J Mol Model (2009) 15:113–122
DOI 10.1007/s00894-008-0381-1

ORIGINAL PAPER

An interrupted beta-propeller and protein disorder: structural bioinformatics insights into the N-terminus of alsin

Dinesh C. Soares · Paul N. Barlow · David J. Porteous · Rebecca S. Devon

Received: 21 August 2008 / Accepted: 5 November 2008 / Published online: 21 November 2008

Abstract Defects in the human *ALS2* gene, which encodes the 1,657-amino-acid residue protein alsin, are linked to several related motor neuron diseases. We created a structural model for the N-terminal 690-residue region of alsin through comparative modelling based on regulator of chromosome condensation 1 (RCC1). We propose that this alsin region contains seven RCC1-like repeats in a seven-bladed beta-propeller structure. The propeller is formed by a double clasp arrangement containing two segments (residues 1–218 and residues 525–690). The 306-residue insert region, predicted to lie within blade 5 and to be largely disordered, is poorly conserved across species. Surface patches of evolutionary conservation probably indicate locations of binding sites. Both disease-causing missense mutations—Cys157Tyr and Gly540Glu—are buried in the propeller and likely to be structurally disruptive. This study aids design of experimental studies by highlighting the importance of construct length, will enhance interpretation of protein–protein interactions, and enable rational site-directed mutagenesis.

Keywords Alsin · Beta-propeller · Comparative modeling · Fold recognition · Protein disorder · RCC1-repeat

Electronic supplementary material The online version of this article (doi:10.1007/s00894-008-0381-1) contains supplementary material, which is available to authorized users.

D. C. Soares (✉) · D. J. Porteous · R. S. Devon
Medical Genetics Section, Molecular Medicine Centre,
Institute of Genetics and Molecular Medicine,
Western General Hospital, University of Edinburgh,
Crewe Road,
Edinburgh EH4 2XU, UK
e-mail: Dinesh.Soares@ed.ac.uk

D. C. Soares · P. N. Barlow
School of Chemistry, University of Edinburgh,
Joseph Black Building, Kings Buildings, West Mains Road,
Edinburgh EH9 3JJ, UK

P. N. Barlow
Institute of Structural and Molecular Biology,
University of Edinburgh,
Michael Swann Building, Kings Buildings, Mayfield Road,
Edinburgh EH9 3JR, UK

Introduction

Mutations in alsin result in one of three related conditions that cause early onset of motor neuron disease (MND): amyotrophic lateral sclerosis 2 (ALS2), infantile onset ascending hereditary spastic paraplegia, and juvenile onset primary lateral sclerosis. In affected individuals, abnormal motor development is usually noticed in infancy. These children either never attain independent walking, or lose the ability during childhood. There follows a slow progression towards paraplegia. Pathologically, MND is characterized by the degeneration and death of large motor neurons in the cerebral cortex, brainstem and spinal cord [1].

The causes and molecular mechanisms of MND are poorly understood and there is no effective treatment. About 10% of MND is familial. Causative mutations are known in only a small minority amongst these 10% of cases. Therefore the identification and understanding of a protein such as alsin, which causes a Mendelian autosomal recessive form of the disease is important in the elucidation of biochemical pathways leading to the degeneration of motor neurons. The 1,657-residue neuronally expressed alsin protein [2] is predicted to act as a regulator of vesicle trafficking, particularly in early endosome pathways. Two domains—a Dbl-homology/pleckstrin-homology domain located centrally in the protein, and a vacuolar protein

sorting-9 domain at the C-terminus (Fig. 1a)—are the most widely studied. This has led to the description of alsin as a guanine nucleotide exchange factor (GEF) or activator for the small GTPase Rab5 [3, 4]. Alsin also acts either as a GEF [3, 5, 6], or alternatively as an effector, of Rac1 [7]. Rab5 and Rac1 are crucial regulators in the sorting and trafficking of endosomes, and in the maintenance of the actin cytoskeleton.

At the N terminus of alsin is the RCC1-like domain, which is homologous to the protein "regulator of chromosome condensation 1" (RCC1). RCC1 itself acts as a GEF for the nuclear GTPase Ran, but the cytoplasmic local-

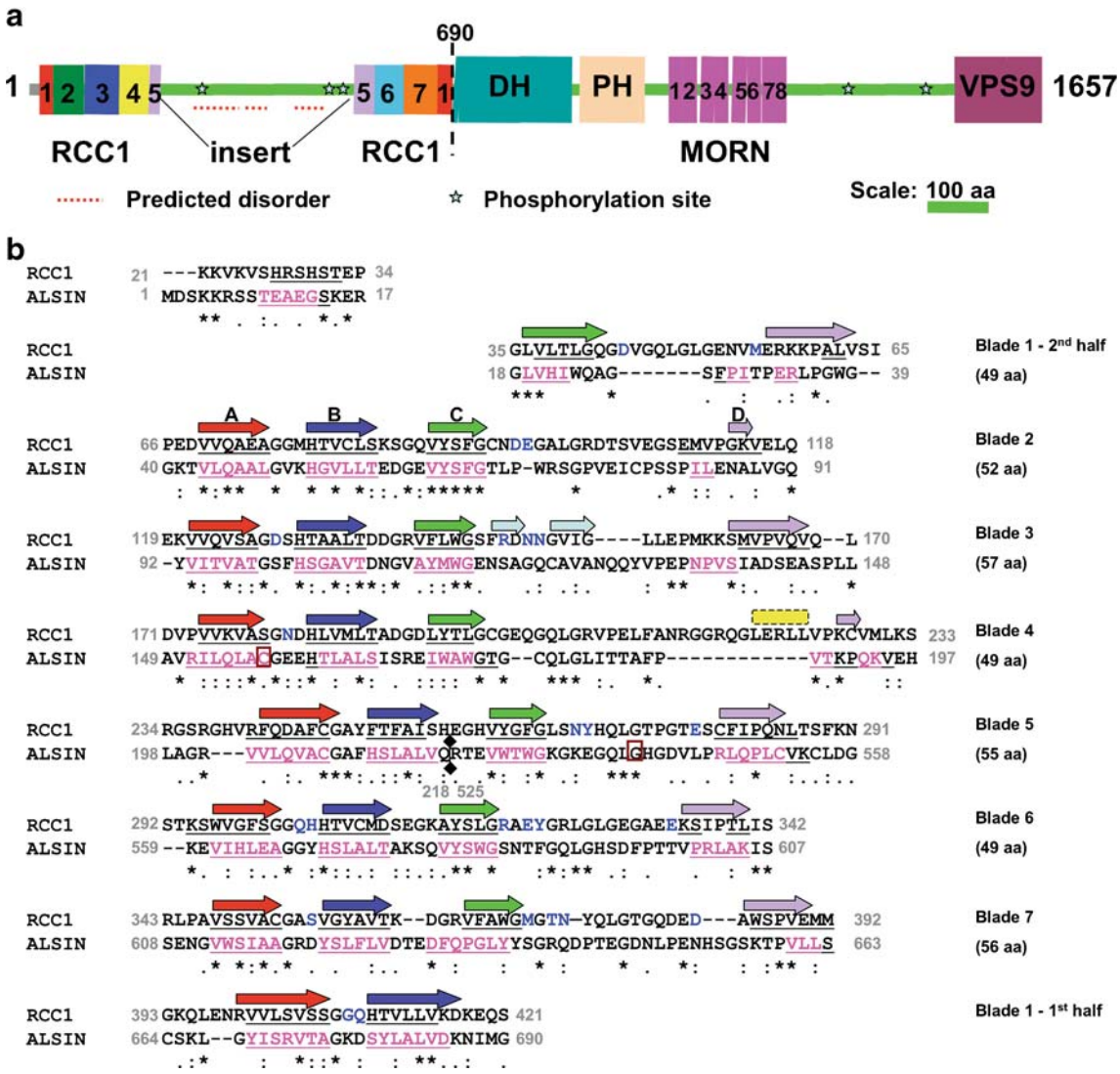

Fig. 1a,b Domain architecture and RCC1-like region alignment in alsin. **a** Protein domains are colour-coded and labelled along the protein length according to domain boundaries predicted from this study. *RCC1* Regulator of chromosome condensation 1, *DH* Dbl-homology domain, *PH* Pleckstrin-homology domain, *MORN* membrane occupation and recognition nexus motif, *VPS9* vacuolar protein sorting-9 domain, *GEF* guanine nucleotide exchange factor. **b** Target–template alignment. Optimal pairwise target (alsin)–template (RCC1) alignment corresponding to the region of the beta-propeller, i.e. the N-terminal tail and seven RCC1-like repeats or blades (numbered and labelled) used for modelling purposes. The blade boundaries are as per Renault et al. [11]. Blade 5 in human alsin is composed of two sequence segments, 198–218 and 525–558, the break in sequence number is indicated. *Asterisks* Completely conserved residues, *colons/dots* conservatively substituted residues. The PsiPred-predicted secondary structure (beta-strands) for human alsin are shown *underlined* along its sequence, and the clear majority consensus core beta-strands derived from Fig. S1 are coloured *pink*. Dictionary of protein secondary structure (DSSP [28])-identified secondary structure for the human RCC1 template is shown by *arrows* (beta-strands) above the sequence and colour-coded according to their position within each blade (*red* strand A, *blue* strand B, *green* strand C, *magenta* strand D). The two short extra beta-strands (*light blue*) in blade 3 and single short alpha-helix (*yellow*) in blade 4 are also indicated for the template. PsiPred-predicted secondary structure for the template RCC1 sequence is also indicated to aid alignment. The two disease-causing missense mutations C157Y (blade 4) and G540E (blade 5) are shown within a *brown box*. The Ran-binding residues for human RCC1 according to Renault et al. [12] are marked *blue*

isation of alsin and its lack of detectable Ran-GEF activity [4] render a similar function unlikely. Instead, the RCC1-like domain in alsin has been suggested to play a role in subcellular localisation and endosomal association [3, 4, 7–9], and to provide surfaces for protein–protein interactions. Indeed, this domain interacts with glutamate receptor interacting protein 1 (GRIP1) [10] and also, in vitro, with a C-terminal construct of alsin itself [7].

No three-dimensional (3-D) structural information exists for any part of alsin. The crystal structure of the human RCC1 protein, however, has been solved, revealing a seven-bladed beta-propeller [11, 12]. The beta-propeller (or 'super barrel') structure is present in many proteins (reviewed in [13–15]). It is comprised of four to eight blades (most commonly seven [16]) corresponding to sequence repeats each composed of a four-stranded antiparallel beta-sheet. The blades are arranged in a radial fashion around a central tunnel. While it has been postulated that, like RCC1, alsin may form a seven-bladed beta-propeller [3, 17, 18], previous sequence-based attempts at delineating the RCC1-like repeats and domain boundaries in alsin have reached conflicting conclusions [3, 17, 19].

We describe below comparative sequence alignment and modelling for the N-terminal domain of alsin. We conclude that this domain forms a seven-bladed beta-propeller. The structure closely matches that of RCC1, but accommodates a 306-residue insert that we predict to be largely disordered. We thus reveal an unusual relationship between sequence and structure that highlights the necessity of using adequate construct length in biochemical studies of this domain. We also predict the molecular effects of known disease-causing mutations and propose regions of likely protein–protein interaction.

Methods

Homology searching for the N-terminal 690-amino acid sequence of human alsin (SwissProt: ALS2_HUMAN) was performed using a standard protein BLAST (tblastn) [20, 21] search against the translated non-redundant database, via the NCBI www-server (http://www.ncbi.nlm.nih.gov/BLAST/) with default parameters. Multiple sequence alignments were generated by ProbCons version 1.5 [22] but subjected to manual editing based on conservation of residues and continuity of secondary structure elements as predicted by PsiPred version 2.5 [23, 24]. All-against-all sequence identities were calculated using the percentage identity matrix (PIM) option under ClustalX [25].

Fold recognition for the 384-amino acid residue sequence (1–218 and 525–690) corresponding to the seven putative RCC1-like repeats in human alsin was performed by the BioInfoBank Meta Server (http://meta.bioinfo.pl/) [26] and 3D-PSSM [27]. Optimal alignment between the alsin target and the RCC1 template is critical for success of the modelling exercise, and in this case was generated based upon conservation of both sequence and secondary structure of target and template in consideration with the alsin-orthologue multiple sequence alignment. Secondary structure for the template was identified by the dictionary of protein secondary structure (DSSP) [28]. This approach helped us with, for example, establishing the register of the third beta-strand in blades 1 and 7 (see Fig. 1b) despite the fact that these repeats in alsin lack the sequence motif [F/W] G that occurs at the ends of the third strands in the other five blades. Secondary structure considerations also led us to disregard an apparently conserved proline residue that occurs at the start of the fourth beta-strand—were an alignment of these proline residues to be forced, the secondary structural alignment would be compromised (alternative alignment not shown).

Fifty models were generated using Modeller 8v2 [29] and the five models with lowest objective function scores [29] were evaluated using PROCHECK v3.5.4 [30]. The one with most appropriate stereochemistry was selected as the representative model. Non-identical side-chain residues were further optimised using SCWRL v3 [31, 32]. The model was protonated under SYBYL v6.9 (Tripos Associates, St. Louis, MO), subjected to energy minimisation to reduce clashes and bad geometries, and evaluated using PROCHECK [30]. The packing quality [33] of the model was assessed using WHATIF [34], and model quality additionally assessed using ProQ [35] and the MetaMQAP server [36]. Solvent-accessibility calculations were performed by GETAREA v1.1 [37]. Electrostatic surface representation of the model was generated using GRASP [38]. PyMol (DeLano Scientific LLC, Palo Alto, CA) was used for visualisation and analysis. Side-chain mutations were undertaken using SCWRL v3 [31, 32], and their effects analysed using WHATIF [34] along with visual analysis using the 'mutagenesis wizard' under PyMol.

Disorder prediction was performed using seven different prediction servers/methods: DisEMBL v1.5 [39]; DISPROT-VSL2 [40, 41]; FoldIndex [42]; PreLink [43]; DRIP-PRED (http://www.sbc.su.se/~maccallr/disorder/); PONDR [44]; and DISOPRED2 [45]. In the case of DisEMBL, the prediction output option from REMARK-465 (missing coordinates in X-ray structure defined by REMARK-465 entries in PDB) was selected. In the case of VSL2 and PONDR, the predictor models "VSL2B" and "VLXT", respectively, were selected. In order to prevent over-prediction, a disorder consensus was derived. This was achieved by applying a simple majority rule for each position, i.e. disorder was predicted for a sequence position if four or more servers/methods agreed.

Results and discussion

The N terminus of alsin encodes seven RCC1-like repeats that are conserved across species

Ten alsin orthologues (including the human sequence) were identified and retrieved by sequence homology to the N-terminal 690-amino acid sequence of human alsin: GenBank identifiers (gi): *Homo sapiens* gi|15823635; *Pan troglodytes* gi|60686966; *Macaca mulatta* gi|109100596; *Canis familiaris* gi|74005605; *Bos taurus* gi|76610251; *Mus musculus* gi|15823639; *Rattus norvegicus* gi|61740638; *Gallus gallus* gi|118093427; *Takifugu rubripes* gi|60686962; *Danio rerio* gi|68361893. Other alsin-like sequences were detected but these were incomplete and hence discarded.

The ten sequences were then split into two segments for multiple alignment with the human RCC1 sequence (Fig. 1b). The two segments corresponded to (i) RCC1-like sequence regions (residues 1–218 and 525–690 in *Homo sapiens*) (Fig. S1), and (ii) an inserted sequence region (residues 219–524 in *Homo sapiens*) (Fig. S2). Seven RCC1-like repeats of appropriate blade-forming length (49–57 amino acid residues) were identified in segment (i). Four beta-strands (labelled A–D) were confidently predicted within each repeat, with beta-strands A–C showing the greatest degree of consensus (Fig. S1a, Fig. 1b).

The alsin RCC1-like repeats are highly conserved among orthologues. These regions are 99% identical between human and chimpanzee, and 54% identical between human and zebra fish (Fig. S1b). Only two residue differences, both of which lie in the RCC1-like regions, exist between the N-terminal 690 amino acid residues of human and chimpanzee alsin; these are the conservative substitutions Ile94Val and Ile615Val. Of the RCC1-like repeats, the 4th, 5th and 6th are the most highly conserved among species; for example, repeat 5 is identical in human, chimpanzee, macaque, dog, cow, mouse and rat (Fig. S3). Repeat 5 also shows high sequence conservation between the human alsin target and the RCC1 template. Conservation within repeat 5 extends into regions immediately adjacent to the inserted sequence (Fig. 1b), which helps delineate the potential start and end positions for the 306-amino acid inserted region. Overall, the target and template share 23.2% pairwise sequence identity and 44% sequence similarity. The target–template alignment reveals that the 25 amino acid residues of RCC1 known to bind Ran [12] are not conserved in alsin (Fig. 1b) consistent with its lack of observed binding to Ran [4].

The RCC1-like repeats of alsin form a seven-bladed beta-propeller stabilised by a double molecular clasp

As expected, fold recognition searches for the 384-amino acid residue RCC1-like region [segment (i)] of alsin resulted in a significant hit with the two human RCC1 X-ray crystal structures (protein data bank, PDB [46] ID: 1A12 and PDB ID: 1I2M, [11, 12]). The 3-D model of alsin was built based upon the higher-resolved set of PDB co-ordinates, 1A12 (chain A), using the optimal target–template alignment. A global target-to-template alignment approach was applied because the sequence and secondary structure similarity between the two was sufficient over the length of the propeller region. Several reports of beta-propeller modelling have been published based upon the global target-to-template approach [47–51]. An alternative fragment-based approach where each blade is modelled individually and then assembled would be a less-optimal procedure in this case since it would struggle to emulate and maintain the inter-blade contacts (packing quality) and continuity/connectivity of the resulting model. Stereochemical analysis of our model revealed good dihedral statistics (Ramachandran plot scores: 89.3% residues in most favoured regions, 7.6% in additional allowed regions, 2.8% in generously allowed regions, 0.3% in disallowed regions). The packing quality of the model attained an overall structural average quality control score of −1.47. To place this in context, incorrect models give a score of less than −3.0, and for poor models the score is less than −2.0 [33, 34]. The packing quality is very similar to a published model of another beta-propeller with an inserted domain [52]. Finally, ProQ and the MetaMQAP scores confirm the validity of the model (ProQ: LGscore 5.56, “extremely good model”; MetaMQAP: GDT_TS 61.65, RMSD 2.94 Å).

As expected for template-based homology models, the core regions of the model closely resemble the RCC1 structure. Thus, like human RCC1 [11, 12], alsin has a pseudo-seven-fold symmetry with an overall appearance of a propeller made up from seven blades each corresponding to a RCC1-like repeat (Fig. 2). Strand A of each sheet is located facing the central tunnel of the propeller, while strand D lies on the outer surface. The N- and C-terminal tails of the RCC1-like domain lie on the same face of the propeller but project in opposite directions. Neither blade 1 nor blade 5 is composed from contiguous sequence. Within blade 1, the C and D strands are provided by the N terminus of the region (residues 18–39) and the A and B strands are contributed by the C terminus (residues 664–690). Strands A, B and C, D from within blade 5 residues 198–218 and 525–558, respectively. Hydrogen-bonding between these regions in blade 1 (a 2N + 2C strand closure) and in blade 5 (a 2 + 2 strand closure either side of the insertion) thus acts as a double “molecular clasp” or “Velcro” to stabilise the circular arrangement [11, 14, 15, 51, 53–55].

An alignment of the seven repeats based on the 3-D model structure (Fig. S4) shows, as in RCC1, the presence of an invariant glycine residue in a tight-turn position at the end of each strand A, plus highly and semi-conserved

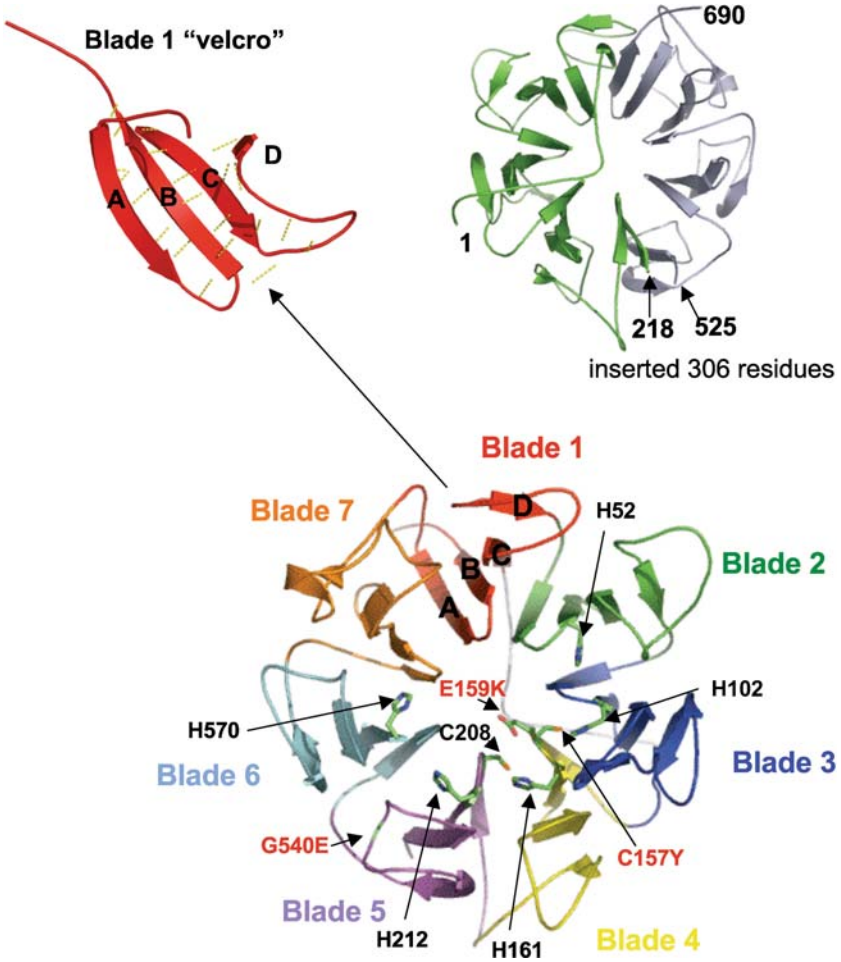

Fig. 2 Three-dimensional (3-D) model of alsin RCC1 domain. The main frame shows a cartoon representation of the seven-bladed beta-propeller of the RCC1-like domain alsin model. The four strands (*A*, *B*, *C* and *D*) present in each blade are labelled on the strands of blade 1, and all seven blades are coloured and labelled individually. The locations of the two missense disease-causing mutations (side-chain only for Cys157, and Gly540 shown in *green*) are indicated on blades 4 and 5 along with the SNP Glu159Lys (E159K) in *red*. The side-chains of the histidine residues that line the central tunnel and connect the blades are also shown and labelled. All loops have been rendered smooth for clarity. The *top-left frame* shows the domain-stabilising "Velcro" in blade 1 between strands B and C from the C- and N-termini, respectively (note: for clarity a similar "Velcro"-like H-bond network occurring within blade 5 is not shown). The main-chain hydrogen-bonds within this blade are indicated by *dashed yellow lines* (all main-chain H-bonds for the model are shown in Fig. S4). The *top right frame* depicts the two segments of the alsin 3-D model (rotated 180° about the *y*-axis from the main frame figure), which correspond to the two halves of the propeller labelled 1–218 (*green*) and 525–690 (*blue-grey*), with the point of insertion indicated

hydrophobic residues within strands A, B and C (Fig. 1b, Figs. S1a, S4). Additionally, buried histidine residues form the start of strand B in blades 2, 3, 4, 5 and 6 (His52, His102, His161, His212, His570); these line the central tunnel and participate in inter-blade H-bonds (Fig. 2). A motif at the end of strand C usually contains an aromatic residue (often Trp) and an adjacent glycine. This motif is involved in forming part of the hydrophobic core between the blades. In the model, the "WG" motif is part of a largely conserved VYxWGT RCC1-like consensus motif [11]. Even in the absence of this motif, blades 1 and 7 contain four-predicted strands and retain the structurally important glycine at the top of strand A, plus other key non-polar residues within strands A–C, ideal for blade formation [13, 14, 56]. Superimposition of the seven blades (Fig. S4) also demonstrates that most of the variation may be found in the loop between strand C and strand D, and in the loop connecting strand D of one blade with strand A of the next. Thus, as in other beta-propellers, strand D is variable and irregular (compared to strands A, B and C) [15] reflecting its location on the outer surface; it does not play a major structural role as evidenced in the beta-propeller structure of beta-lactamase inhibitor protein-II [54]. Indeed this variability in strand D has been suggested to play a role in providing versatility for protein–protein interactions and function [57].

The identification of the structural repeats and inference of a seven-bladed propeller on the basis of our model do

not coincide with previously proposed sequence repeats in alsin. For example, Hadano et al. [17] identified three RCC1-like regions while Yang et al. [19] proposed six RCC1-like repeats and SwissProt [58] lists five repeats. Topp et al. [3] identified five RCC1-like repeats arranged in a configuration of three then two repeats interrupted by an ~300 residue insert and suggested that two other repeats may exist. These discrepancies may have arisen for three reasons: (1) the circular nature of the propeller and the molecular clasp whereby N- and C-terminal sequences contribute to a single blade; (2) the presence of the 306-residue insert (residues 219–524) that partitions the fifth repeat, obscuring the obvious "signal" of a continuous seven-bladed RCC1-like beta-propeller; (3) repeats 1 and 7 show structural similarity to RCC1 in the absence of obvious sequence similarity (Figs. 1b, 2).

Implications for truncated forms of alsin

In addition to the full-length form, there exists a short transcript of the *ALS2* gene, comprising exons 1–4 only (instead of 1–34) [17]. There is no experimental evidence that this short form is translated into the predicted 396 amino acid peptide [2, 4, 9, 59, 60], although its existence has often been assumed. If translated, the proposed short form would contain only the first half of the modelled beta-propeller followed by 154 residues of the insert region and 26 novel residues. Because alsin short-form protein is not biochemically detectable [2, 4, 9, 59, 60], it is likely that the short-form cannot fold appropriately and is rapidly degraded. However, beta-propellers composed of a variable number of blades do exist, and are thought to evolve by blade duplication and deletion [61]. There are therefore at least three other possibilities: (1) the terminal blades of the short form (i.e. the second half of blade 1 and first half of blade 5) could engage with one another to form a smaller four-bladed propeller; (2) the truncated "half-propeller" with four blades could form a symmetrical homodimer, creating an eight-blade propeller; or (3) an "open" ring-propeller could be formed as precedented by the C-terminal domain of topoisomerase IV ParC subunit [62]. In our model, the C-terminus of the RCC1-like domain lies at amino acid 690. This is entirely consistent with the observation that the equivalent to the following residue is the first residue of the ALS2 C-terminal-like protein; this lacks the RCC1-like domain but otherwise shows a high degree of homology to the rest of alsin [2, 63].

Several published investigations of the function or subcellular localisation of the alsin N-terminal domain have used artificial short constructs for in vitro studies. Such RCC1-like constructs consisted of residues 1–666 [10], 1–680 [4, 7] and 1–705 [3]. Our model reveals the importance of the inclusion of the entire domain 1–690 in such constructs, so as to allow formation of blade 1 and the completion of the stable propeller structure. There is evidence that excising one blade of a seven-bladed propeller (prolyl oligopeptidase) leads to lower stability and artefactual dimerisation [64]. Thus we conclude that arbitrary alsin construct length, particularly in light of the obligatory molecular clasp arrangement, would not allow for appropriate folding, causing artefactual higher order structures, potentially skewing results. We suggest that previous results based upon RCC1-like constructs shorter than 690 amino acids require verification and that future experimentation should utilise a construct of at least amino acids 1–690. It is possible that the published interaction between residues 1–666 of alsin and GRIP1 [10] may occur within the insert region, instead of within the beta-propeller.

The beta-propeller model reveals highly conserved and negatively charged surface 'patches', and a potential disulfide bond

In addition to those conserved buried residues required to maintain structure and blade architecture, there are several exposed residues that are conserved among orthologues (highlighted on the surface of the model, Fig. 3b). Some of these residues cluster in prominent patches that could correspond to regions that are important for protein–protein interactions or other functions; these could form the focus of site-directed mutagenesis experiments. Most notable are a 'C'-shaped patch surrounding the central tunnel on one face of the propeller, and an elongated region at the extremity of the propeller on the opposite face (Fig. 3b). It is worth noting that the central tunnel of beta-propellers often binds prosthetic groups and harbours catalytic residues [14, 15]. This region in alsin features a negatively charged region (Fig. 3a) that in part overlaps with the 'C'-shaped patch of conserved residues. Similar negatively charged residue-containing clusters in other proteins have been shown to bind to metals [65, 66]. In this respect it could be relevant that the protein superoxide dismutase 1 (SOD1), which is mutated in 10–20% of familial ALS cases [67] binds both copper and zinc [68]; and alteration of its zinc-binding capacity may play a role in ALS pathogenesis, via a toxic gain-of-function [69]. Alsin has been shown to physically interact with mutant SOD1 [70, 71]. Given the effects of dietary metals on animal models of MND [72–76], the prospect of a common functional metal-binding link is tantalising.

There are two strictly conserved cysteine residues in the model, Cys157 (blade 4) and Cys208 (blade 5), whose alpha carbons are sufficiently close (5.8 Å) that their side-chains could participate in a disulfide bond [77] (we did not set side-chain disulfide bond restraints in the modelling protocol). Such an inter-blade disulfide bond could other-

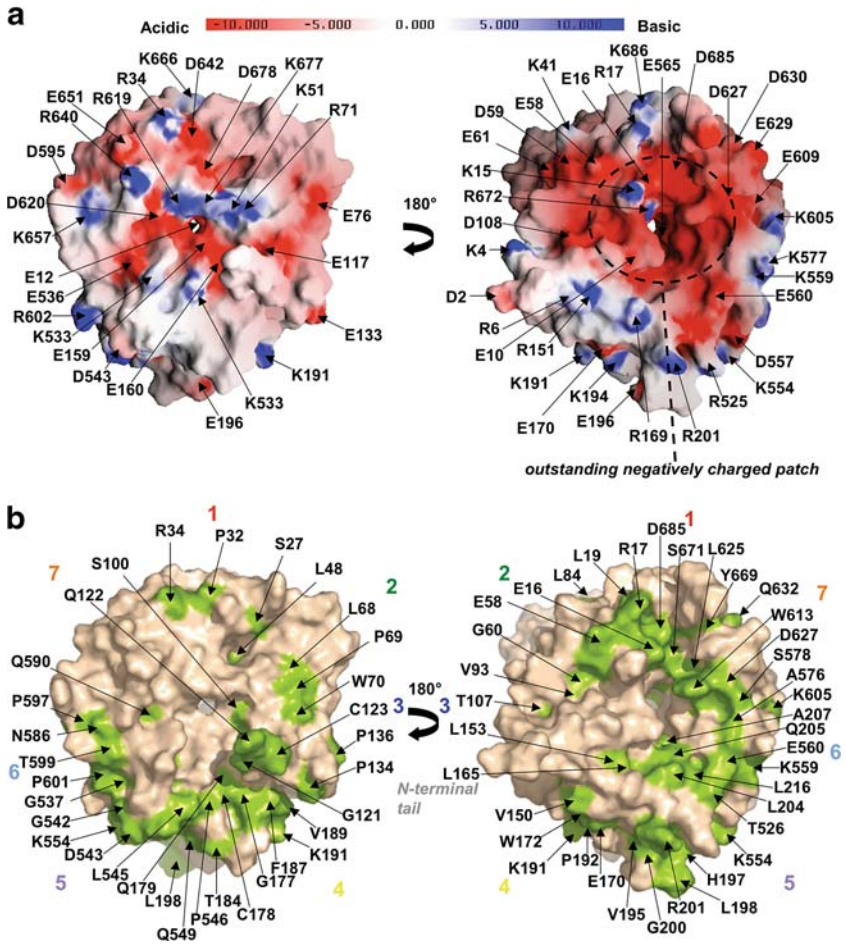

Fig. 3a,b Surface property representations of the alsin RCC1 3-D model. **a** Two views rotated by 180° about the *y*-axis of a GRASP [38] electrostatic surface representation of the model. The molecule appears to expose many charged residues (*labelled*). Negative charge is coloured *red* and positive charge *blue*, ranging from −10 kT to +10 kT (k=Boltzmann's constant; T=temperature in Kelvin). **b** Sequence conservation mapped onto model surface. Surface representation of conserved patches is based upon the orthologue multiple sequence alignment of RCC1-like regions in Fig. S1, where strictly conserved residues are shown in *green*. The 'C'-shaped patch is observable on the right-hand side. The surface images are in equivalent orientations to each other, and to those in Fig. 2

wise play a crucial role in stabilising the circular arrangement of blades. Indeed, the four-bladed haemopexin and collagenase beta-propellers use a disulfide bond to seal and stabilise the N- and C-termini circular array (instead of the typical Velcro) (reviewed in [14, 78]). Although the intracellular reducing environment in which alsin is localised would normally preclude formation of a disulfide, we note that the SOD1 protein exhibits a rare intracellular disulfide bond [68, 79, 80]; hence a similar occurrence in alsin cannot be ruled out entirely. Interestingly, Cys157 is the location of the Cys157Tyr ALS2-linked missense mutation (see below).

Known missense mutations in alsin would disrupt the propeller structure

The model enables us to make predictions about the structural effects of the two known alsin missense mutations: Cys157Tyr (previously incorrectly numbered Cys156) [81] and Gly540Glu [8], and other reported single nucleotide polymorphisms (SNPs) such as Ile94Val, Glu159Lys, and Met368Val [82].

Cys157 is strictly conserved among orthologues and is located at the end of strand A in blade 4. It is deeply buried (Fig. 2, Fig. S5a) suggesting a structural role. In silico substitution of this cysteine residue with tyrosine resulted in abnormally short inter-atomic distances [34] between the tyrosine side-chain and four residues from blade 3 (Ala97, His102, Gly104 and Met114). Thus such a mutation at the inter-blade interface is likely to disrupt structure.

Gly540 is also strictly conserved amongst orthologues. It occurs in the loop between strands C and D of blade 5 (Fig. 2) and is buried at its interface with blade 6. In silico mutation to Glu results in clashes of the Glu540 side-chain with the Trp583 side-chain (Trp583 is part of the VYxWG motif within blade 6). Hence this substitution (Fig. S5b)

probably destabilises the structure. Such an inference is consistent with proteolytic susceptibility and hence with the observed functional effects of the Gly540Glu mutation, namely protein delocalisation and cytotoxicity [8].

Glu159 is a highly conserved residue that lies within the loop between strand A and B in blade 4 with its side-chain partially exposed. Substitution with a Lys residue would alter surface charge and likely have a potential effect on function rather than structure. This SNP is detected in both ALS patients and controls, so is unlikely to contribute to disease pathogenesis, although since it is so rare (~1 in 200 alleles in both populations) it is difficult to draw meaningful conclusions. Of other SNPs: Ile94Val (within blade 3) is a conservative replacement of a non-buried residue (and this position is in any case occupied by a valine in several species; Fig. S1a) and is unlikely to have a major impact on structure or function; Met368Val corresponds to a non-conserved residue within the insert region (see below; Supplementary Fig. S2) so it is difficult to speculate on any structural or functional role. Note, however, that this residue occurs as a Val in macaque and cow.

The beta-propeller in alsin is interrupted by a 306 amino acid insertion that is largely disordered

The sequence encoding the RCC1-like beta-propeller in alsin is interrupted, between strands B and C of blade 5, by a 306-amino acid insertion (residues 219–524). Note that this fifth RCC1-repeat exhibits a remarkable degree of cross-species conservation, which is presumably necessary to maintain a structural scaffold for accommodation of the 306-residue insert. The region corresponding to the insert has lower levels of cross-species conservation than the RCC1-like regions (for example, 47% sequence identity between human and zebra fish) (Fig. S2) and contains few predicted secondary structure elements (Fig. S6).

The insert sequence was submitted to seven prediction servers to gauge its regions of order or disorder (Fig. S6). More than 50% of residues (157 out of 306) are predicted to be disordered; these cluster in three major regions: residue positions 266–339 (74 residues), 354–388 (35 residues) and 433–480 (48 residues). All seven methods agreed on disorder for 44 out of 306 residue positions (~14%). By contrast, the region spanning residue positions 400–428 was predicted to be "ordered" by all methods thus implying that the insert region possesses at least some residual structure. In support of this, PsiPred predicted two alpha-helices within this region.

There is precedence for propeller structures having inserted domains. For instance, the sequence of the beta-propeller in neuraminidase [83] is interrupted by almost 200 residues, and both the leech trans-sialidase propeller [84] and the integrin alpha M subunit [52, 85, 86] have large propeller insertions. In our alsin-RCC1 model, the insertion occurs entirely within a loop region (Figs. 1b, 2). This is consistent with insertions in other proteins [87, 88] and is likely indicative of a gene insertion event [87]. Notwithstanding the number of disordered residues, it is possible that more than one domain resides within this 306 amino acid insert region, since 80% of inserted domains are less than 175 amino acids in length [87]. Indeed, the Meta-DP server [89] predicted the possibility of two domains for this region. Domain and fold recognition searches, however, did not reveal any significant homologies.

The disordered regions of the insert will lack specific tertiary structure and will exist in a range of conformations [90]. Disorder is common among eukaryotic proteins (33% contain at least some disordered regions) and, in general, the function of disordered proteins involves binding to a ligand [90] accompanied by a structural transition into a folded form. Such a transition was proposed to permit a rapid response and play a role in transmission of cellular signals in numerous processes including endocytosis [90], which is not inconsistent with the proposed role of alsin as an endocytotic regulator [18]. The reduced sequence conservation of the insert region amongst alsin orthologues might imply that there are subtle inter-species differences in the ligand interactions it confers, which may play a role in the unexpectedly mild phenotype of alsin-null mice [18]. Interestingly, three out of the five experimentally determined phosphorylation sites within alsin [5] lie within regions that were predicted to be disordered by one or more method (Ser277, Ser492 and Thr510). This agrees with previous evidence that phosphorylation frequently occurs within intrinsically disordered protein regions [91, 92].

Finally, it is often difficult to express disordered proteins in sufficient quantities, and their purification is challenging [39]. Alternative strategies such as co-expression with a known binding partner, or deletion of potentially disordered segments to increase expression, stability and foldability of the protein construct, are needed to successfully crystallise the protein. The identification of these disordered regions embedded in the alsin N-terminal beta-propeller offers a rational route towards future attempts at structure determination.

Conclusions

This work sheds light on the structure and function of the N-terminal 690 amino acids of alsin, and gives rise to hypotheses that would benefit from experimental confirmation. The convergence of multiple lines of evidence from sequence analysis, secondary structure prediction and the good quality of the 3-D model provides strong support that the RCC1-like region of alsin forms a seven-bladed beta-propeller stabilised by a double clasp. The misfolding/

folding possibilities of the proposed short form of alsin are presented. The propeller surface possesses regions of high sequence conservation that represent promising sites for mutagenesis experiments to probe residues involved in protein–protein interactions. The presence of a prominent negatively charged region at the entrance to the central tunnel may indicate a metal-binding capability. A large insertion sequence that occurs within the best conserved repeat in the propeller is largely disordered and may play an important role in ligand binding. Both missense mutations previously linked to disease will disrupt the structure.

Acknowledgements We would like to thank Dr. Andrew F. Coulson for critical reading of the manuscript, and Dr. Alice J. Walmesley and Prof. Lindsay Sawyer for helpful discussions. D.C.S. acknowledges funding from a Strategic Research Development Grant. The model co-ordinates are available from D.C.S. upon request.

References

1. Brown RH Jr (1995) Cell 80:687–692. doi:10.1016/0092-8674(95)90346-1
2. Devon RS, Schwab C, Topp JD, Orban PC, Yang YZ, Pape TD, Helm JR, Davidson TL, Rogers DA, Gros-Louis F, Rouleau G, Horazdovsky BF, Leavitt BR, Hayden MR (2005) Neurobiol Dis 18:243–257. doi:10.1016/j.nbd.2004.10.002
3. Topp JD, Gray NW, Gerard RD, Horazdovsky BF (2004) J Biol Chem 279:24612–24623. doi:10.1074/jbc.M313504200
4. Otomo A, Hadano S, Okada T, Mizumura H, Kunita R, Nishijima H, Showguchi-Miyata J, Yanagisawa Y, Kohiki E, Suga E, Yasuda M, Osuga H, Nishimoto T, Narumiya S, Ikeda JE (2003) Hum Mol Genet 12:1671–1687. doi:10.1093/hmg/ddg184
5. Tudor EL, Perkinton MS, Schmidt A, Ackerley S, Brownlees J, Jacobsen NJ, Byers HL, Ward M, Hall A, Leigh PN, Shaw CE, McLoughlin DM, Miller CC (2005) J Biol Chem 280:34735–34740. doi:10.1074/jbc.M506216200
6. Kanekura K, Hashimoto Y, Kita Y, Sasabe J, Aiso S, Nishimoto I, Matsuoka M (2005) J Biol Chem 280:4532–4543. doi:10.1074/jbc.M410508200
7. Kunita R, Otomo A, Mizumura H, Suzuki-Utsunomiya K, Hadano S, Ikeda JE (2007) J Biol Chem 282:16599–16611. doi:10.1074/jbc.M610682200
8. Panzeri C, De Palma C, Martinuzzi A, Daga A, De Polo G, Bresolin N, Miller CC, Tudor EL, Clementi E, Bassi MT (2006) Brain 129:1710–1719. doi:10.1093/brain/awl104
9. Yamanaka K, Vande Velde C, Eymard-Pierre E, Bertini E, Boespflug-Tanguy O, Cleveland DW (2003) Proc Natl Acad Sci USA 100:16041–16046. doi:10.1073/pnas.2635267100
10. Lai C, Xie C, McCormack SG, Chiang HC, Michalak MK, Lin X, Chandran J, Shim H, Shimoji M, Cookson MR, Huganir RL, Rothstein JD, Price DL, Wong PC, Martin LJ, Zhu JJ, Cai H (2006) J Neurosci 26:11798–11806. doi:10.1523/JNEUROSCI.2084-06.2006
11. Renault L, Nassar N, Vetter I, Becker J, Klebe C, Roth M, Wittinghofer A (1998) Nature 392:97–101. doi:10.1038/32204
12. Renault L, Kuhlmann J, Henkel A, Wittinghofer A (2001) Cell 105:245–255. doi:10.1016/S0092-8674(01)00315-4
13. Pons T, Gomez R, Chinea G, Valencia A (2003) Curr Med Chem 10:505–524
14. Paoli M (2001) Prog Biophys Mol Biol 76:103–130. doi:10.1016/S0079-6107(01)00007-4
15. Jawad Z, Paoli M (2002) Structure 10:447–454. doi:10.1016/S0969-2126(02)00750-5
16. Murzin AG (1992) Proteins 14:191–201. doi:10.1002/prot.340140206
17. Hadano S, Hand CK, Osuga H, Yanagisawa Y, Otomo A, Devon RS, Miyamoto N, Showguchi-Miyata J, Okada Y, Singaraja R, Figlewicz DA, Kwiatkowski T, Hosler BA, Sagie T, Skaug J, Nasir J, Brown RH Jr, Scherer SW, Rouleau GA, Hayden MR, Ikeda JE (2001) Nat Genet 29:166–173. doi:10.1038/ng1001-166
18. Hadano S, Kunita R, Otomo A, Suzuki-Utsunomiya K, Ikeda JE (2007) Neurochem Int 51:74–84. doi:10.1016/j.neuint.2007.04.010
19. Yang Y, Hentati A, Deng HX, Dabbagh O, Sasaki T, Hirano M, Hung WY, Ouahchi K, Yan J, Azim AC, Cole N, Gascon G, Yagmour A, Ben-Hamida M, Pericak-Vance M, Hentati F, Siddique T (2001) Nat Genet 29:160–165. doi:10.1038/ng1001-160
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) J Mol Biol 215:403–410
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389–3402. doi:10.1093/nar/25.17.3389
22. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) Genome Res 15:330–340. doi:10.1101/gr.2821705
23. Jones DT (1999) J Mol Biol 292:195–202. doi:10.1006/jmbi.1999.3091
24. McGuffin LJ, Bryson K, Jones DT (2000) Bioinformatics 16:404–405. doi:10.1093/bioinformatics/16.4.404
25. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) Nucleic Acids Res 25:4876–4882. doi:10.1093/nar/25.24.4876
26. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) Bioinformatics 19:1015–1018. doi:10.1093/bioinformatics/btg124
27. Kelley LA, MacCallum RM, Sternberg MJ (2000) J Mol Biol 299:499–520. doi:10.1006/jmbi.2000.3741
28. Kabsch W, Sander C (1983) Biopolymers 22:2577–2637. doi:10.1002/bip.360221211
29. Sali A, Blundell TL (1993) J Mol Biol 234:779–815. doi:10.1006/jmbi.1993.1626
30. Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) J Appl Cryst 26:283–291. doi:10.1107/S0021889892009944
31. Bower MJ, Cohen FE, Dunbrack RL Jr (1997) J Mol Biol 267:1268–1282. doi:10.1006/jmbi.1997.0926
32. Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) Protein Sci 12:2001–2014. doi:10.1110/ps.03154503
33. Vriend G, Sander C (1993) J Appl Cryst 26:47–60. doi:10.1107/S0021889892008240
34. Vriend G (1990) J Mol Graph 8:52–56. doi:10.1016/0263-7855(90)80070-V
35. Wallner B, Elofsson A (2003) Protein Sci 12:1073–1086. doi:10.1110/ps.0236803
36. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM (2008) BMC Bioinformatics 9:403
37. Fraczkiewicz R, Braun W (1998) J Comput Chem 19:319–333. doi:10.1002/(SICI)1096-987X(199802)19:3<319::AID-JCC6>3.0.CO;2-W
38. Nicholls A, Sharp KA, Honig B (1991) Proteins 11:281–296. doi:10.1002/prot.340110407
39. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003) Structure 11:1453–1459. doi:10.1016/j.str.2003.10.002
40. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK (2005) Proteins 61(Suppl 7):176–182. doi:10.1002/prot.20735
41. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) BMC Bioinformatics 7:208. doi:10.1186/1471-2105-7-208
42. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL (2005) Bioinformatics 21:3435–3438. doi:10.1093/bioinformatics/bti537

43. Coeytaux K, Poupon A (2005) Bioinformatics 21:1891–1900. doi:10.1093/bioinformatics/bti266
44. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Proteins 42:38–48. doi:10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3
45. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) J Mol Biol 337:635–645. doi:10.1016/j.jmb.2004.02.002
46. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235
47. Sengupta J, Nilsson J, Gursky R, Spahn CM, Nissen P, Frank J (2004) Nat Struct Mol Biol 11:957–962. doi:10.1038/nsmb822
48. Rakotobe D, Violot S, Hong SS, Gouet P, Boulanger P (2008) Virol J 5:32. doi:10.1186/1743-422X-5-32
49. Gaudermann P, Vogl I, Zientz E, Silva FJ, Moya A, Gross R, Dandekar T (2006) BMC Microbiol 6:1. doi:10.1186/1471-2180-6-1
50. Durand A, Villard C, Giardina T, Perrier J, Juge N, Puigserver A (2003) J Protein Chem 22:183–191. doi:10.1023/A:1023431215558
51. Gifford ML, Robertson FC, Soares DC, Ingram GC (2005) Plant Cell 17:1154–1166. doi:10.1105/tpc.104.029975
52. Oxvig C, Springer TA (1998) Proc Natl Acad Sci USA 95:4870–4875. doi:10.1073/pnas.95.9.4870
53. Neer EJ, Smith TF (1996) Cell 84:175–178. doi:10.1016/S0092-8674(00)80969-1
54. Lim D, Park HU, De Castro L, Kang SG, Lee HS, Jensen S, Lee KJ, Strynadka NC (2001) Nat Struct Biol 8:848–852. doi:10.1038/nsb1001-848
55. Stevens TJ, Paoli M (2008) Proteins 70:378–387. doi:10.1002/prot.21521
56. Appleton BA, Wu P, Wiesmann C (2006) Structure 14:87–96. doi:10.1016/j.str.2005.09.013
57. Hadjebi O, Casas-Terradellas E, Garcia-Gonzalo FR, Rosa JL (2008) Biochim Biophys Acta 1783:1467–1479. doi:10.1016/j.bbamcr.2008.03.015
58. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) Nucleic Acids Res 31:365–370. doi:10.1093/nar/gkg095
59. Hadano S, Benn SC, Kakuta S, Otomo A, Sudo K, Kunita R, Suzuki-Utsunomiya K, Mizumura H, Shefner JM, Cox GA, Iwakura Y, Brown RH Jr, Ikeda JE (2006) Hum Mol Genet 15:233–250. doi:10.1093/hmg/ddi440
60. Yamanaka K, Miller TM, McAlonis-Downes M, Chun SJ, Cleveland DW (2006) Ann Neurol 60:95–104. doi:10.1002/ana.20888
61. Chaudhuri I, Soding J, Lupas AN (2008) Proteins 71:795–803. doi:10.1002/prot.21764
62. Hsieh TJ, Farh L, Huang WM, Chan NL (2004) J Biol Chem 279:55587–55593. doi:10.1074/jbc.M408934200
63. Hadano S, Otomo A, Suzuki-Utsunomiya K, Kunita R, Yanagisawa Y, Showguchi-Miyata J, Mizumura H, Ikeda JE (2004) FEBS Lett 575:64–70. doi:10.1016/j.febslet.2004.07.092
64. Juhasz T, Szeltner Z, Fulop V, Polgar L (2005) J Mol Biol 346:907–917. doi:10.1016/j.jmb.2004.12.014
65. Gregory DS, Martin AC, Cheetham JC, Rees AR (1993) Protein Eng 6:29–35. doi:10.1093/protein/6.1.29
66. Lin CT, Lin KL, Yang CH, Chung IF, Huang CD, Yang YS (2005) Int J Neural Syst 15:71–84. doi:10.1142/S0129065705000116
67. Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, Donaldson D, Goto J, O'Regan JP, Deng HX et al (1993) Nature 362:59–62. doi:10.1038/362059a0
68. Hart PJ, Liu H, Pellegrini M, Nersissian AM, Gralla EB, Valentine JS, Eisenberg D (1998) Protein Sci 7:545–555
69. Lyons TJ, Liu H, Goto JJ, Nersissian A, Roe JA, Graden JA, Cafe C, Ellerby LM, Bredesen DE, Gralla EB, Valentine JS (1996) Proc Natl Acad Sci USA 93:12240–12244. doi:10.1073/pnas.93.22.12240
70. James PA, Talbot K (2006) Biochim Biophys Acta 1762:986–1000
71. Kanekura K, Hashimoto Y, Niikura T, Aiso S, Matsuoka M, Nishimoto I (2004) J Biol Chem 279:19247–19256. doi:10.1074/jbc.M313236200
72. Oyanagi K, Kawakami E, Kikuchi-Horie K, Ohara K, Ogata K, Takahama S, Wada M, Kihira T, Yasui M (2006) Neuropathology 26:115–128. doi:10.1111/j.1440-1789.2006.00672.x
73. Divers TJ, Cummings JE, de Lahunta A, Hintz HF, Mohammed HO (2006) Am J Vet Res 67:120–126. doi:10.2460/ajvr.67.1.120
74. Ermilova IP, Ermilov VB, Levy M, Ho E, Pereira C, Beckman JS (2005) Neurosci Lett 379:42–46. doi:10.1016/j.neulet.2004.12.045
75. Kihira T, Yoshida S, Yase Y, Ono S, Kondo T (2002) Neuropathology 22:171–179. doi:10.1046/j.1440-1789.2002.00441.x
76. Kihira T, Yoshida S, Kondo T, Yase Y, Ono S (2004) J Neurol Sci 219:7–14. doi:10.1016/j.jns.2003.11.010
77. Mallick P, Boutz DR, Eisenberg D, Yeates TO (2002) Proc Natl Acad Sci USA 99:9679–9684. doi:10.1073/pnas.142310499
78. Fulop V, Jones DT (1999) Curr Opin Struct Biol 9:715–721. doi:10.1016/S0959-440X(99)00035-4
79. Strange RW, Antonyuk S, Hough MA, Doucette PA, Rodriguez JA, Hart PJ, Hayward LJ, Valentine JS, Hasnain SS (2003) J Mol Biol 328:877–891. doi:10.1016/S0022-2836(03)00355-3
80. Banci L, Bertini I, Cantini F, D'Amelio N, Gaggelli E (2006) J Biol Chem 281:2333–2337. doi:10.1074/jbc.M506497200
81. Eymard-Pierre E, Yamanaka K, Haeussler M, Kress W, Gauthier-Barichard F, Combes P, Cleveland DW, Boespflug-Tanguy O (2006) Ann Neurol 59:976–980. doi:10.1002/ana.20879
82. Hand CK, Devon RS, Gros-Louis F, Rochefort D, Khoris J, Meininger V, Bouchard JP, Camu W, Hayden MR, Rouleau GA (2003) Arch Neurol 60:1768–1771. doi:10.1001/archneur.60.12.1768
83. Crennell S, Garman E, Laver G, Vimr E, Taylor G (1994) Structure 2:535–544. doi:10.1016/S0969-2126(00)00053-8
84. Luo Y, Li SC, Chou MY, Li YT, Luo M (1998) Structure 6:521–530. doi:10.1016/S0969-2126(98)00053-7
85. Springer TA (1997) Proc Natl Acad Sci USA 94:65–72. doi:10.1073/pnas.94.1.65
86. Lu C, Oxvig C, Springer TA (1998) J Biol Chem 273:15138–15147. doi:10.1074/jbc.273.24.15138
87. Aroul-Selvam R, Hubbard T, Sasidharan R (2004) J Mol Biol 338:633–641. doi:10.1016/j.jmb.2004.03.039
88. Selvam RA, Sasidharan R (2004) Nucleic Acids Res 32:D193–D195. doi:10.1093/nar/gkh047
89. Saini HK, Fischer D (2005) Bioinformatics 21:2917–2920. doi:10.1093/bioinformatics/bti445
90. Fink AL (2005) Curr Opin Struct Biol 15:35–41. doi:10.1016/j.sbi.2005.01.002
91. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) Nucleic Acids Res 32:1037–1049. doi:10.1093/nar/gkh253
92. Karlin D, Longhi S, Receveur V, Canard B (2002) Virology 296:251–262. doi:10.1006/viro.2001.1296