1085-9195/01/35/35–47/$13.25

ORIGINAL ARTICLE

Predicted Structures of Two Proteins Involved in Human Diseases

Huan-Xiang Zhou and Guoli Wang*

Department of Physics, Drexel University, Philadelphia, PA 19104

Abstract

Structures of 79 proteins involved in human diseases were predicted by sequence alignments with structural templates. The predicted structures for ALDP and CSA, proteins responsible for adrenoleukodystrophy and the Cockayne syndrome, respectively, were analyzed to elucidate the molecular basis of disease mutations. In particular we positioned residue P484 of ALDP in the homodimer interface. This positioning is consistent with a recent experimental finding that the mutation P484R significantly decreases the self-interaction of ALDP and suggests that the disease mechanism of this mutation lies in the impaired ALDP dimerization. We identified two new WD repeats in CSA and suggest that one of these forms part of the interaction surface with other proteins.

Index Entries: Adrenoleukodystrophy; the Cockayne syndrome; disease mutation; sequence alignment; homology modeling.

INTRODUCTION

With the sequencing of the human genome near completion, the next challenges are to build structural models for the estimated 30,000–35,000 proteins *(1,2)* and to elucidate the molecular basis of their functioning. It is clear that automated structure prediction methods will play an essential role. We have carried out a systematic structural characterization of proteins involved in human diseases and predicted structures for 79 such proteins. The predicted structures of two disease proteins, ALDP and CSA, were analyzed in detail in order to elucidate the molecular basis of disease mutations.

ALDP, located in the peroxisomal membrane, is an ATP-binding cassette (ABC) half-transporter with a hydrophobic domain and a nucleotide-binding domain *(3)*. In the present study we focus on the nucleotide-binding domain. A number of mutations (P484R, S515F, R518W, S606L, and R617C) within this domain are known to contribute to adrenoleukodystrophy (OMIM 300100). This neurodegenerative disorder is characterized by the abnormal metabolism of very long-chain fatty acids *(4)*.

CSA is a protein with 396 amino acids that has been proposed to contain five WD repeats *(5)*. It serves to rectify UV sensitivity and defective chloramphenicol acetyl-transferase gene reactivation. Deletions of C-terminal portions

* Author to whom all correspondence and reprint requests should be addressed. E-mail: hxzhou@einstein.drexel.edu

(V282-E374, Y322-C terminal, and E348-E374) of CSA are known to be causes of the Cockayne syndrome (OMIM 216400).

MATERIALS AND METHODS

The structure prediction was carried out by a recently developed method called COBLATH *(6).* Given a set of protein sequences (termed queries), structure prediction by COBLATH consists of two steps. First, each sequence from a complete set of representatives from the Protein Data Bank (PDB) is matched against the set of queries by PSI-BLAST *(7)* to search for hits. For a particular query, if a hit with a PDB sequence (now called the template) is indeed scored, the final query-template alignment is obtained by a threading procedure.

The disease proteins studied in this paper came from two lists (http://genome.nhgri.nih.gov/clone/ and http://ncbi.nih.gov/disease/). For the complete set of representative sequences from the PDB, we chose the FSSP library *(8),* in which PDB sequences with identities below 25% are treated as unique entries. The edition (August 5, 2000) of the FSSP library used in the present study has 2261 entries.

Assignment of Structural Templates

Two sessions of PSI-BLAST searches were carried out to search for matches between the 2261 representative PDB sequences and the disease-protein sequences. Each of the sequences in the FSSP library was first used to search for hits in a database of 348,901 sequences from Swissprot. Its structural alignments in the FSSP library were used as seeding. All the resulting alignments collected in the first PSI-BLAST session were then used as seeding for a search within the set of disease proteins. Swissprot and disease-protein sequences were filtered for low-complexity regions. Both e and h were set to 0.5×10^{-3}. Seventy-nine disease proteins were matched with the FSSP entries.

Validation of Template-Assignment Protocol

A separate experiment was designed to assess the reliability of the template-assignment protocol. The strategy was to use an earlier edition of the FSSP library for template assignments and then check these assignments against newly deposited PDB entries. Specifically, template assignments were made for the open reading frames (ORFs) of 32 completely sequenced genomes with the November 30, 1999 edition of the FSSP library (1907 entries) by the same protocol as described previously. Two of the genomes included were *Mycoplasma genitalium* (MG) and *Saccharomyces cerevisiae* (SC). Out of a total of 91,287 ORFs, 32,644 (36%) were assigned templates.

Since November 30, 1999, 3884 new protein chains have been deposited in the PDB. We used each of these new PDB sequence to search for hits among the 32 genomes by PSI-BLAST with an extremely strict criterion: with e set to 10^{-30} and h set to 10^{-35}. This way we can be assured that whenever there is a match between a PDB protein and an ORF, they will have similar structures. A total of 5354 hits were scored. Among them 4871 were with ORFs that were assigned templates from the 1907 entries of the FSSP library. When the newly assigned PDB templates were compared to the older FSSP templates, all but one were found to be structural neighbors in the later (August 5, 2000 edition) FSSP library. The new PDBs thus resoundingly confirmed the reliability of the FSSP templates.

Additional assessment of reliability was made by comparing with template assignments by other methods and in other studies for the MG and SC genomes. One hundred sixty-two of the 479 MG ORFs annotated by PSI-BlAST were also assigned templates by threading with a scoring function composed of sequence profile and predicted secondary structure and solvent exposure *(6).* Of these all but one had consistent template assignments. Jones assigned templates for 2107 of the 6337 SC

ORFs (web page: http://globin.bio.warwick.ac.uk/genome) via a different threading procedure *(9)*. Of these 1802 were assigned templates from the 1907 FSSP entries; and the template assignments were consistent in all but 10 cases. In general, PSI-BLAST has higher reliability.

Refinement of Query-Template Alignment

Once a match between a query and a template was obtained by PSI-BLAST, the accuracy of the sequence alignment was further improved by a threading procedure *(6)*. The scoring function is the sum of the sequence profile from PSI-BLAST and a substitution matrix relating the predicted and actual secondary structure and solvent exposure. The threading was implemented by dynamic programming, without penalty for unmatched N and C terminals of either the query or the template (i.e., the local-local algorithm). Penalty scores for alignment gaps were 5 for opening and 0.3 per residue for extension.

RESULTS AND DISCUSSION

Structural Annotations of 79 Disease Proteins

We were able to obtain structural templates for 79 disease proteins. In 19 cases the template assignment is trivial since a structure for either a region or the whole sequence has already been deposited in the PDB and the best match was with that particular PDB entry. In the other 60 cases, at least one region of the sequence was matched to a homologue. In particular, the N-terminal nuclease domain of WRN (responsible for Werner syndrome) first predicted by Mushegian et al. *(10)* is confirmed. The annotations of all the 60 disease proteins are listed in Table 1.

In three disease proteins new structural domains were detected (*see* Table 1). Pendrin, the protein responsible for Pendred syndrome (and the homologous diastrophic dysplasia protein) has been known as a sulfate transporter *(11)*. According to our structural annotation, 96 residues near the C terminal of pendrin form a structure similar to the transcription regulator SpoIIAA *(12)*. We predict that pyrin, responsible for familial Mediterranean fever, has an SH2 domain in the N terminal half. The PKD1 protein, responsible for type 1 polycystic kidney disease, is predicted to have a lipase domain near the C terminal.

Predicted Structure of ALDP

The structure model for the nucleotide-binding domain of ALDP was built on the PDB template 1b0uA *(13)* and is shown in Fig. 1. Though the overall sequence identity between ALDP and 1b0uA is only 23%, residues important for the functioning of the proteins are highly conserved. For example, the P-loop sequence in 1b0uA, GSSGSGKS, is aligned to GPNGCGKS in ALDP. Other residues of 1b0uA that interact with ATP (Q100, D178, E179, and H211) are strictly conserved in ALDP. The motif C sequence of 1b0uA, LSGGQQQRVSIAR, is aligned to LSGGEKRIGMAR in ALDP.

A number of mutations (P484R, S515F, R518W, S606L, and R617C) on ALDP are known to contribute to adrenoleukodystrophy (OMIM 300100). These can now be rationalized by the structure model. Residue P484 is located in a tight turn connecting two β strands. By analogy to 1b0uA, these strands are part of the interface of ALDP homodimerization, which is an essential step for the functioning of ALDP as an ABC transporter. Mutation P484R will lead to rearrangement of the β strands, which in turn will impair ALDP homodimerization. Indeed experiments have shown that this mutation significantly decreases the self-interaction of ALDP *(14)*. S515 and R518 are near the ATP binding site and may interact with ATP transiently. Their mutations to nonpolar residues may interfere with ATP binding. S606 and R617 are the second and last residues in motif C, which interacts with the EAA-like motif in the N-terminal half *(15)*. Mutations S606L and R617C will affect this interaction.

Table 1
Structural Annotation of 60 Disease Proteins

Qs	Qe	Template	Ts	Te	Tl	Match	identity	Fold[a]
1. Aarskog-Scott Syndrome								
>gi | 595425 | gb | AAA57004.1 | (U11690) FGD1 (961 aa)								
372-	691	1dbh_	2-	338	340	301	18	PH domain-like
663-	789	1dvpA	103-	214	217	109	33	PIP binding domain
830-	919	1mai_	16-	117	119	90	8	PH domain-like
2. Achondroplasia								
>gi | 182569 | gb | AAA52450.1 | (M58051) fibroblast growth factor receptor (806 aa)								
74-	246	1cvsC	42-	211	211	152	23	Immunoglobulin-like
458-	756	1fgka	1-	278	278	278	86	Protein kinase-like
3. Adenomatous Polyposis Coli								
>gi | 190165 | gb | AAA60353.1 | (M73548) polyposis locus-encoded protein (2844 aa)								
325-	726	3bct_	12-	454	457	379	23	alpha-alpha superhelix
4. Adrenoleukodystrophy, X-linked								
>gi | 38591 | emb | CAA79922.1 | (Z21876) ALD protein (745 aa)								
474-	715	1b0uA	7-	254	261	204	23	P-loop NTPase
5. Aniridia								
>gi | 189353 | gb | AAA59962.1 | (M77844) oculorhombin (422 aa)								
4-	136	6paxA	1-	133	133	133	98	DNA-binding 3-helical
213-	267	1au7A	74-	128	130	55	31	DNA-binding 3-helical
6. Ataxia Telangiectasia								
>gi | 1497931 | gb | AAB38309.1 | (U55757) ataxia-telangiectasia (3056 aa)								
1067-	1742	1qbkB	3-	686	879	634	12	alpha-alpha superhelix
2681-	3014	1qmmA	555-	837	841	272	22	Protein kinase-like
7. Autoimmune Polyglandular Syndrome								
>gi | 2696619 | dbj | BAA23990.1 | (AB006684) AIRE-1 (545 aa)								
294-	344	1zbdB	53-	108	134	51	22	PIP binding domain
8. Bloom Syndrome								
>gi | 1072122 | gb | AAA87850.1 | (U39817) Bloom's syndrome protein (1417 aa)								
672-	1027	1d2mA	176-	548	552	337	15	P-loop NTPase
9. Breast Cancer, Type 1								
>gi | 555932 | gb | AAA73985.1 | (U14680) breast/ovarian cancer (1863 aa)								
22-	121	1rmd_	24-	115	116	92	28	Classic zinc finger
10. Choroideremia								
>gi | 2950156 | emb | CAA55011.1 | (X78121) Rab geranylgeranyltransferase (653 aa)								
174-	545	1gnd_	25-	390	430	351	25	FAD/NAD(P)-binding
11. Chondrodysplasia Punctata								
>gi | 791004 | emb | CAA58556.1 | (X83573) ARSE (589 aa)								
37-	585	1auk_	2-	480	481	457	34	Phosphatase/sulphatase

(continues)

Table 1
(Continued)

Qs	Qe	Template	Ts	Te	Tl	Match	identity	Fold[a]
12. Chronic Myeloid Leukemia								
>gi | 487345 | gb | AAB60388.1 | (U07000) breakpoint cluster region (1271 aa)								
498-	863	1dbh_	3-	333	340	293	19	PH domain-like
911-	1006	1rsy_	26-	118	135	84	26	C2 domain-like
1049-	1231	1tx4A	4-	185	196	158	24	GTPase activation
>gi | 514268 | gb | AAB60393.1 | (U07563) tyrosine-protein kinase (1149 aa)								
85-	516	1fmk_	5-	425	438	401	45	SH3-like barrel
13. Cockayne Syndrome								
>gi | 975302 | gb | AAA82605.1 | (U28413) CSA protein (396 aa)								
32-	362	1gotB	44-	340	295	17		7-bladed β-propeller
>gi | 182181 | gb | AAA52397.1 | (L04791)excision repair protein (1493 aa)								
506-	989	1d2mA	8-	545	552	466	12	p-loop NTPase
14. Congenital Adrenal Hyperplasia								
>gi | 30322 | emb | CAA41709.1 | (X58906) steroid 21-monooxygenas (495 aa)								
25-	484	1bu7A	2-	451	455	425	19	Cytochrome P450
15. Congenital Nephrotic Syndrome								
>gi | 3025699 | gb | AAC39687.1 | (AF035835) nephrin (1241 aa)								
34-	113	1cs6A	297-	365	382	68	25	Immunoglobulin-like
146-	542	1cs6A	15-	381	382	346	16	Immunoglobulin-like
579-	935	1cs6A	40-	378	382	314	20	Immunoglobulin-like
845-	1036	1fnhA	2-	184	269	183	16	Immunoglobulin-like
16. Cystic Fibrosis								
>gi | 180332 | gb | AAA35680.1 | (M28668) membrane conductance regulator (1480 aa)								
437-633	1b0uA	14-	236	257	195	26	p-loop NTPase	
1208-	1440	1b0uA	1-	246	257	231	20	p-loop NTPase
17. Diastrophic Dysplasia								
>gi | 549988 | gb | AAA70081.1 | (U14528) sulfate transporter (739 aa)								
627-	722	1auz_	25-	116	116	91	16	SpoIIaa-like
18. DiGeorge Syndrome								
>gi | 1136388 | dbj | BAA11480.1 | (D79985) putative hydrophobic domain (550 aa)								
30-	69	1cr8A	4-	42	42	39	49	LBD of LDLP receptor
110-	267	1htn_	19-	153	156	135	17	C-type lectin-like
19. Duchenne Muscular Dystrophy								
>gi | 181857 | gb | AAA53189.1 | (M18533) dystrophin (3685 aa)								
15-	240	1qagA	1-	225	225	205	72	Calponin-homology
20. Epidermolytic Palmoplantar Keratoderma								
>gi | 435476 | emb | CAA82315.1 | (Z29074) cytokeratin 9 (623 aa)								
152-	467	1c1gD	17-	283	284	260	17	Parallel coiled-coil
21. Familial Mediterranean Fever								
>gi | 4588600 | gb | AAD26152.1 | (AF111163) pyrin (781 aa)								
188-	286	2cblA	220-	305	305	85	18	SH2-like

(continues)

Table 1
(Continued)

Qs	Qe	Template	Ts	Te	Tl	Match	identity	Fold[a]
22. Glycerol Kinase Deficiency								
>gi | 5834428 | emb | CAB54859.1 | (AJ252550) glycerol kinase (559 aa)								
15-	521	1glcG	5-	488	489	469	48	Ribonuclease H-like
23. Hereditary Non-Polyposis Colon Cancer								
>gi | 466462 | gb | AAA17374.1 | (U07418) homolog of E. coli mutL (756 aa)								
8-	334	1b63A	5-	332	333	327	35	Ribosomal protein S5
24. Holt-Oram Syndrome								
>gi | 1772561 | emb | CAA70592.1 | (Y09445) transcription factor (513 aa)								
52-	238	1xbrA	2-	182	184	180	52	diphtheria toxin
25. Huntington Disease								
>gi | 454415 | gb | AAB38240.1 | (L12392) Huntington's Disease protein (3144 aa)								
131-	273	1qgrA	330-	471	871	142	22	alpha-alpha superhelix
601-	1155	1qgrA	45-	562	871	500	8	alpha-alpha superhelix
26. Juvenile Glaucoma								
>gi | 2104789 | gb | AAC51725.1 | (AF001620) trabecular meshwork-induced (504 aa)								
63-	181	1c1gC	144-	250	284	107	21	Parallel coiled-coil
27. Kallman Syndrome								
>gi | 307080 | gb | AAA59202.1 | (M97252) KAL (680 aa)								
183-	381	1qg3A	1-	180	195	173	17	Immunoglobulin-like
443-	654	1qg3A	18-	194	195	174	20	Immunoglobulin-like
28. Limb-Girdle Muscular Dystrophy Type 2B								
>gi | 3600028 | gb | AAC63519.1 | (AF075575) dysferlin (2080 aa)								
2-	91	1djxB	436-	531	561	87	22	C2 domain-like
221-	319	1djxB	435-	542	561	95	25	C2 domain-like
383-	485	1djxB	438-	531	561	88	23	C2 domain-like
1107-	1276	1djxB	395-	559	561	154	19	C2 domain-like
1557-	1667	1djxB	415-	529	561	109	21	C2 domain-like
1813-	1944	1rsy_	28-	132	135	105	28	C2 domain-like
29. Long-QT Syndrome								
>gi | 2465531 | gb | AAC51776.1 | (AF000571) voltage-dependent K^+ channel (676 aa)								
257-	356	1bl8A	2-	97	97	96	33	Membrane all-alpha
>gi | 4156239 | dbj | BAA37096.1 | (AB009071) long QT syndrome (LQT2) (1159 aa)								
26-	135	1byw_	1-	110	110	97	100	Profilin-like
551-	667	1bl8A	7-	96	97	89	15	Membrane all-alpha
762-	847	2cgpC	15-	104	200	85	27	Double-stranded β-helix
30. Lowe Oculocerebrorenal Syndrome								
>gi | 189356 | gb | AAA59964.1 | (M88162) inositol polyphosphate-5-phophatase (968 aa)								
319-	606	1ako_	3-	266	268	252	16	DNase I-like
782-	959	1tx4A	3-	189	196	178	22	GTPase activation

(continues)

Table 1
(Continued)

Qs	Qe	Template	Ts	Te	Tl	Match	identity	Fold[a]
31. Lung Cancer								
>gi | 3603418 | gb | AAC63525.1 | (AF083439) phosphatase regulatory subunit (601 aa)								
21-	601	1b3uA	8-	588	588	581	86	alpha-alpha superhelix
32. Marfan Syndrome								
>gi | 1335064 | emb | CAA45118.1 | (X63556) fibrillin (2871 aa)								
529-	602	1emn_	3-	73	82	71	49	Knottins
1028-	1107	1emn_	4-	79	82	76	51	Knottins
2054-	2125	1apj_	3-	74	74	72	100	TB module/8-cys domain
2124-	2205	1emn_	1-	82	82	82	99	Knottins
33. Menkes Disease								
>gi | 854136 | emb | CAA57777.1 | (X82335) Menkes disease (1500 aa)								
8-	75	1aw0_	3-	70	72	68	53	Ferredoxin-like
165-	236	1aw0_	1-	68	72	68	46	Ferredoxin-like
278-	336	1aw0_	4-	62	72	59	46	Ferredoxin-like
375-	446	1aw0_	1-	72	72	72	100	Ferredoxin-like
489-	557	1aw0_	4-	72	72	69	35	Ferredoxin-like
567-	631	1aw0_	6-	70	72	65	37	Ferredoxin-like
763-	1059	1eulA	70-	366	994	277	25	
1212-	1386	1eulA	583-	789	994	175	31	
34. Miller-Dieker Lissencephaly								
>gi | 349826 | gb | AAA02881.1 | (L13386) Miller-Dieker lissencephaly (410 aa)								
122-	410	1bpoA	40-	330	487	271	12	7-bladed β-propeller
35. Myotonic Dystrophy								
>gi | 976144 | gb | AAA75236.1 | (L00727) myotonin-protein kinase, Form I (639 aa)								
56-	382	1cdkA	11-	325	343	305	35	Protein kinase-like
36. Myotubular Myopathy 1, X-linked								
>gi | 2224683 | dbj | BAA20826.1 | (AB002369) KIAA0371 (1198 aa)								
1115-	1178	1dvpA	151-	214	217	64	36	PIP binding domain
37. Neurofibromatosis, Type 1								
>gi | 292354 | gb | AAA59925.1 | (M89914) neurofibromin (2839 aa)								
1206-	1552	1wer-	3-	322	324	319	26	GTPase activation
38. Neurofibromatosis, Type 2								
>gi | 3980301 | emb | CAA76993.1 | (Y18000) NF2 protein (615 aa)								
21-	313	1ef1A	1-	294	294	293	61	
520-	579	1ef1C	16-	74	89	59	31	
39. Opitz G/BBB Syndrome								
>gi | 3462509 | gb | AAC33001.1 | (AF041209) midline 1 fetal kidney isoform 2 (667 aa)								
2-	88	1rmd_	18-	89	116	72	25	RING finger domain
380-	485	1cfb_	103-	204	205	97	15	Immunoglobulin-like

(continues)

Table 1
(Continued)

Qs	Qe	Template	Ts	Te	Tl	Match	identity	Fold[a]
40. Pancreatic Cancer								
>gi | 1163234 | gb | AAA91041.1 | (U44378) Dpc4 (552 aa)								
9-	138	1mhdB	1-	132	132	128	53	SMAD MH1 domain
319-	552	1ygs_	1-	234	234	234	100	SMAD/FHA domain
41. Paroxysmal Nocturnal Hemoglobinuria								
>gi | 219994 | dbj | BAA02019.1 | (D11466) PIG-A protein (484 aa)								
44-	401	1f0kA	8-	347	351	323	13	Flavodoxin-like
42. Pendred Syndrome								
>gi | 2654005 | gb | AAC51873.1 | (AF030880) pendrin (780 aa)								
654-	726	1auz_	41-	109	116	69	20	SpoIIaa-like
43. Polycystic Kidney Disease, Type 1								
>gi | 904223 | gb | AAC37576.1 | (L33243) polycystic kidney disease 1 (4302 aa)								
50-	139	1ds9A	54-	139	198	86	24	Leucine-rich repeat
275-	334	1b4r_	1_	80	80	80	100	Immunoglobulin-like
404-	534	1b6e_	1-	120	121	120	19	C-type lectin-like
1136-	1208	1b4r_	8-	78	80	70	39	Immunoglobulin-like
2069-	2131	1b4r_	8-	68	80	61	39	Immunoglobulin-like
3120-	3226	1ca1_	258-	368	370	106	13	Lipase/lipooxygenase
44. Polycystic Kidney Disease, Type 2								
>gi | 3126905 | gb | AAC16004.1 | (AF004873) polycystic kidney disease (968 aa)								
708-	778	1sra_	62-	137	151	71	15	EF Hand-like
45. Prader-Willi Syndrome								
>gi | 190247 | gb | AAA60151.1 | (J04564) snRNP polypeptide B (285 aa)								
7-	87	1d3bB	1-	81	81	81	100	Sm motif of SNRNP
104-	223	1jvr_	6-	121	137	116	20	Retroviral matrix
46. Retinitis Pigmentosa, X-linked								
>gi | 2204218 | emb | CAA66258.1 | (X97668) XLRP3 (815 aa)								
10-	363	1a12A	13-	393	401	332	23	7-bladed β-propeller
47. Situs Inversus, X-linked								
>gi | 2957266 | gb | AAC05594.1 | (AF028706) ZF protein of the cerebellum 3 (467 aa)								
250-	410	2gliA	1-	153	155	150	53	Classic zinc finger
48. Spinocerebellar Ataxia Type 7								
>gi | 2370155 | emb | CAA04154.1 | (AJ000517) spinocerebellar ataxia 7 (892 aa)								
400-	485	1jvr_	48-	135	137	87	20	Retroviral matrix
49. Stargardt's Disease								
>gi | 3243082 | gb | AAC23915.1 | (AF000148) ABC transporter (2273 aa)								
931-	1174	1b0uA	5-	257	257	223	25	P-loop NTPase
1938-	2176	1b0uA	3-	255	257	237	20	P-loop NTPase

(continues)

Table 1
(Continued)

Qs	Qe	Template	Ts	Te	Tl	Match	identity	Fold[a]
50. Tuberous Sclerosis								
>gi | 1063586 | gb | AAB41564.1 | (L48546) tuberin (1807 aa)								
43-	552	1ggrA	220-	709	871	485	9	alpha-alpha superhelix
>gi | 2331281 | gb | AAC51674.1 | (AF013168) hamartin (1164 aa)								
727-	965	1c1gC	44-	256	284	210	17	Parallel coiled-coil
51. Waardenburg Syndrome								
>gi | 431254 | gb | AAC50053.1 | (U02368) PAX3 protein (836 aa)								
35-	163	6paxA	2-	129	133	128	63	DNA-binding 3-helical
205-	277	1fj1A	4-	76	81	73	81	DNA-binding 3-helical
340-	460	2hfh_	3-	105	109	101	27	DNA-binding 3-helical
52. Werner Syndrome								
>gi | 6272686 | gb | AAF06162.1 | (AF181897) WRN (1432 aa)								
67-	363	1xw1_	3-	308	580	290	12	Ribonuclease H-like
527-	896	1d2mA	150-	545	552	360	11	P-loop NTPase
53. Williams Syndrome								
>gi | 1432164 | gb | AAB17545.1 | (U62293) LIM-kinasel (647 aa)								
19-	138	1b8tA	4-	170	192	118	24	GR-like (DNA-binding)
139-	271	1i16_	9-	130	130	118	18	PDZ domain-like
338-	605	1fmk_	185-	422	438	237	35	Protein kinase-like
54. Wilms Tumor								
>gi | 37978 | emb | CAA35956.1 | (X51630) Krueppel-like ZF protein (575 aa)								
447-	566	1tf6A	2-	122	179	118	34	Classic zinc finger
55. Wilson Disease								
>gi | 1947035 | gb | AAB52902.1 | (U03464) ATP7B (1465 aa)								
59-	127	1aw0_	4-	72	72	69	41	Ferredoxin-like
145-	210	1aw0_	5-	70	72	66	45	Ferredoxin-like
258-	329	1aw0_	4-	72	72	69	48	Ferredoxin-like
360-	426	1aw0_	4-	70	72	67	60	Ferredoxin-like
488-	557	1aw0_	3-	72	72	70	37	Ferredoxin-like
566-	632	1aw0_	5-	71	72	67	37	Ferredoxin-like
746-	1081	1eu1A	70-	392	994	307	21	
1185-	1352	1eu1A	590-	789	994	168	32	
56. Zellweger Syndrome								
>gi | 695566 | emb | CAA59324.1 | (X84899) peroxisomal C-terminal targeting (639 aa)								
346-	434	1a17_	55-	145	159	89	11	alpha-alpha superhelix
483-	589	1a17_	2-	111	159	107	17	alpha-alpha superhelix

[a] Following each disease heading the Genbank entry for the disease protein and the total number *N* of residues, in the form of "(*N* aa)", are listed. Alignment information for each domain is then listed. This includes the starting and ending residue numbers of the disease sequence (Qs and Qe), the PDB code of the template, and starting and ending residues of the template (Ts and Te), the total number of residues of the template (T1), and total number of aligned residues (Match), the sequence identity (%) of the alignment, and the structural fold of the template *(18).* In four diseases, two genes were involved.

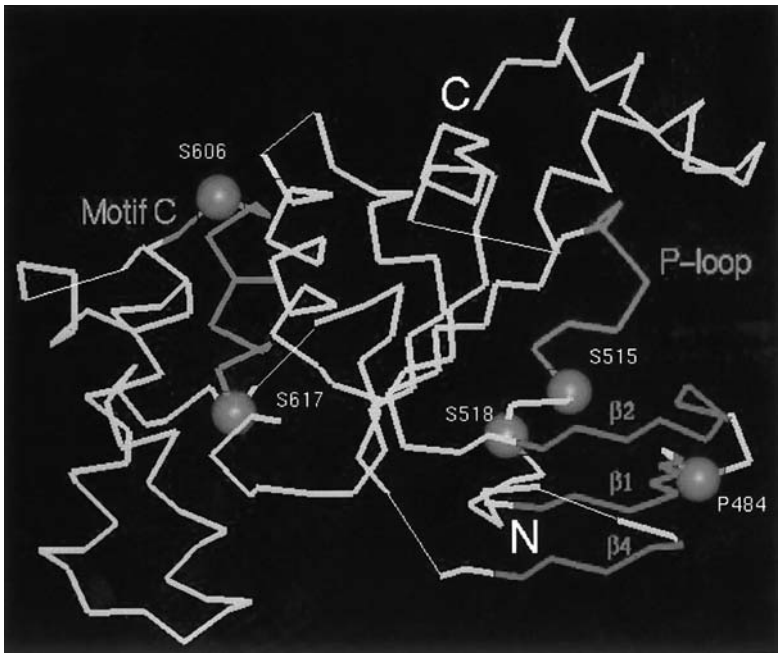

Fig. 1. Structure model for the ATP-binding domain of ALDP. The P-loop and motif C are highlighted. The β sheet consisting of β2, β1, and β4 forms the interface upon dimerization. The disease-causing mutation sites are denoted by spheres. This model was constructed by taking the corresponding C_α positions of 1b0uA using the alignment shown. In this alignment, the P-loop and motif C sequences are underlined and the conserved residues interacting with ATP are in bold.

```
1b0uA:   7  LHVIDLHKRYG-GHEVLKGVSLQARAGDVISIIGSSGSGKS   46
ALDP: 474  IICENIPIVTPSGEVVVASLNIRVEEGMHLLITGPNGCGKS  514

1bouA: 47  TFLRCINFLEKPSEGAIIVN-----GQNINLVRDKDGQLKV   82
ALDP: 515  SLFRILGGLW-PTYGGVLYKPPPDR----------------  538

1b0uA: 83  ADKNQLRLLRTRLTMVFQHFNLWSHMTVLENVMEAPIQVLG  123
ALDP: 539  ------------MFYIPQRP-YMSVGSLRDQVIYPDSVEDM  566

1b0uA:124  LSKHDARERALKYLAKVGIDERAQGK---------YPVHLS  155
ALDP: 567  QRKGYSEQDLEAILDVVHL-HHILQREGGWEAMCDWKDVLS  606

1b0uA:156  GGQQQRVSIARALAMEPDVLLFDEPTSALDPELVGEVLRIM  196
ALDP: 607  GGEKQRIGMARMFYHRPKYALLDECTSAVS----IDVEGKI  643

1b0uA:197  QQLAEE-GKTMVVVTHEMGFARHV-----------------  219
ALDP: 644  FQAAKDAGIALLSITHRPSLWKYHTHKKQFDGEGGWKFEKL  684

1b0uA:220  -SSHVIFLHQGKIEEEGDPEQVFGNPQSPRLQQFLKGSLK-  258
ALDP: 685  D----------------SAARLSLTEEKQRLEQQLAGIPKM  709

1b0uA:259  ---KLE  261
ALDP: 710  QRRLQE  715
```

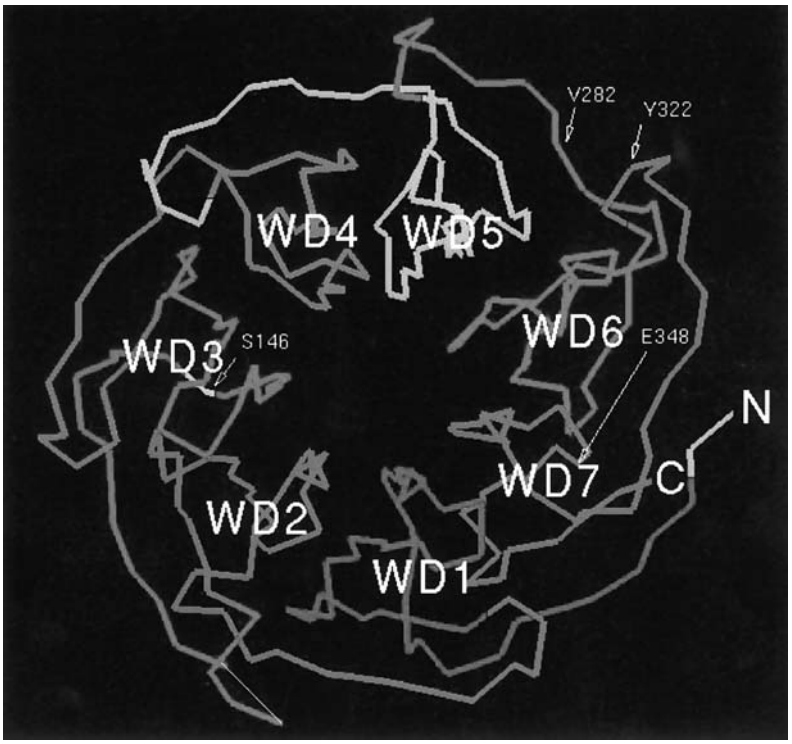

Fig. 2. Structure model for the CSA protein. The seven WD repeats are labeled. The disease-causing deletion sites are shown by arrows. This model was constructed by taking the corresponding C_α positions of 1gotB using the alignment shown. In this alignment the last two residues of all the seven WD repeats are in bold.

```
1gotB:  44  QMRTRRTLRGHLAKIYAMHW-GTDSRLLLSASQDGKLII WD   83
CSA:    32  LNKDRDVERIHGGGINTLDIEPVEGRYMLSGGSDGVIVL YD   72

1gotB:  84  SYTTNKVHAIPLRS-------------SWVMTCAYAPSGN-  110
CSA:    73  LENSSRQSYYTCKAVCSIGRDHPDVHRYSVETVQWYPHDTG  113

1gotB: 111  YVACGGLDNICSI YN LKTREGNVRVSRELAGHTGYLSCCRF  151
CSA:   114  MFTSSSFDKTLKV WD TNTLQ--TADVFNFEETVYSHHMSPV  152

1gotB: 152  LDD-NQIVTSSGDTTCAL WD IETGQQTTTFTGHTGDVMSLS  191
CAS:   153  STKHCLVAVGTRGPKVQL CD LKSGSCSHILQGHRQEILAVS  193

1gotB: 192  LAPD-TRLFVSGACDASAKL WD VREGM--------------  217
CSA:   194  WSPRYDYILATASADSRVKL WD VRRASGCLITLDQHNGKKS  234

1gotB: 218  --CRQTFTGHESDINAICFFPNGNAFATGSDDATCRL FD LR  256
CSA:   235  qaVESANTAHNGKVNGLCFTSDGLHLLTVGTDNRMRL WN SS  275

1gotB: 257  ADQELMTYS---HDNIICGITSVSFSKSGRLLLAGYDDFNC  294
CSA:   276  NGENTLVNYGKVCNNSKKGLKFTVSCGCSSEFVFVPYGSTI  316

1gotB: 295  NV WD ALKADRAGVLAGHDNRVSCLGVTDDGMAVATGSWDSF  335
CSA:   317  AV YT VYSGEQITMLKGHYKTVDCCVFQSNFQELYSGSRDCN  357

1gotB: 336  LKI WN  340
CSA:   358  ILA WV  362
```

Predicted Structure of CSA

The CSA protein has been proposed to contain five WD repeats *(5).* Based on the alignment with the PDB template 1gotB *(16),* we suggest that this protein is a seven-bladed β-propeller with seven WD repeats (*see* Fig. 2). The two newly proposed WD repeats, repeats 3 and 6, end with residues CD and YT, respectively. Interactions of the CSA protein with the CSB protein (listed in Table 1) and the p44 protein (a subunit of the RNA polymerase II basal transcription factor TFIIH) have been implicated in the disease mechanism of Cockayne syndrome *(5).* The N-terminal 146 residues were shown to be involved in these interactions. These residues make up the first three WD repeats (*see* Fig. 2). Noting that the interactions of 1gotB with 1gotA (the β and α subunits of the G protein transducin) occur exclusively in WD repeats 2 and 3 of 1gotB *(17),* we suggest that the interactions sites of the CSA protein with the CSB and p44 proteins are restricted to residues 73 to 172, which make up WD repeats 2 and 3.

Deletions of C-terminal portions (V282-E374, Y322-C terminal, and E348-E374) are known to be causes of the Cockayne syndrome (OMIM 216400). It is clear from Fig. 2 that these deletions will disrupt the structural integrity of the CSA protein. The last deletion may also interfere with the interactions between the CSA protein and its partners due to the proximity of the deleted residues and the putative interaction sites.

The aforementioned refined understanding on the mechanisms of adrenoleukodystrophy and the Cockayne syndrome illustrates the value of the predicted structure models. As the number of structural templates determined experimentally rapidly increases, the automated prediction methods will become more successful and accurate. It is expected that the *in silico* approach will soon take over as the primary tool for structural characterization of the proteins in the human genome. As an immediate extension of the present study, structural characterization of the over 1000 disease genes compiled under OMIM (Online Mendelian Inheritance in Man; www.ncbi.nlm.nih.gov) *(19,20)* is underway.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant GM58187.

REFERENCES

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.
3. Mosser, J., Douar, A.-M., Sarde, C.-O., Kioschis, P., Feil, R., Moser, H., et al. (1993) Putative X-linked adrenoleukodystrophy gene shares unexpected homology with ABC transporters. *Nature* **361,** 726–730.
4. Moser, H. W. (1997) Adrenoleukodystrophy: phenotype, genetics, pathogenesis and therapy. *Brain* **120,** 1485–1508.
5. Henning, K. A., Li, L., Iyer, N., McDaniel, L. D., Reagon, M. S., Legerski, R., et al. (1995) The Cockayne syndrome group A gene encodes a WD repeat protein that interacts with CSB protein and a subunit of RNA polymerase II TFIIH. *Cell* **82,** 555–564.
6. Shan, Y., Wang, G., and Zhou, H.-X. (2001) Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. *Proteins* **42,** 23–37.
7. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleio Acids Res.* **25,** 3389–3402.
8. Holm, L. and Sander, C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25,** 231–234.
9. Jones, D. J. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287,** 797–815.
10. Mushegian, A. R., Bassett, D. E., Boguski, M. S., Bork, P., and Koonin, E. V. (1997) Proc.

Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl. Acad. Sci. USA* **94,** 5831–5836.
11. Everett, L. A., Glaser, B., Beck, J. C., Idol, J. R., Buchs, A., Heyman, M., et al. (1997) Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genet.* **17,** 411–422.
12. Kovacs, H., Comfort, D., Lord, M., Campbell, I. D., and Yudkin, M. D. (1998) Solution structure of SpoIIAA, a phosphorylatable component of the system that regulates transcription factor sigmaF of Bacillus subtilis. *Proc. Natl. Acad. Sci. USA* **95,** 5067–5071.
13. Hung, L.-W., Wang, I. X., Nikaido, K., Liu, P.-Q., Ames, G. F.-L., and Kim, S.-H. (1998) Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature* **396,** 703–707.
14. Liu, L. X., Javier, K., Berteaux-Lecellier, V., Cartier, N., Benarous, R., and Aubourg, P. (1999) Homo- and heterodimerization of peroxisomal ATP-binding cassette half-transporters. *J. Biol. Chem.* **274,** 32,738–32,743.
15. Shani, N., Sapag, A., and Valle, D. (1996) Characterization and analysis of conserved motifs in a peroxisomal ATP-binding cassette transporter. *J. Biol. Chem.* **271,** 8725–8730.
16. Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E., and Sigler, P. B. (1996) Crystal structure of a G-protein beta gamma dimer at 2.1A resolution. *Nature* **379,** 369–374.
17. Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E., and Sigler, P. B. (1996) The 2.0 A crystal structure of a heterotrimeric G protein. *Nature* **379,** 311–319.
18. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540.
19. Antonarakis, S. E. and McKusick, V. A. (2000) OMIM passes the 1,000 disease gene mark. *Nature Genet.* **25,** 11.
20. Jimennez-Sanchez, G., Childs, B., and Valle, D. (2001) Human disease genes. *Nature* **409,** 853–855.