tr>
<td>ldlr, model of LDLR_HU1†</td>
<td>264</td>
<td>−5.34</td>
<td>−0.569</td>
<td>−3.65</td>
</tr>
<tr>
<td>lrp1, model of LRPI.CH§</td>
<td>260</td>
<td>−6.70</td>
<td>−0.651</td>
<td>−3.65</td>
</tr>
<tr>
<td>nido, model of NIDO_HU1†</td>
<td>266</td>
<td>−6.22</td>
<td>−0.549</td>
<td>−3.48</td>
</tr>
<tr>
<td>Sec-13, 6-bladed WD40 domain modeled here§</td>
<td>282</td>
<td>−8.89</td>
<td>−0.876</td>
<td>−4.07</td>
</tr>
<tr>
<td>Sec-13, 6-bladed WD40 domain published model§</td>
<td>282</td>
<td>−7.43</td>
<td>−1.360</td>
<td>−5.04</td>
</tr>
<tr>
<td>lrp1-mis, control mistransferred model§</td>
<td>260</td>
<td>−5.73</td>
<td>−1.793</td>
<td>−3.75</td>
</tr>
</tbody>
</table>

* Prosa II combined C β pairwise and surface potential Z-scores relative to the pII3.0.shrt.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.

† 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.

§ 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.

Agreement with experimental data

Model predictions were tested against biochemical data. Circular dichroism of protein, 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 4A 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 A 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
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 \( \beta \)-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 395 currently known protein folds arranges a single polypeptide into a single domain with six structurally similar units, the six-bladed \( \beta \)-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 \( \beta \)-propeller domain, with an expectation that this could occur by chance alone of \( 1.6 \times 10^{-3} \) (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 \textit{sevenless} and \textit{c-ros}; it should be noted that a relationship has been reported with only two isolated YWTD repeats in \textit{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 \( \beta \)-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 \( \beta \)-sheets of the \( \beta \)-propeller. The C1124–C1135 disulfide bridge between \( \beta \)-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 \( \beta \)-strands 2 and 3, and between \( \beta \)-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 \( 5.6 \times 10^{-4} \).

The structure quality evaluation programs used here readily “flunk” deposited X-ray structures that have one residue structure to sequence frame-shifts, 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 7.** Frequency of sequence-structure distances permissive for disulfide bond formation

<table>
<thead>
<tr>
<th>Disulfide</th>
<th>Residues</th>
<th>Sequence separation</th>
<th>Structure distance (Å)</th>
<th>Frequency*</th>
</tr>
</thead>
<tbody>
<tr>
<td>C987–C1205(^b)</td>
<td>C987, L1202</td>
<td>215</td>
<td>C(^a) &lt; 15</td>
<td>0.033</td>
</tr>
<tr>
<td>C1124–C1135(^a)</td>
<td>C1124, C1135</td>
<td>11</td>
<td>C(^b) &lt; 6</td>
<td>0.017</td>
</tr>
</tbody>
</table>

* Distance distributions were calculated by Azat Badreddinov and Andrej Sali, Rockefeller U., with a non-redundant set of all chains in the protein database that are <30% identical in sequence and have \( \geq 200 \) residues. These 456 chains are structurally diverse and represent all of the structure classes.

\(^a\) 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 editing 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\(^a\) 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\(^b\) atom distance distributions were calculated for all residues that have more than 10% surface exposure, and are 210 to 230 sequence positions apart. Results were similar with a small set of all \( \beta \) structures.
<table>
<thead>
<tr>
<th>Evidence</th>
<th>Conclusion</th>
<th>Probability, expectation, frequency, or strength of conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Cysteine-to-size rule</td>
<td>YWTD repeats are not separate domains</td>
<td>High</td>
</tr>
<tr>
<td>(2) Always six YWTD repeats</td>
<td>Six YWTD repeats form a single domain</td>
<td>Moderate</td>
</tr>
<tr>
<td>(3) Exons of same phase at domain boundaries, not at repeat boundaries</td>
<td>Modular unit is group of six repeats, not individual repeats</td>
<td>High</td>
</tr>
<tr>
<td>(4) Nidogen electron microscopy</td>
<td>YWTD repeats form a single globular domain</td>
<td>Very high</td>
</tr>
<tr>
<td>(5) YWTD repeats have sequence homology</td>
<td>Repeats have similar structure</td>
<td>Very high</td>
</tr>
<tr>
<td>(6) (1, 2, 3, 4) + 5: Single, globular domain with 6 structurally similar units</td>
<td>Consistent with only one of 393 known folds: 6-bladed β-propeller domain</td>
<td>$F = 0.003$</td>
</tr>
<tr>
<td>(7) Prediction of groups of six YWTD repeats in <em>sevenless/c-ros</em></td>
<td>Hypothesis has heuristic value</td>
<td>High</td>
</tr>
<tr>
<td>(8) YWTD domains have 24 predicted β-strands, with 4/repeat</td>
<td>Agrees with 6-bladed β-propeller domain</td>
<td>High</td>
</tr>
<tr>
<td>(9) Fold prediction with TOPTIFS: 7 of 8 top hits are β-propeller</td>
<td>β-Propeller domain</td>
<td>High</td>
</tr>
<tr>
<td>(10) Circular dichroism of nidogen shows all-β-structure</td>
<td>Agrees with β-propeller domain</td>
<td>Moderate</td>
</tr>
<tr>
<td>(11) THREADER normalized Z-score for LDLR-related YWTD domains of 4.52</td>
<td>β-Propeller domain</td>
<td>$E = 1.6 \times 10^{-3}$</td>
</tr>
<tr>
<td>(12) THREADER normalized Z-score for <em>sevenless/c-ros</em> YWTD domains of 4.20</td>
<td>β-Propeller domain</td>
<td>$E = 7.2 \times 10^{-3}$</td>
</tr>
<tr>
<td>(13) Model of nidogen agrees with experimentally determined disulfides</td>
<td>Six-bladed β-propeller domain with largely correct sequence alignment</td>
<td>$P = 5.6 \times 10^{-4}$</td>
</tr>
<tr>
<td>(14) Representative models have good structural quality</td>
<td>Six-bladed β-propeller domain with largely correct sequence alignment</td>
<td>Extremely high</td>
</tr>
</tbody>
</table>
YWTD Repeats Fold into a Six-bladed β-Propeller

YWTD Repeats Fold into a Six-bladed β-Propeller

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 0.005 of all folds; the probability that these three lines of evidence would all yield the same fold by chance alone is $3 \times 10^{-9}$. 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 $3 \times 10^{-9}$ and $2 \times 10^{-11}$. 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 LR1, LR2, and yolkes, 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 $\beta$-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 $\beta$-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 subunit $\alpha_2\beta_1$ appears to bind laminin through loops on the top of its predicted $\beta$-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 $\beta$-VLDL binding sites in the class A repeats, for resistance of the receptor to degradation under basal conditions or during endocytosis of $\beta$-VLDL, and for acid-induced dissociation of the degradation-$\beta$-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 $\beta$-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 $\beta$-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 $\beta$-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 $\beta$-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 $\beta$-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.
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, 211875.

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 initiation codon, were obtained by querying the Human Gene Mutation Database, Cardiff UK (http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html) (Krąwczak & 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 Cα 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.
vitelligenin 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; LR1P1_CH, LR1P1_CHICK; LR1P1_HU, LR1P1_HUMAN; LR1P1_MO, 49442; LR2P2_RA, LR2P2_RAT; LR2P2_CAE, LR2P2_CEAEL; NIDO_HR, NIDO_MOUSE; NIDO MOUSE; ONID_HU, 1449167 human osteonidogen; YL_DM, YL_DROME, NIDO_MO, NIDO_MOUSE; ONID_HU, 1449167 human osteonidogen; NIDO_HU, NIDO_HUMAN; cynthia roretzi LR11_CH, 1552268, chick LR11; LR11_HU, 1552324, human LR11; LR11_RB, X97001 chick LR8B; LR11_HU, 1552324, human LR11; LR11_RB, X. laevis vitellogenin receptor of 858 D. melanogaster sevenless SEV_DM, 7LES_DROME, Drosophila melanogaster ros SEV_CH, TVCHSR, chicken c- SEV_DV, 7LES_DROVI, Drosophila virilis sevenless; SEV_CH, TVCHSR, chicken c-ros; SEV_HV, TVHURS, human c-ros; SEV_MO, 211769; 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 LRPI, designated here LRPI_CE, is predicted to be encoded by the cosmids F47B3 (U97017) and T21E3 (AF003135), 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 genomic sequences, respectively.

A homologue of LDVR, designated here LDVR_CE here, is encoded by cosmid T13C2 (U40030). The top four BLAST2 hits were LDVR_MOUSE, LDVR_RAT, LDVR_RAABT, 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 NPXY 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, CAAAGgtat, 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 (Z75555) 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; LRPI; LR2P2; LR8; LR11; and EGF precursor. In LRPI, LR2P2, and yokless, 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 (Gotth, 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 (a = 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 (Gotth, 1996). The eight phylogenetically most distant sequences had <21% identity to the three target sequences for structure prediction; these sequences, LRPI_CE1, LRPI_CE4,
content: LRPI_CE5, LRPI_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 LRP's, 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. identity were EGF_HU2, n = 79, x = 31(±9)%; LDLR_HU1, n = 79, x = 36(±14)%; NIDO_HU1, n = 73, x = 29(±5)%; consensus, n = 81, x = 44(±6)%. These alignments were submitted to PHD (Rost, 1996) (http://www.emb地狱.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, 1bg, 2bat, 2sil, 4ah, 1eur, 1got, 1ldo, 1lm, 1kit, 1nir, 1ncn, 1uxo, 1tit, 1wit, a 1.28 Å structure of cytochrome cF from Thioplasma pantotroph (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 of the z axis, and permuted to place residues 45 to 54 after residues 55 to 340, so that all sheet of 7 was modeled at the end of the pdb file (1bg_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 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 1bg_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 Modeller (Sali & Blundell, 1993) (http://guitar.rockefeller.edu/modeller/modeller.html). Two alignments were used to generate restraint files: A B C D E 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_m60.

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 (m60) model, using framework residues that align with the KIYAMHW, LLLS, KLIWID, 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_m60_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 m60_s2 and m60_s4.

Final models were made using the sequence-structure alignment shown in Figure 7, with either w1-6_on_1tbg_per_sym_166_m60_s2 or m60_s2 as templates. A model of yeast Sec-13 used m60_s4 as template, except that the template pdb file was permitted 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; LRPI_CH7, 1px; NIDO_HU1, 1npx.

Acknowledgments

I am deeply grateful to Azat Badredtinov 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 TOPTITS, and to Osamu Gotoh for support of PRRP. I appreciate very much 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


*Edited by P. E. Wright*

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