Molecular Vision 2000; 6:30-9 <http://www.molvis.org/molvis/v6/a6>
Received 1 November 1999 | Accepted 28 February 2000 | Published 7 April 2000

Prediction of structural and functional relationships of Repeat 1 of human interphotoreceptor retinoid-binding protein (IRBP) with other proteins

Eleanore A. Gross[1], Gui Ru Li[1], Ze-Yu Lin[1], Sarah E. Ruuska[1], Jeffrey H. Boatright[1], I. Saira Mian[2], John M. Nickerson[1]

[1]Department of Ophthalmology, Emory University, Atlanta, GA; [2]Department of Cell and Molecular Biology, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA

Purpose: We compared the structure and function of interphotoreceptor retinoid-binding protein (IRBP) related proteins and predicted domain and secondary structure within each repeat of IRBP and its relatives. We tested whether tail specific protease (Tsp), which bears sequence similarity to IRBP Domain B, binds fatty acids or retinoids, and whether IRBP possessed protease activity resembling Tsp's catalytic function. These tests helped us to learn whether the primary sequence similarities of family members extended to higher order structural and functional levels.
Methods: Predictions derived from multiple sequence alignments among IRBP and Tsp family members and secondary structure computer programs were carried out. The first repeat of human IRBP (EcR1) and Tsp were expressed, purified, and tested for binding properties. Tsp was examined for fluorescence enhancement of retinol or 16-anthroyloxy-palmitic acid (16-AP) to test for ligand binding. IRBP was tested for protease activity.
Results: Tsp did not exhibit fluorescence enhancement with retinol or 16-AP. IRBP did not exhibit protease activity. The positions of critical residues needed for the ligand binding properties of retinol were predicted. Primary sequence and three-dimensional similarity was found between Domain A of IRBP Repeat 3 and eglin c.
Conclusions: The sequence similarity of Tsp and IRBP raised the possibility that each might share the function of the other protein: IRBP might possess protease activity or Tsp might possess retinoid or fatty acid binding activity. Our studies do not support such a shared function hypothesis, and suggest that the sequence similarity is the result of maintenance of structure. The finding of similarity to eglin c in Domain A suggests the possibility of a tight interaction between Domain A and Domain B, possibly implying the need for Domain A in retinoid-binding, and suggesting that both Domains should be present in testing mutations. The positions of predicted critical amino acids suggest models in which a large binding pocket holds the retinoid or fatty acid ligand. These predictions are tested in a companion paper.

Mammalian interphotoreceptor retinoid-binding protein (IRBP) is a high molecular weight glycoprotein of approximately 140 kDa with a four-fold repeat structure. Each repeat has two domains; the first 80 residues of each repeat form "Domain A," possibly involved in the regulation of ligand binding, and the remaining 220 amino acids form "Domain B," which contains the ligand binding sites. Domain B is separated from Domain A by a stretch rich in prolines [1]. Domain A corresponds to the initial part of Repeat 4 encoded by Exon 1, Domain B is formed from the remaining three exons that constitute Repeat 4. Baer, et al. [2] found no retinol binding activity in an Exon 1/Domain A protein, but they found retinol binding activity in partial Domain B protein representing protein from Exons 2 plus 3. Also, they found limited binding activity in an Exon 4 protein [2].

A weak but statistically significant primary sequence similarity exists between IRBP Domain B and *E. coli* tail specific protease (Tsp), other bacterial and eukaryotic proteases, and the archeon Thermoplasma acidophilium Tricorn Protease [3,4]. Generally, the proteases of this family are serine proteases that cleave C-terminal hydrophobic amino acids, known as C-terminal processing proteases (CTPs). There is no similarity between Domain A and these proteases, and until this present report there were no identified similarities of Domain A to any other protein.

Members of a family of proteins sharing a common ancestor often exhibit shared functions. Because of the sequence similarity between IRBP and the CTPs, we asked whether IRBP exhibited any protease activity, as many members of the CTP-IRBP domain family apparently are C-terminal proteases. Mutational studies of Tsp suggest that the region equivalent to IRBP Domain B has a role in Tsp's catalytic function [5]. Proteolytic activity has not been examined in IRBP. Though IRBP does not retain any of the conserved amino acids impli-

Correspondence to: John M. Nickerson, PhD, Room B5602, Emory Eye Center, Emory University, 1365B Clifton Road, N.E., Atlanta, GA, 30322; Phone: (404) 778-4411; FAX: (404) 778-2231; email: litjn@emory.edu

Dr. Gross is now at the Picower Institute for Medical Research, 350 Community Drive, Manhasset, NY, 11030

cated in Tsp active site formation [5], Tricorn Protease has one of the three conserved amino acids [4], suggesting that it is reasonable to test IRBP for protease activity.

The Domain B/Tsp sequence similarity may suggest that Tsp may possess retinol and fatty acid binding capabilities. IRBP binds various long chain fatty acids [6,7] and retinoids [7-11]. Here we report that Tsp does not bind either retinol or 16-AP (16-anthroyloxy-palmitic acid, a fluorescent fatty acid analog), suggesting that Tsp cannot bind either tested ligand. Also, IRBP lacks general protease activity with casein as a substrate. Thus, neither family member seems to retain shared functions, implying that the conserved amino acids help to maintain the same structure.

We additionally report that a predicted tertiary structure of Domain A appears to be quite similar to the established tertiary structure of eglin c, a small canonical serine protease inhibitor. Canonical protease inhibitors are proteins that bind to the active site of a protease thereby competing with the substrate. The inhibitor may be cleaved by the protease. Thus, it may be useful for predictive purposes to consider a single IRBP repeat as a tightly bound complex of a protease (Domain B) and protease inhibitor (Domain A).

Given this structural and functional information on the IRBP repeat, we predicted amino acids needed for retinoid binding and tertiary structure. The similarity in structure between Domain A and eglin c, implying a possible tight interaction between Domains A and B, suggest that some conserved sequences are needed to ensure tight domain-to-domain contacts. Without shared functions between Tsp and IRBP, we predicted that conserved sequences between the two proteins are needed to maintain secondary and tertiary structure, and thus amino acids shared among IRBP Repeats but not conserved across Tsp to IRBP are more likely to be important for retinoid-binding in IRBP. These data led us to predict that several critical residues in Domain B are involved with retinoid binding. These predictions were tested in a companion paper [12].

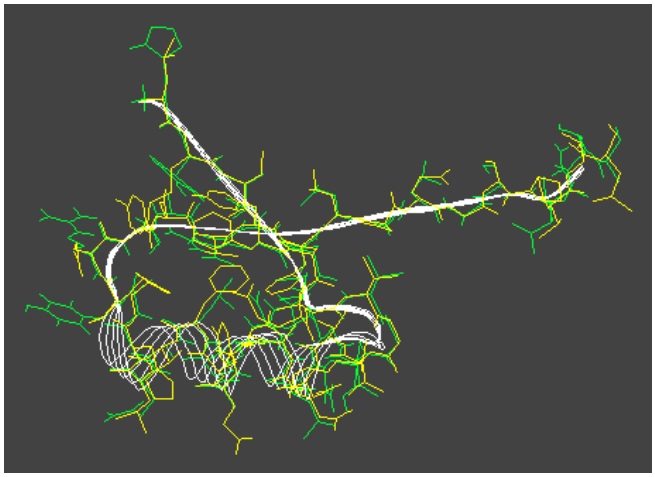

Figure 1. Domain A modelled using the three dimensional structure of eglin c. The yellow lines represent eglin c (entry ID 1EGP from PDB) and the green structure is Domain A from Repeat 3. Domain A represents the first 80 amino acids of Repeat 3 from human IRBP.

METHODS

Protease assays: To test for protease activity, the Pierce (Rockford, IL) Quanticleave Protease Assay kit was employed as per the manufacturer's instructions.

Tsp purification: An *E. coli* Tsp clone was the kind gift of Dr. Robert Sauer (Massachusetts Institute of Technology, Cambridge, MA). The protein was expressed and purified according to Keiler and Sauer [5]. Protein expression from pKK101 was induced in *E. coli* with isoproyl-β-D-thiogalactoside (IPTG), and after 3 h, bacteria were harvested. Tsp was recovered by 6 M guanidine hydrochloride extraction, the soluble proteins were passed through a metal ion affinity chromatography column, and the bound proteins were eluted with 250 mM imidazole. This fraction was passed through a QAE column, and the resulting Tsp was about 90% pure [5].

Original wild type Repeat 1 construct for E. coli*expression:* As described previously [13], a *BamH*I restricted PCR product representing the first repeat of human IRBP was subcloned into an *E. coli* expression vector, pLEX (Invitrogen, La Jolla, CA) to create pLexR1. The *E. coli* expressed Repeat 1 (EcR1) protein product has a 7 amino acid N terminal extension derived from the pLEX expression vector, followed by a 5 amino acid propeptide, 300 first repeat amino acids, and a C terminal 6 histidine amino acid tag, for a total of 318 amino acids. This construct does not contain the signal peptide or the initial methionine of the native IRBP. The calculated molecular mass of wild type EcR1 is 34,656.5 Da. The protein was produced and purified as previously described [13].

Ligands: All-*trans*-retinol, (ε = 46,000 at 325 nm, >95% pure, Sigma, St. Louis, MO) and 16-AP (ε = 8000 at 363 nm, >90% pure, Molecular Probes, Inc., Portland, OR) were dissolved in ethanol and used under subdued red light (Kodak 1A safelight, Eastman Kodak, Rochester, NY).

Fluorescence measurements: Equilibrium fluorescence measurements were performed at room temperature in subdued red light on solutions of EcR1 and Tsp at 1 μM, according to previously described methods [13,14].

RESULTS

The results are presented in two sections. In Section 1, we describe computer analyses of human IRBP and make predictions derived from these analyses. In Section 2, we compare Tsp and recombinant human IRBP in tests of ligand binding and protease activity.

Section 1: Computer analyses of human IRBP: Several programs were used to examine the primary sequence of human IRBP. The value of their output was in predicting the secondary structure, tertiary structure, and family relationships of the query sequences. We examined the output in order to

predict specific amino acids needed in binding retinol, and to predict domains of IRBP that are not engaged in binding retinol. The programs included PredictProtein [15], used to find proteins related in primary, secondary, and tertiary structure; hydrophobic cluster analysis (HCA) [16] used to predict common secondary structural features; BLAST [17] used to find related sequences; Protein Structure Analysis (PSA) [18] used to predict secondary structure and class of the protein; JPred [19] used to predict secondary structure; and Hidden Markov Models (HMMs, [20]), used to predict related sequences at the primary sequence level.

Possible tertiary structure of Domain A from Repeat 3: Based on primary sequence similarity, PredictProtein suggested that Domain A of Repeat 3 was similar to eglin c, a small canonical serine protease inhibitor with a well known tertiary structure. A comparison of the predicted tertiary structures of Domain A and eglin c is shown in Figure 1. PredictProtein detected only eglin c from PDB as similar to Domain A in the database, and, in general, the PredictProtein program is conservative [15]. The program Swiss-Model [21] was used to carry out homology modelling and to create Figure 1. Figure 1 shows the general close alignment of the alpha

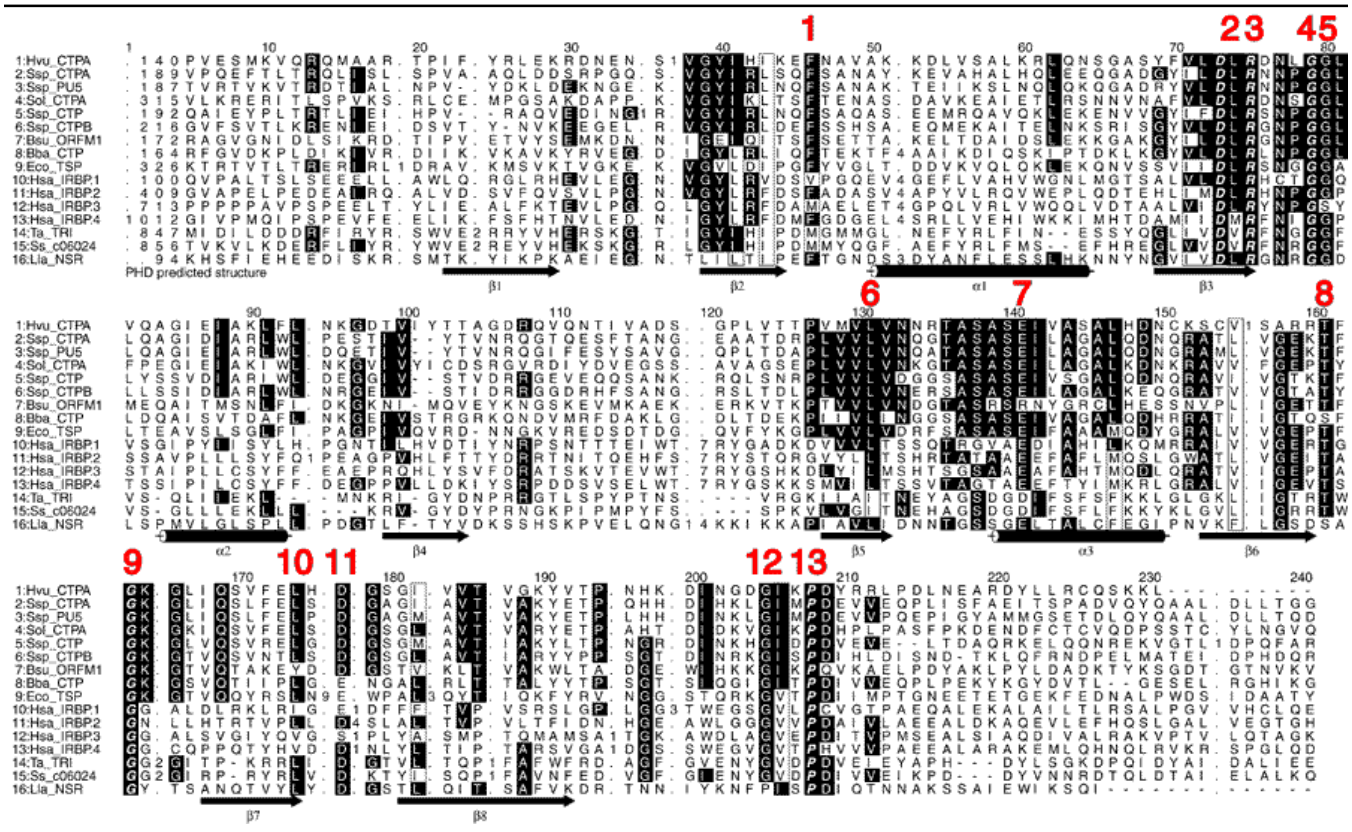

Figure 2. HMM multisequence alignment of the catalytic Domain of Tail Specific Protease and Family members, including IRBP Domain B. Within the 300 amino acid long IRBP repeat unit, there appear to be identifiable structural or folding domains. Domain A in IRBP corresponds to the first 80 amino acids of a repeat. Domain B in IRBP begins near the boundary of Exon 1 and 2 in Repeat 4 and contains amino acids 80-300. Tsp lacks the equivalent eglin c-like structure of Domain A. However, in Tsp and the other carboxy-terminal proteases, they contain a PDZ domain in place of the eglin-like Domain A. Carbohydrate attachment sites are found in the repeat mainly at exon exon boundaries. Predicted secondary structures are labeled on the figure as predicted by PHD. Very highly conserved amino acids are marked by black boxes. All these sequence features are superimposed on an alignment of sixteen different family members. Organisms represented include human, plants, bacteria, cyanobacteria, and archea. Numbers correspond to amino acids changed for mutants 1 through 13 in human Repeat 1 (which are described in a companion paper [12]), sequence 10: Hsa_IRBP.1. Amino acid changes of human Repeat 1 made in the companion paper [12] are shown in red numbers: 1: V116N, 2: L147A, 3: R148D, 4: G152A, 5: G153A, 6: L208A, 7: E218A, 8: T237A, 9: G239T, 10: I249A, 11: E251A, 12: G278A, 13: P281A. The proteins shown are as follows: Hvu_CTPA is the C-terminal peptidase (Ctp) from the vascular plant *Hordeum vulgare*(X90929); Ssp_CTPA is the C-terminal processing proteinase precursor from cyanobacteria *Cyanobacterium synechocystis* (A53964); Ssp_PU5 is a hypothetical protein from the petBD operon (P42784); Ssp_CTP (D90906) and Ssp_CTPB (X96490) are other CTPs from that same organism; Sol_CTPA (X90558) is the D1 precursor CTP from the spinach plant (*Spinacia oleracea*); Bsu_ORFM1 is a CTP from *Bacillus subtilis*(AF015775 and X98341); Bba_CTP is a CTP from Bartonella bacilliformis (L37094); Eco_TSP (D90827 also known as prc, D00674), is Tail-Specific Protease of *Escherichia coli*; Hsa_IRBP.1 is Repeat 1 of human IRBP; Hsa_IRBP.2 is Repeat 2 of human IRBP; Hsa_IRBP.2 is Repeat 3 of human IRBP; Hsa_IRBP.4 is Repeat 4 of human IRBP (the accesion number of human IRBP is M22453); Ta_TRI is the Tricorn Protease of archaeon *Thermoplasma acidophilium*(U72850); Ss_c06024 (Y08256) is a genome sequence from the archaeon *Sulfolobus solfataricus*. Lla_NSR is the nisin-resistance protein from gram-positive bacterium *Lactococcus lactis* (U25181).

carbons and the similar orientation of the side chains of almost every amino acid. This general agreement in the positions of corresponding amino acids between eglin c and the model of Domain A suggested that Domain A possesses the same three-dimensional structure as eglin c.

Primary sequence alignments: A similarity score between Tsp, a C-terminal processing protease (CTP) and IRBP was first reported by Silber, et al. [3] to be 8.7 standard deviations greater than the mean similarity calculated for other database sequences. Using Domain B of human Repeat 1 as the

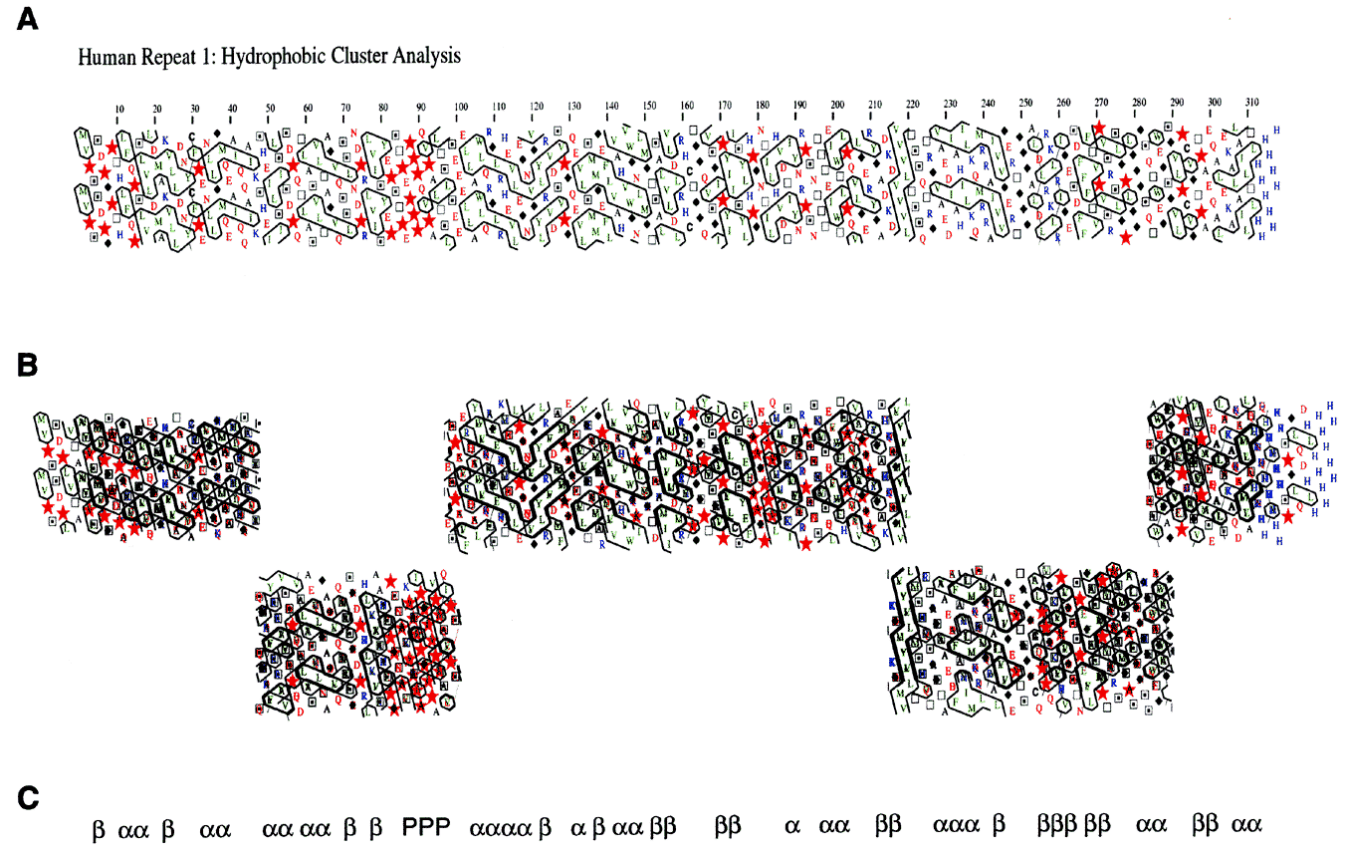

Figure 3. Hydrophobic cluster analysis (HCA) of the four repeats of human IRBP. A primary amino acid sequence is written downward at an angle of about 12.5° from vertical with 7 or 8 amino acids per line, representing about two turns of an α-helix. A second copy of the amino acid sequence is also printed, but it is shifted in phase by 3.5 amino acids. This representation displays amino acids adjacent to each other on the horizontal dimension that might be near each other if they were found in an α-helix. Hydrophobic amino acids are displayed in green and clusters of these amino acids, which include V, I, L, F, M, Y, and W, are boxed by black contour lines. Other amino acids are represented as follows: Red stars, P; black diamonds, G; open boxes, T; boxes with a black dot in the center, S; blue coloring represents basic amino acids (R, K, and H); red letters indicate the acidic amino acids (D, E) and their uncharged counterparts (Q, N). Black amino acids include A and C. The patterns of the contour lines in certain cases are strongly associated with either α-helix or β-strand [16,23]. **A** shows the hydrophobic cluster analysis of EcR1. Note the clear separation of putative Domain A (amino acids 1-80) from Domain B (amino acids 90-310) by the proline-rich region at about position 85. Positions 100 to 300 correspond to the sequence Hsa_IRBP.1 shown in Figure 2. **B** shows the conservation of hydrophobic clusters in an alignment of HCAs from all four Repeats (EcR1, EcR2, EcR3 and EcR4). The alignment of the four sequences was done in five blocks to allow four gaps to be introduced at positions likely to contain loops or turns of variable length among the four different sequences. The heavy black lines indicate overlapping contours that are identically positioned among the four sequences. Conserved clusters become obvious and many are associated with one type of secondary structure as predicted in **C**. The secondary structure assignments were based on the 17 classes identified by Lemesle-Varloot, et al. [16]. For example, the vertical stripes of hydrophobic amino acids at positions 215-220 in EcR1, and well conserved in the other three repeats, was classified as Code 1111, which has a preference ratio of 2.8 to 1, β over α. In the region from 30-70 in Repeat 1 and corresponding regions in the other three repeats, we predict that there may be an α-helix-turn-α-helix structure bounded by a β-strand or extended structure at the N-terminus and another β-strand C-terminal to the last α-helix. In the region from about 160-215 in Repeat 1 and the corresponding regions of the other three repeats there is the same periodicity of 4 prolines (**Pro**): **Pro**-hydrophobic cluster-**Pro**-hydrophobic cluster-**Pro**-hydrophobic cluster-hydrophilic cluster-hydrophobic cluster-**Pro**-β-strand. There is a glycine-rich region at about 250 in Repeat 1 and corresponding regions in the other repeats. It is followed by a β-strand at 255-260, a hydrophilic region from 260-265, a glycine-proline rich region from 275-280, hydrophilic regions from 280-290 and 295-305 leading into a possible amphipathic α-helix near the end of the repeat. A possible assignment of the secondary structure based on the conservation of hydrophobic clusters is compiled in **C**.

query in an ungapped BLAST search of the most current database (September 27, 1999), we found that the highest raw score of any protein in the nonredundant protein database (other than known IRBP orthologs) is 40 bits with an expected value of 0.02. This match was to the CTP of *Cyanobacterium synechocystis*. Other related CTPs possess raw scores ranging from 37 to 32 bits with E-values from 0.16 to 4.0. The sequence homology of the CTP family was extended to include the archaeon *Thermoplasma acidophilum* Tricorn Protease [4,5,22]. Keiler and Sauer [5] found two regions of similarity between Tsp and Domain B of IRBP with 53% identity; the rest of the two proteins have little similarity. Because 8.7 standard deviations and a best E-value of 0.02 may seem weak, a multisequence alignment technique was used to study the apparently distant sequence relationship of IRBP and the CTPs.

Hidden Markov models: The statistical method that we have employed to model the CTP-IRBP domain is a Hidden Markov Model (HMM). HMMs provided a means to model, align, and discriminate families of related sequences (Sequence Alignment and Modeling System). Figure 2 shows an HMM-generated multiple sequence alignment of 31 putative orthologs of CTP and IRBP that includes proteins from archaea, eukaryotes, cyanobacteria, bacteria and plants. The statistical analysis of the HMM suggests that the alignment is from a single family of proteins. For a significance level of 0.01 and a (nonredundant protein) database containing 186440 different sequences, a threshold value of -22.4 discriminates family members from non-family members. A more negative number for a given sequence denotes increasing confidence in the relatedness of the sequence to those used to create the HMM while a more positive number indicates greater similarity due to chance. The 31 IRBP-like sequences had scores ranging from -519 to -72 while all other sequences in the database had scores of -10 or greater (the vast majority of the 186440 database sequences had large positive scores). Thus, the sequence similarity between Domain B of the Repeat of eukaryotic IRBPs and a domain present in the CTPs is unlikely to be random and is significant [20]. The HMM analysis validates more clearly the family relationship of IRBP and the CTPs and extends the range of the family members to new biological kingdoms. The alignment also defines well conserved amino acids needed for mutation analysis described in the companion paper [12].

Hydrophobic cluster analysis: Hydrophobic cluster analysis [16,23] of the four human repeats was carried out at DrawHCA and the results are shown in Figure 3A. The explanation of these profiles is more fully considered in the Discussion, but briefly, one of the exceptional advantages of HCA is the ability to depict and recognize conserved patterns in related proteins that are not collinear with the primary sequence. Patterns separated by 3 or 4 amino acids (or multiples of 3 or 4) can be recognized. Several features in the plots are worth noting: (1) A distinct boundary between Domains A and B occurs at about position 80-90 in each Repeat. (2) Distinctive β-strands are predicted at positions marked with β, and distinct α-helices are marked with α on Figure 3C. (3) In Fig-

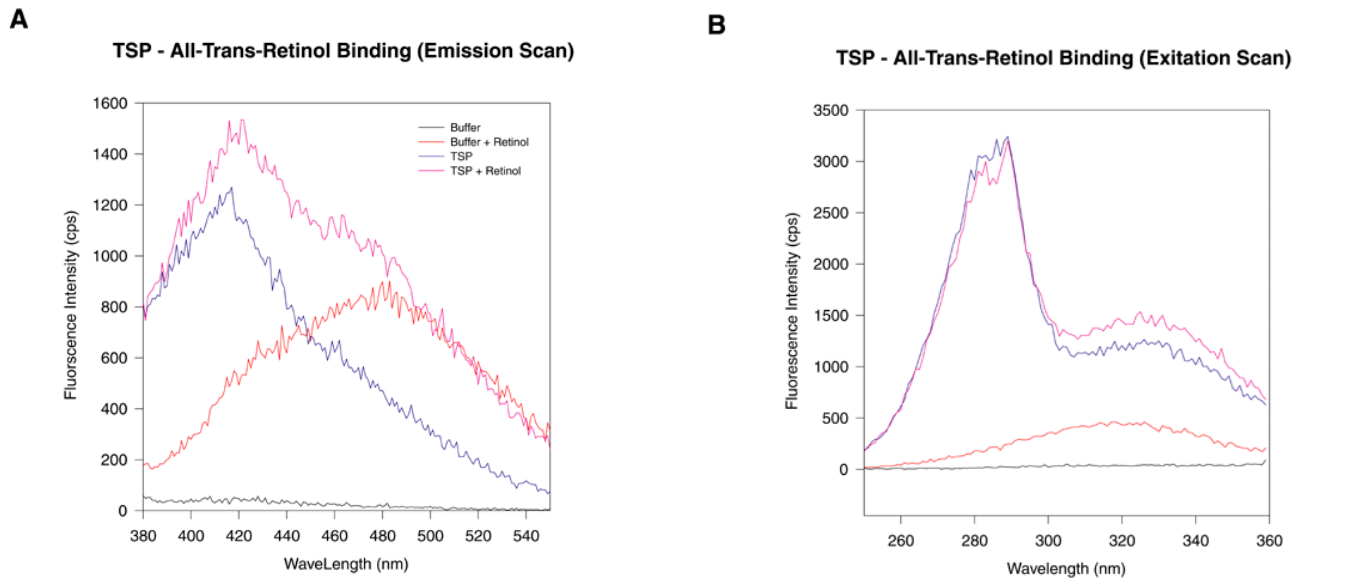

Figure 4. Does retinol bind to Tsp?. Tsp (1 μM) and retinol (10 μM) were mixed together and incubated to allow potential binding to occur. **A** shows the results of an emission wavelength scan from 380 to 550 nm while holding the excitation constant at 333 nm. The spectrum was compared with the same concentration of protein alone and 10 μM all-*trans*-retinol alone. The scans show that the mixture of ligand and protein was no different from the sum of the ligand alone and the protein alone, suggesting that retinol exhibited no fluorescence enhancement when mixed with the protein (indicating that Tsp does not bind retinol). **B** shows the results of an excitation scan while holding the emission constant at 479 nm. The results were the same as with the emission scan and showed no fluorescence enhancement of retinol in the presence of Tsp, again suggesting that retinol did not bind to Tsp.

ure 3B, well conserved non-linear patterns are depicted by overlaying parts of the four Repeats. The heavy-lined boxes represent the boundaries of conserved patches of hydrophobicity and hydrophilicity that occur in the two-dimensional representation of the sequences (not a collinear relationship), whereas less conserved areas have light lines. Many distinctive heavy-lined patterns are conserved among the four Repeats. (4) Other heavy features, including identical amino acids, indicate conserved motifs. (5) The bottom line (Figure 3C) shows a prediction of the secondary structure based on hydrophobic patterns recognized in the bidimensional arrays of the sequences. The patterns associated with α-helix and β-strands are as defined in reference [16].

PSA: Secondary structure and class predictions for Repeats 1-4 were obtained via the PSA program [18]. The composite prediction derived from α-helix, β-strand, and turn from PSA can be interpreted as eight alternating α-β motifs with occasional extra β-strands and occasional replacement of an α-helix with a β-strand (data not shown). The same conserved pattern of helices, turns, and strands occurred regardless of the specific Repeat with few exceptions. The pattern is similar to the pattern obtained from the HCA analysis, above (Figure 3C).

JPred: We employed JPred as another method to predict secondary structure. The query structure was a multiple sequence alignment of Repeat 1 from twelve species. There is a good agreement among this Repeat 1 Jpred consensus (data not shown) and the results from HCA and PSA secondary structure predictions.

To summarize Section 1, we found: (1) a structural similarity between eglin c and Domain A, (2) a conserved pattern of alternating α-helices and β-strands, and (3) a weak but statistically significant match among the Domain B structures of individual Repeats and the CTPs (including Tsp) extending across most of the biological kingdoms.

Section 2: Comparisons of functional characteristics of Tsp and EcR1: Given the amino acid sequence similarity between Tsp and IRBP, it was an obvious question to ask whether the similarity extended to the physiological level. Two experiments were carried out to test whether IRBP and Tsp share functional properties. We tested whether IRBP exhibited any protease activity (a known functional activity of Tsp) and whether Tsp possessed any retinol or fatty acid binding activity (a putative function of IRBP).

Does Tsp bind retinol or 16-AP? Figure 4 and Figure 5 show the results of fluorescence wavelength scans of Tsp with and without added retinol (Figure 4) or 16-AP (Figure 5). In both excitation and emission scans, a mixture of the protein and ligand was the sum of the scans of the ligand and protein separately scanned. This suggested that there was no fluorescence enhancement with either ligand and thus no evidence of Tsp binding either ligand. This contrasts with the results observed with IRBP and IRBP repeats where there is a ten-fold enhancement of fluorescence upon binding retinol and an 8-fold enhancement on binding 16-AP as described in the literature [2,13,14,24-28].

Does IRBP have protease activity? To determine whether IRBP possesses any protease activity, we incubated

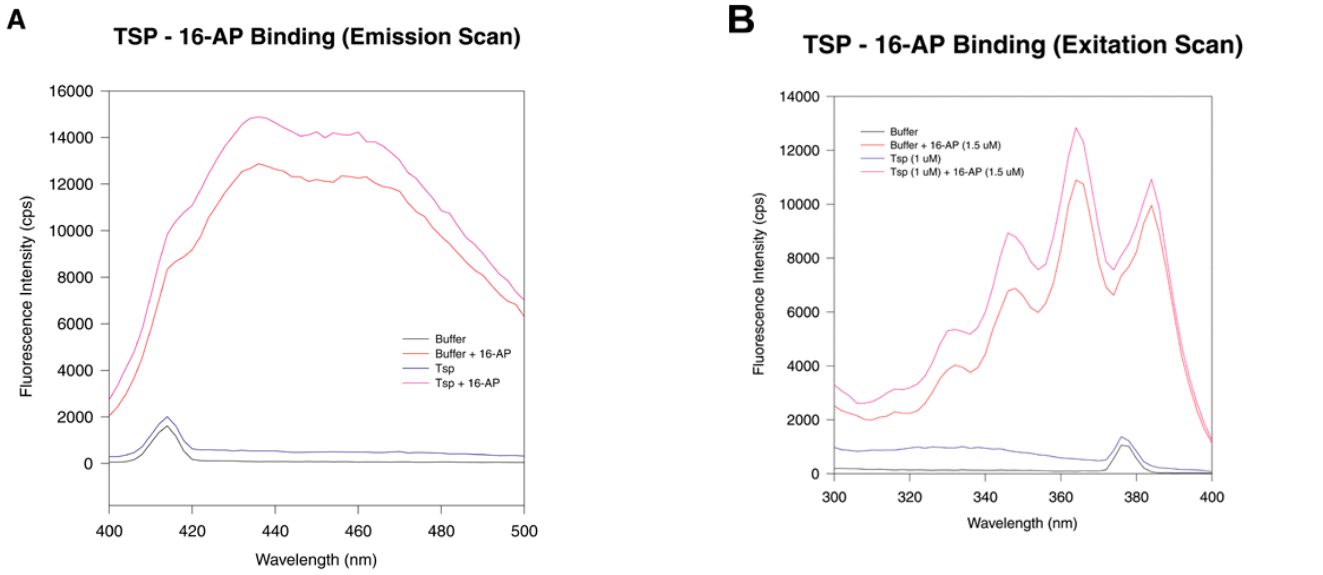

Figure 5. Does Tsp bind 16-AP?. Tsp (1 μM) and 16-AP (1.5 μM) were mixed together and incubated to allow possible binding to occur. In **A** we show the results of the emission wavelength scan from 400 to 500 nm while holding the excitation constant at 362 nm. The spectrum was compared with the protein alone (1 μM) and 16-AP alone (1.5 μM). The scans show that the mixture of ligand and protein was no different from the sum of the ligand alone and the protein alone, indicating that 16-AP exhibited no fluorescence enhancement when mixed at equilibrium with the protein (indicating that Tsp does not bind 16-AP). **B** shows the results of an excitation scan while holding the emission at 432 nm. The results were the same as with the excitation scan and showed no fluorescence enhancement of 16-AP in the presence of Tsp, also suggesting that 16-AP did not bind to Tsp.

recombinant human IRBPs with succinylated casein. When digested with a protease, the cleaved casein product will have a primary amino group that reacts with trinitorbenzenesulfonic acid, producing an orange color detected by absorbance at 450 nm. Casein was chosen because many different proteases exhibit enzymatic activity with this substrate [29]. After incubation overnight at 37 °C, we were not able to detect any protease activity in IRBP (Table 1), indicated by the Student's t-test (p = 0.458). Positive controls showed large A_{450} measurements even with traces of trypsin by comparison (Table 1). These results suggest that IRBP does not possess protease activity. These experiments do not rule out the possibility that IRBP has a specific protease activity that is not active on this casein-based substrate.

To summarize Section 2, despite the sequence similarity of Domain B of IRBP and the CTPs, we found no evidence that Tsp and IRBP share functions, suggesting that the sequence similarity is the result of maintenance of the same tertiary structure. These results suggest that mutation of common amino acids between Tsp and IRBP would alter conformation, but that amino acids shared only among the IRBP Repeats, and not with the CTPs, would be better candidates for retinol-binding functional mutants.

Table 1. Test of Protease Activity in IRBP

	A_{450} before incubation	A_{450} after incubation
IRBP (n = 14)	0.0802 ± 0.0314	0.0860 ± 0.0297
3.90 ng Trypsin	0.031	0.491
1.95 ng Trypsin	0.026	0.336
0.980 ng Trypsin	0.029	0.299
0.488 ng Trypsin	0.030	0.260
0.244 ng Trypsin	0.029	0.185
0.122 ng Trypsin	0.033	0.112
Buffer (no protein)	0.037	0.049

About 20 ng of IRBP (n = 14) was incubated with succinylated casein substrate for 25 h at 37 °C. Succinylation blocks primary amines in the protein. Should there be protease activity in a tested protein preparation, the casein would be digested, exposing unblocked primary amines, which react with trinitrobenzenesulfonic acid and produce a colored product. Absorbance at 450 nm was measured before and after the incubation. The value for IRBP represents the mean ± the standard deviation. In this experiment, the IRBP sample used was R12+ [13]. These results are typical of those obtained with other human IRBP recombinant proteins. By t-test, the difference between A_{450} before and after incubation (0.0058) of the IRBP samples is not significant (p = 0.458), indicating no protease activity in this IRBP preparation. Trypsin, even at the lowest level tested of 0.12 ng, gave a detectable increase in absorbance (0.079) over the starting value. Buffer with substrate gave a small increase in absorbance of 0.012 after incubation.

DISCUSSION

We undertook an analysis of the sequences of the four repeats of human IRBP to predict essential structures and functional domains within the Repeat 1 protein. We found sequence similarity between IRBP Domain A and eglin c, a protease inhibitor that functions by binding tightly to the active site of a protease. We extended the previous studies of sequence similarities between Domain B of IRBP and Tsp and related proteases, and we examined the possibility of shared functions between Tsp and IRBP. The results suggest no shared functions despite sequence similarities, implying that the conservation of amino acid sequence is to maintain the same structure, not function. As a consequence of the above studies, we could mutate amino acids that might be involved in the function of binding ligands. In a companion paper [12], we made single point substitutions considering these to be less intrusive in changing tertiary structure in contrast to truncations of several or many amino acids that might greatly affect conformation or tertiary structure.

Section 1: Computer analyses and the prediction of retinol contact points

Analysis of the predicted relationship of Domain A to eglin c: There is primary sequence similarity between eglin c and IRBP, which justifies three-dimensional model building. The reasons for reporting this information here are threefold: First, the eglin c/IRBP similarity is novel and interesting as no homologous structure (of IRBP Domain A) had been known before. Second, it allows us to build a three-dimensional model of part of IRBP, previously not possible with any part of IRBP. Third, eglin c binds proteases and this may suggest a similar tight Domain A-Domain B interaction exists in a single IRBP Repeat. Thus, Domain A might interact with Domain B and affect binding in a single Repeat although Domain A does not directly bind retinol [2].

The data shown in Figure 1 demonstrate a structural similarity between eglin c and Domain A of Repeat 3. Eglin c is a small protease inhibitor [30]. We can speculate on the meaning of the resemblance between eglin c and IRBP Domain A as follows: First, eglin c and IRBP are both secretory proteins, and perhaps the addition of an eglin c like domain by exon-shuffling to the N-terminal end of Domain B of IRBP led to the original formation of a secreted monomeric progenitor of IRBP. Second, the three-dimensional model of Domain A may hint at potential functional roles of IRBP: Domain A may be a protease inhibitor like eglin c. The tight interaction of proteins such as eglin c with the active site of a protease, such as leucocyte elastase, may imply an analogous structure or homologous regulatory/binding function in IRBP. Alternatively, Plantner and coworkers [31] found that IRBP and matrix metalloproteinases (MMPs) can co-purify when isolated from the IPM. This may suggest a function for Domain A in binding to MMPs, analogous to eglin c binding proteases. A single molecule of complete IRBP, with four Domain A's, might serve as a scaffold for four enzyme molecules. Also, we speculate that Domain A of each IRBP Repeat may function as an in-

hibitor of a separate protease in the IPM. Another possibility is that Domain A may bind to an active site (possibly a ligand binding site) in a putative IRBP receptor on the RPE apical membrane or photoreceptor plasma membrane. We further speculate that such interactions of Domain A with other proteins might be reversible (as it is with eglin c and elastase complex). As eglin c binds proteases, this may suggest a similar tight Domain A-Domain B interaction in a single IRBP Repeat. Thus, point mutations in Domain A might exert effects on Domain B and cause a loss of binding in a single Repeat. This new information may imply that Domain A is relevant to IRBP functions in binding retinol. This information also highlights the need to include Domain A and Domain B in an expressed protein to allow this putative interaction to occur. Consequently, in a companion paper [12] we included Domain A sequences in constructs designed to test whether single point mutations in Domain B can affect retinol binding function.

Analysis of the HMM predictions: An HMM of the IRBP-CTP domain was created (Figure 2) and used to assess whether the observed sequence similarity is statistically significant. Importantly, the statistics derived from the HMM showed convincingly ($p < 0.01$) that each member of the alignment is a member of a single protein family, and implies that there was just one ancestor from which the present day family members all descended. The alignment highlights amino acids that are strongly conserved. The PHD secondary structure prediction underneath this alignment indicates that IRBP consists of an $\alpha + \beta$ class of protein.

These statistics expand and extend the prior findings [3,5], and the results of a PILEUP analysis (GCG, [32]) reported by Baer, et al. [2] of the fourth repeat of IRBP from assorted species and CTPs. Jointly, these findings suggest an ancient family originating before the divergence of archea and bacteria, perhaps 3.5 billion years ago.

The HMM analysis excluded other retinoid- and fatty acid-binding proteins as members of this family. For example, no match was found to the serum retinol binding protein or cellular retinol binding protein. The isolation of the IRBP-CTP family from other retinoid binding protein families may imply a different function for IRBP, not related to the transport or substrate presentation functions of these other binding protein families.

As shown in Figure 2, within the 300 amino acid IRBP repeat unit, there appear to be identifiable structural or folding domains. Domain B in IRBP begins near the boundary of Exon 1 and 2 in Repeat 4. Tsp lacks the equivalent eglin c-like structure of Domain A. However, Tsp (and the other carboxy-terminal proteases) contain a PDZ domain (unpublished observation, I. S. Mian) in place of the eglin c-like Domain A. Carbohydrate attachment sites are found in the repeat mainly at exon-exon boundaries. These suggest possible boundaries between different structural domains within the repeat.

A structure of the Repeat has emerged as one Domain A followed by one Domain B. This was supported by the apparent separation of these two domains by an obvious high-proline content and low complexity sequence between the two domains. This motif could be a hinge [33] or a tether [34]. The region is very readily detected by the hydrophobic cluster analysis, discussed next, and is obvious in all four repeats.

Analysis of the HCA predictions: Hydrophobic cluster analysis was designed to display primary amino acid sequences in two-dimensions so that amino acids that may be close together in two- or three-dimensions may be clustered in a two-dimensional graphical depiction [16,23]. The clusters of amino acids suggest well conserved secondary structure, and display the boundaries between adjacent globular domains. The separation of Domain A from Domain B was clearly made by a proline rich region at about position 85 (Figure 3). The prediction of secondary structure and the identification of well conserved nonlinear hydrophobic patches suggest conserved conformation and patterns of structures among the four human IRBP Repeats. While it is not possible solely with this predictive device to determine which patterns are involved with structure and which others are critical for ligand binding, the HCA analysis allowed us to select candidates for these roles.

Analysis of the PSA predictions: A different approach to identifying secondary structure within the IRBP Repeat was to use the PSA program. PSA computes the probability of an amino acid being in a particular secondary structure by making use of predefined structural class models, which are used to make Discrete State Space Models. The predictions derived from the PSA utility show extensive similarity among the predictions for the four human IRBP Repeats. The predictions generate 10 α-helices and 11 β-strands with intermediate turns. While the predictions have lost some of their phasing because of small additions or gaps in loops, the peak heights and valleys show close agreement, as do the lengths of the predicted structures. The predicted secondary structures thus show widespread agreement in the conformation of the four repeats of IRBP and show extensive similarity with the secondary structure of Tsp and the other CTPs.

Analysis of the JPred predictions: JPred develops a consensus among several programs used to predict secondary structures based on different algorithms and principles. We employed JPred to obtain a multiple sequence alignment of Repeat 1 from 12 different species, including human. When viewed alongside the HCA and PSA predictions, there is a good correspondence in the predicted secondary structure. Overall, the secondary structure predictions allow us to predict (with about 75% accuracy) a putative secondary structure for the IRBP Repeat. We used the consensus secondary structure to predict targets for mutation that might be involved in ligand binding, hypothesizing that a binding site might be a group of amino acids near clustered C-terminal ends of β-strands. This location for an active site is typical of TIM barrel proteins and Rossman fold proteins [35].

Secondary structure predicted by these very different approaches (PHD, HCA, PSA, and JPred) show similar charac-

teristics. These agreed with estimates of α-helix and β-strand content from circular dichroism of each repeat in human IRBP and the prediction of an α+ β class protein for IRBP [13]. We suggest that Domain A consists of two α-helices with one β-strand and two turns. Domain B seems to consist of 6 or 7 α-helices and 8 or 9 β-strands generally alternating (See also the PHD structure predication in Figure 2). Several of the amino acids that we subsequently chose to mutate in Repeat 1 are located near the C-terminal ends of β-strands.

Section 2: Analysis of potential shared functional characteristics of Tsp and EcR1

Members of a family of proteins sharing a common ancestor often exhibit shared functions. Because of the sequence similarity between IRBP and the CTPs, we asked whether *E. coli* Tsp (a CTP) exhibits the ability to bind retinol or 16-AP. Also, we tested whether IRBP exhibited any protease activity, as many members of the CTP-IRBP domain family apparently are C-terminal proteases. The results show that Tsp does not exhibit fluorescence enhancement when at equilibrium with retinol or 16-AP, and this suggests that it does not bind retinoids or fatty acid analogs (Figure 4 and Figure 5). Similarly, we found that IRBP does not appear to possess protease activity. These results imply that the relationship of IRBP and Tsp lies in the structure of the proteins, not in their functional capabilities. Thus, even though IRBP and Tsp share a tertiary structure or fold (which is versatile and useful), these proteins long ago diverged in function. As described below, we relied on and employed this information to predict specific amino acid changes in EcR1 that might affect function without changing the structure of the protein.

Summary and Conclusions: The lack of shared functions between Tsp and IRBP suggests that the conservation of amino acid sequences between these two proteins is important to the maintenance of an evolutionarily successful tertiary structure that can be adapted for very different biological functions. The apparent fusion of an eglin c-like domain to Domain B of an IRBP/Tsp ancestor may have led to the transformation in function from a protease to a retinoid-binding protein, and this fusion might be an example of exon-shuffling [36,37]. The Domain A-eglin c similarity, coupled with the already known tight binding of eglin c to proteases, leads us to propose a close interaction between Domain A and Domain B in a single IRBP repeat. The above results placed constraints on the choices of amino acids to mutate that would affect retinol-binding functions without interfering with the overall structure of Repeat 1: Mutation of amino acids shared between IRBP and Tsp might cause changes in structure. Also, amino acids conserved among IRBP repeats might be structurally conserved contact points between Domains A and B. However, conservation of β-strand positions can be used to predict the locations of active site residues and consequently to select candidate amino acids of the retinol binding site.

In a companion paper [12], we consider whether the binding site for retinoids and fatty acids is (A) large or small (B) whether the site is a surface hydrophobic patch (C) the dependence of ligand binding on a heat-sensitive conformation. The data in the present paper allow us to rationally choose point mutations permitting us to resolve these questions.

ACKNOWLEDGEMENTS

Parts of this work were presented at the 1996, 1997, 1998, and 1999 annual meetings of the Association for Research in Vision and Ophthalmology. We thank Dr. Robert Sauer for supplying the *E. coli* Tsp clone. These studies were funded by NIH R01 EY10553 (to JMN), P30 EY06360, and T32 EY 07092; a Center grant from the Foundation Fighting Blindness; and an unrestricted grant to the Emory Eye Center from Research to Prevent Blindness; and the Director, Office of Science, Office of Basic Energy Sciences, U.S. Department of Energy, under Contract number DE-AC03-76SF0098 (to ISM).

REFERENCES

1. Borst DE, Redmond TM, Elser JE, Gonda MA, Wiggert B, Chader GJ, Nickerson JM. Interphotoreceptor retinoid-binding protein. Gene characterization, protein repeat structure, and its evolution. J Biol Chem 1989; 264:1115-23.
2. Baer CA, Retief JD, Van Niel E, Braiman MS, Gonzalez-Fernandez F. Soluble expression in E. coli of a functional interphotoreceptor retinoid-binding protein module fused to thioredoxin: correlation of vitamin A binding regions with conserved domains of C-terminal processing proteases. Exp Eye Res 1998; 66:249-62.
3. Silber KR, Keiler KC, Sauer RT. Tsp: A tail-specific protease that selectively degrades proteins with nonpolar C termini. Proc Natl Acad Sci U S A 1992; 89:295-9.
4. Tamura T, Tamura N, Cejka Z, Hegerl R, Lottspeich F, Baumeister W. Tricorn protease-the core of a modular proteolytic system. Science 1996; 274:1385-9.
5. Keiler KC, Sauer RT. Identification of active site residues of the Tsp protease. J Biol Chem 1995; 270:28864-8.
6. Bazan NG, Reddy TS, Redmond TM, Wiggert B, Chader GJ. Endogenous fatty acids are covalently and noncovalently bound to interphotoreceptor retinoid-binding protein in the monkey retina. J Biol Chem 1985; 260:13677-80.
7. Chen Y, Houghton LA, Brenna JT, Noy N. Docosahexaenoic acid modulates the interactions of the interphotoreceptor retinoid-binding protein with 11-cis-retinal. J Biol Chem 1996; 271:20507-15.
8. Saari JC, Teller DC, Crabb JW, Bredberg L. Properties of an interphotoreceptor retinoid-binding protein from bovine retina. J Biol Chem 1985; 260:195-201.
9. Adler AJ, Martin KJ. Retinol-binding proteins in bovine interphotoreceptor matrix. Biochem Biophys Res Commun 1982; 108:1601-8.
10. Fong SL, Liou GI, Landers RA, Alvarez RA, Bridges CD. Purification and characterization of a retinol-binding glycoprotein synthesized and secreted by bovine neural retina. J Biol Chem 1984; 259:6534-42.
11. Putilina T, Sittenfeld D, Chader GJ, Wiggert B. Study of a fatty acid binding site of interphotoreceptor retinoid-binding protein using fluorescent fatty acids. Biochemistry 1993; 32:3797-803.
12. Gross EA, Li GR, Ruuska SE, Boatright JH, Mian IS, Nickerson JM. Effects of dispersed point substitutions in Repeat 1 of human interphotoreceptor retinoid-binding protein (IRBP). Mol Vis 2000; 6:40-50 <http://www.molvis.org/molvis/v6/a7>.

13. Lin ZY, Li GR, Takizawa N, Si JS, Gross EA, Richardson K, Nickerson JM. Structure-function relationships in interphotoreceptor retinoid-binding protein (IRBP). Mol Vis 1997; 3:17 <http://www.molvis.org/molvis/v3/p17>.
14. Nickerson JM, Li GR, Lin ZY, Takizawa N, Si JS, Gross EA. Structure-function relationships in the four repeats of human interphotoreceptor retinoid-binding protein (IRBP). Mol Vis 1998; 4:33 <http://www.molvis.org/molvis/v4/p33>.
15. Rost B. Better 1D predictions by experts with machines. Proteins 1997; Suppl 1:192-7.
16. Lemesle-Varloot L, Henrissat B, Gaboriaud C, Bissery V, Morgat A, Mornon JP. Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences. Biochimie 1990; 72:555-74.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10.
18. Stultz CM, White JV, Smith TF. Structural analysis based on state-space modeling. Protein Sci 1993; 2:305-14.
19. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998; 14:892-3.
20. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994; 235:1501-31.
21. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997; 18:2714-23.
22. Shestakov SV, Anbudurai PR, Stanbekova GE, Gadzhiev A, Lind LK, Pakrasi HB. Molecular cloning and characterization of the ctpA gene encoding a carboxyl-terminal processing protease. Analysis of a spontaneous photosystem II-deficient mutant strain of the cyanobacterium Synechocystis sp. PCC 6803. J Biol Chem 1994; 269:19354-9.
23. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci 1997; 53:621-45.
24. Adler AJ, Evans CD, Stafford WF 3rd. Molecular properties of bovine interphotoreceptor retinol-binding protein. J Biol Chem 1985; 260:4850-5.
25. Okajima TI, Pepperberg DR, Ripps H, Wiggert B, Chader GJ. Interphotoreceptor retinoid-binding protein: role in the delivery of retinol to the pigment epithelium. Exp Eye Res 1989; 49:629-44.
26. Chen Y, Saari JC, Noy N. Interactions of all-trans-retinol and long-chain fatty acids with interphotoreceptor retinoid-binding protein. Biochemistry 1993; 32:11311-8.
27. Baer CA, Kittredge KL, Klinger AL, Briercheck DM, Braiman MS, Gonzalez-Fernandez F. Expression and characterization of the fourth repeat of Xenopus interphotoreceptor retinoid-binding protein in E. coli. Curr Eye Res 1994; 13:391-400.
28. Lin ZY, Si JS, Nickerson JM. Biochemical and biophysical properties of recombinant human interphotoreceptor retinoid binding protein. Invest Ophthalmol Vis Sci 1994; 35:3599-612.
29. Hatakeyama T, Kohzaki H, Yamasaki N. A microassay for proteases using succinylcasein as a substrate [published erratum appears in Anal Biochem 1992; 207:359]. Anal Biochem 1992; 204:181-4.
30. Bode W, Huber R. Natural protein proteinase inhibitors and their interaction with proteinases. Eur J Biochem 1992; 204:433-51.
31. Plantner JJ, Quinn TA. Association of matrix metalloproteinases with interphotoreceptor retinoid binding protein. Curr Eye Res 1997; 16:51-5.
32. Devereux J, Haeberli P, Smithies O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 1984; 12:387-95.
33. Beale D, Feinstein A. Structure and function of the constant regions of immunoglobulins. Q Rev Biophys 1976; 9:135-80.
34. Wootton JC, Drummond MH. The Q-linker: a class of interdomain sequences found in bacterial multidomain regulatory proteins. Protein Eng 1989; 2:535-43.
35. Branden C, Tooze J. Introduction to protein structure. New York: Garland Publishing; 1991.
36. Patthy L. Genome evolution and the evolution of exon-shuffling—a review. Gene 1999; 238:103-14.
37. Doolittle RF. The multiplicity of domains in proteins. Annu Rev Biochem 1995; 64:287-314.