

# Detection of Alpha-Rod Protein Repeats Using a Neural Network and Application to Huntingtin

Gareth A. Palidwor<sup>1</sup>, Sergey Shcherbinin<sup>2</sup>, Matthew R. Huska<sup>3</sup>, Tamas Rasko<sup>3</sup>, Ulrich Stelzl<sup>4</sup>, Anup Arumughan<sup>3</sup>, Raphaelae Foulle<sup>3</sup>, Pablo Porras<sup>3</sup>, Luis Sanchez-Pulido<sup>5</sup>, Erich E. Wanker<sup>3</sup>, Miguel A. Andrade-Navarro<sup>1,3\*</sup>

**1** Ottawa Health Research Institute, Ottawa, Ontario, Canada, **2** Medical Imaging Research Group, The University of British Columbia, Vancouver General Hospital, Vancouver, British Columbia, Canada, **3** Max-Delbrück Center for Molecular Medicine, Berlin, Germany, **4** Otto-Warburg Laboratory, Max Planck Institute for Molecular Genetics, Berlin, Germany, **5** Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom

## Abstract

A growing number of solved protein structures display an elongated structural domain, denoted here as alpha-rod, composed of stacked pairs of anti-parallel alpha-helices. Alpha-rods are flexible and expose a large surface, which makes them suitable for protein interaction. Although most likely originating by tandem duplication of a two-helix unit, their detection using sequence similarity between repeats is poor. Here, we show that alpha-rod repeats can be detected using a neural network. The network detects more repeats than are identified by domain databases using multiple profiles, with a low level of false positives (<10%). We identify alpha-rod repeats in approximately 0.4% of proteins in eukaryotic genomes. We then investigate the results for all human proteins, identifying alpha-rod repeats for the first time in six protein families, including proteins STAG1-3, SERAC1, and PSMD1-2 & 5. We also characterize a short version of these repeats in eight protein families of Archaeal, Bacterial, and Fungal species. Finally, we demonstrate the utility of these predictions in directing experimental work to demarcate three alpha-rods in huntingtin, a protein mutated in Huntington's disease. Using yeast two hybrid analysis and an immunoprecipitation technique, we show that the huntingtin fragments containing alpha-rods associate with each other. This is the first definition of domains in huntingtin and the first validation of predicted interactions between fragments of huntingtin, which sets up directions toward functional characterization of this protein. An implementation of the repeat detection algorithm is available as a Web server with a simple graphical output: <http://www.ogic.ca/projects/ard>. This can be further visualized using BiasViz, a graphic tool for representation of multiple sequence alignments.

**Citation:** Palidwor GA, Shcherbinin S, Huska MR, Rasko T, Stelzl U, et al. (2009) Detection of Alpha-Rod Protein Repeats Using a Neural Network and Application to Huntingtin. *PLoS Comput Biol* 5(3): e1000304. doi:10.1371/journal.pcbi.1000304

**Editor:** Philip E. Bourne, University of California San Diego, United States of America

**Received:** November 5, 2008; **Accepted:** January 27, 2009; **Published:** March 13, 2009

**Copyright:** © 2009 Palidwor et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by the Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF) and the Canada Research Chairs (MAA-N). LS-P is supported by an EMBO Long Term Fellowship (ALT 325-2008).

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [Miguel.andrade@mdc-berlin.de](mailto:Miguel.andrade@mdc-berlin.de)

## Introduction

Tandems of repeated protein sequences forming structural domains occur in at least 3% of proteins in eukaryotic organisms [1]. Characterization of these repeats by sequence similarity is sometimes difficult as weak evolutionary constraints cause rapid sequence divergence [2]. In particular, repeats including two alpha helices packed together then stacked to form a flexible rod (denoted here alpha-rod) belong to this category (see an example in Figure 1).

Some of these alpha-rod repeats have been defined in terms of sequence similarity and are widespread in multiple protein families: HEAT [3,4], Armadillo [5] and HAT [6]. Others are evident in just one protein family, for example the PFTA repeats [7]. Some, however, bear no statistically significant sequence similarity and may not have originated from sequence duplication (for example, the all-helical VHS domain in *Drosophila melanogaster* Hrs protein [8], or the subunit H of *Saccharomyces cerevisiae* vacuolar ATP synthase [9]).

This divergence complicates the detection of alpha-rod repeats by methods based on sequence similarity. For example, profile-

based methods used in the protein domain databases PFAM [10] and SMART [11] detect only two of the 14 HEAT repeats of human AP-2 complex subunit beta-1 (Figure 1), and might fail to detect any repeats in other alpha-rod containing sequences.

Despite the heterogeneity of alpha-rod repeats, they have common features (discussed in [4]): length of about 40 amino acids, anti-parallel alpha-helices, and constraints given by the packing of consecutive repeats. This suggests that alpha-rod repeats are a protein structural feature that obeys some physical constraints irrespective of their evolutionary origin and particular sequence. Coiled coils and transmembrane alpha-helices are other examples of such structural features. Statistical methods have been used to predict coiled coils [12] and transmembrane alpha-helices [13] with excellent reliability, using algorithms that learn to recognize these features from amino acid sequences. In particular, back-propagation neural networks [14] have been used with success to predict secondary structure [15,16], transmembrane alpha-helices [17], and protein residue solvent accessibility [18].

We hypothesized that a back-propagation neural network could be better suited than homology based methods for the detection of

## Author Summary

Many proteins have an elongated structural domain formed by a stack of alpha helices (alpha-rod), often found to interact with other proteins. The identification of an alpha-rod in a protein can therefore tell something about both the function and the structure of that protein. Though alpha-rods can be readily identified from the structure of proteins, for the vast majority of known proteins this is unavailable, and we have to use their amino acid sequence. Because alpha-rods have highly variable sequences, commonly used methods of domain identification by sequence similarity have difficulty detecting them. However, alpha-rods do have specific patterns of amino acid properties along their sequences, so we used a computational method based on a neural network to learn these patterns. We illustrate how this method finds novel instances of the domain in proteins from a wide range of organisms. We performed detailed analysis of huntingtin, the protein mutated in Huntington's chorea, a neurodegenerative disease. The function of huntingtin remains a mystery partially due to the lack of knowledge about its structure. Therefore, we defined three alpha-rods in this protein and experimentally verified how they interact with each other, a novel result that opens new avenues for huntingtin research.

different types of alpha-rod repeats, if trained in an appropriate set of sequences containing these repeats. The last ten years have seen the resolution of a sufficient number of protein 3D structures of sequences with alpha-rod repeats to provide a useful training set for such predictions.

## Results

We manually compiled a set of protein sequences with known structures reported to contain structural repetitions forming an alpha-rod composed of stacked repeats (see supplementary Table S1 in Text S1, positives). To reduce redundancy, no two sequences with more than 70% identity were included in the set (after verifying that they were full length homologs). We included one protein from each of three HEAT repeat types [4], two armadillo repeat proteins, and five other unrelated proteins. A similar sized set of sequences adopting a variety of structures but without alpha-rod repeats was compiled as a negative set (Table S1 in Text S1, negatives).

The input window of the neural network was chosen to be 39 amino acids, which is close to the average repeat length. Since these repeats are characterized by two helices of similar size, we chose as the central defining feature the middle residue in the hinge between the two helices. This residue should be equidistant from two secondary structure elements with particular packing features, likely presenting a periodicity of small and hydrophobic residues constrained by the intra-repeat interactions between the two helices and the inter-repeat interactions with the stack of consecutive repeats [4]. Therefore, the network was trained to detect the central residue of the hinge (see Methods). The file with the annotated sequences used for the training is provided as supplementary Dataset S1.

## Analysis of Proteins of Known Structure

The parameters of the method were optimized using the analysis of proteins of known structure. We found that hits above a score of 0.8 were reliable, especially when the protein had several

of them in the appropriate periodicity. Identification of a sequence as containing an alpha-rod was optimal when requiring at least three hits above a score of 0.8 with a minimum spacing of 30 amino acids between hits and a maximum of 135. Further details can be found in the supplementary Text S1.

A total of 87 sequences were selected with this threshold, which can be grouped in 12 protein families of which 8 were not homologous to those used in the training set (Table S2 in Text S1). Since these examples correspond to proteins of known structure, it was easy to visually verify that of those eight families seven were true positives and only one constituted a false positive. Homology of these proteins to the ones used in the training is extremely low or statistically non-significant. Therefore, we concluded that the network was useful in expanding our current knowledge of the occurrences of these repeats and we set to demonstrate this. For simplicity we will denote our methodology as ARD (Alpha-rod Repeat Detection) henceforth.

## Analysis of Complete Genomes

To illustrate the coverage of the method we analyzed the complete protein sets from a series of fully sequenced organisms. The threshold tested in the analysis of PDB was used to select positive sequences. The results of the analysis are in Table 1. The fractions of alpha-rod repeat proteins are around 0.4% for the nine eukaryotic genomes and lower (0.05%–0.21%) in the three prokaryotic organisms tested. No correlation was found between proteome size and fraction of positives.

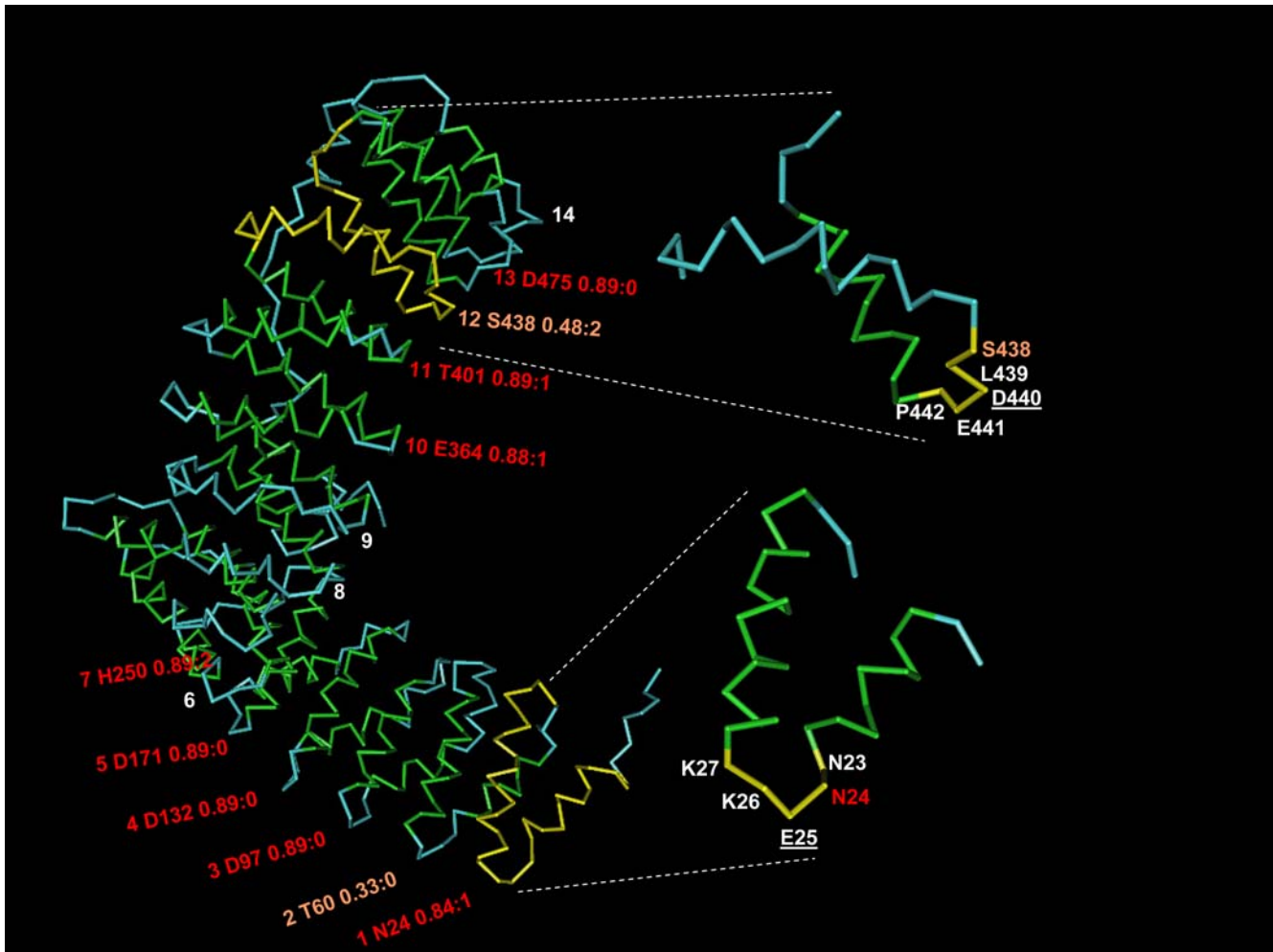
Using ARD we were able to detect protein sequences that PFAM [10] and SMART [11] do not detect or that they detect with multiple profiles (PFAM: Arm, HEAT\_PBS and HEAT; SMART: ARM, EZ\_HEAT and HEAT). Many of these were not described in the literature.

To illustrate the ability of ARD to identify new results we will focus on families with at least one human gene. To illustrate how the method covers various profiles used by SMART and PFAM we will examine results on families with HEAT repeats of the PBS type from fungi, bacteria, and archaea. Finally, we illustrate an experimental application of the method to dissect domains in huntingtin, the protein mutated in Huntington's disease, for which little is known regarding its structure and function.

## Survey of Human Genes

A total of 86 human proteins were found to contain alpha-rod repeats, which we grouped in 52 families on the basis of their sequence similarity. Of those families, at least 16 have not been yet described to contain alpha-rod repeats in the literature, with 9 undetected by both the SMART and PFAM domain detection web tools (see Table 2).

In particular, six families have neither literature nor database repeat assignment; for these, we could verify the repeats using a manually tuned iterative PSIBLAST sequence search [19] of the region with repeats, which showed significant similarity to alpha-rod repeat regions in other protein families. Four of these families encode proteins of unknown function: Serac1, C8orf73, C17orf66, and KIAA0423 (and homolog LOC23116). A fifth family has three members in humans, the stromal antigens 1, 2 and 3 (STAG1-3), subunits of the cohesin complex, which mediates cohesion between sister chromatids [20]. In particular, the phosphorylation of STAG2 is essential for cohesin dissociation during prophase and prometaphase [21]. This family has two homologs in *Xenopus* (demonstrated to form part of two different cohesion complexes [22]), the plant *Arabidopsis thaliana* (Scc3, needed for the orientation of the kinetochores during meiosis [23]) and yeast (Irr1/Scc3, involved in cell wall integrity [24]). The



**Figure 1. Detection of repeats in an alpha-rod protein.** Structure (alpha-backbone trace) of the 591 aa N-terminal fragment of human adaptor-related protein complex 2, beta 1 subunit, as forming part of the AP2 clathrin adaptor core [69] (PDB code 2VGL chain B). Green and blue represent residues in alpha-helix and in disordered conformation, respectively. This structure has no residue in beta-strand conformation and is entirely composed of an alpha-rod of 14 repeats previously classified as HEAT repeats of type ADB [4]. The label for each repeat indicates the following: repeat order, residue detected by the network, score of hit, and position relative to residue used for training. For example, “1 N24 0.84:1” indicates that the residue detected for repeat #1 was N (amino acid code for asparagine) in position 24 of the sequence, with score 0.84, but that the residue in relative position 1 (that is, at 25) was the one used to train the network as being in the hinge. Ten out of the 14 repeats were detected, 8 of them with score  $\geq 0.80$ . The inset shows repeats 12 (right, top) and 1 (right, bottom) with the residue used as positive in the training underscored. A coloured label indicates the residue identified by the network after training, which in both cases is not the one given in the training but others belonging to the hinge (E25 and S438). The figure was generated using NCBI’s linked viewer, Cn3D [70]. doi:10.1371/journal.pcbi.1000304.g001

analysis of the family suggests that their sequences are composed of alpha-rod repeats (Figure 2 and Figure S3A in Text S1).

The sixth novel assignment case is the PSMD family (proteasome 26S subunit, non-ATPase) members 1, and 2, and 5. PFAM/SMART identify these as containing repeats of the Proteasome/cyclosome (PC\_rep), originally predicted to be composed of a beta strand and an alpha helix [25]. However, ARD predicts 5 repeats which overlap with those. Secondary structure predictions (using JPRED3 [26]) and homology to alpha-rod repeats proposed for PSMD1 yeast homolog Sen3/RPN2 [27] clearly suggest that these are alpha-rod repeats, and that the current PC\_rep motif used by PFAM/SMART cuts one of the helices in half. This suggests that the PFAM/SMART domain definition should be revised.

Another family for which a redefinition of the PFAM/SMART profile may be required is RRP12, homolog to the yeast

Ribosomal RNA processing 12, identified as HEAT-repeat containing, Ran binding, and required for the nuclear export of both the 40S and 60S ribosomal subunits in yeast [28]. SMART and PFAM identify only one HEAT repeat in the human sequence because other repeats overlap with domain NUC173, defined as present in several nucleolar proteins [29], whereas ARD identifies 9 repeats.

Three other families remain undetected by PFAM and SMART profiles but have been described to contain alpha-rod repeats in separate publications: these are the MRO (Maestro), which expresses a nucleolar protein of unknown function during male mouse gonad development [30], FRAP1/mTOR, which we described as repeat containing in the first publication defining the HEAT repeats [3] (Figure 2 and Figure S3B in Text S1), and NIPBL (the homolog to *Drosophila* Nipped-B) related to sister chromatid cohesion yeast proteins Scc2 and Mist4 [31].

**Table 1.** Results of predictions in complete genomes.

Organism	Proteins	Hits <sup>1</sup>	Genes	Fraction
<i>Homo sapiens</i>	43797	159	86	0.36%
<i>Mus musculus</i>	32241	125	93	0.39%
<i>Monodelphis domestica</i>	32685	131	81	0.40%
<i>Gallus gallus</i>	22250	102	75	0.46%
<i>Xenopus tropicalis</i>	28324	96	69	0.34%
<i>Danio rerio</i>	36078	116	85	0.32%
<i>Drosophila melanogaster</i>	19789	52	41	0.26%
<i>Saccharomyces cerevisiae</i>	6697	23	23	0.34%
<i>Gibberella zeae</i>	11640	37	37	0.32%
<i>Escherichia coli</i>	4133	2	2	0.05%
<i>Anabaena variabilis ATCC 29413</i>	5634	11	11	0.20%
<i>Methanosarcina mazei</i>	3303	7	7	0.21%

<sup>1</sup>At least three matches with score >= 0.8 and with >= 30 aa spacing.  
doi:10.1371/journal.pcbi.1000304.t001

For ten other gene families, PFAM and SMART suggest the presence of the repeats but their coverage is more limited than that of ARD and this evidence remains unreported in the literature. This is the case of STK36/FU (the homolog to *Drosophila* fused, a mediator of sensitivity to PARP [32]), INTS4 (integrator complex subunit 4, which associates with the C-terminal domain of RNA polymerase II large subunit [33]), and of eight hypothetical proteins: C1orf175, LOC165186, HEATR2, HEATR4, HEATR6, KIAA1468, RTDR1 (deleted in rhabdoid tumour), and TMCO7 (which interacts with MACF1, the microtubule-actin crosslinking factor 1 according to a two-hybrid screening [34]).

The combination of ARD analyses of the human protein homologs in other organisms, secondary structure prediction and definition of regions of amino acid composition bias facilitates the definition of the boundaries of domains composed of repeats sometimes reused in different domain architectures. Here we present three examples.

We found that the LOC165186 and KIAA0423 hypothetical human proteins (mentioned above) define two families whose structured sequence is likely alpha-rods; these two proteins share a C-terminal domain possibly made of more than 10 repeats (Figure 2 and Figure S3C in Text S1). LOC165186, conserved in mammals, has an additional N-terminal composition biased region of around 500 amino acids, whereas KIAA0423, conserved down

**Table 2.** Selected predictions for human genes<sup>a</sup>.

	Representative	Description	A <sup>b</sup>	S	P	R	Ref <sup>c</sup>	Related <sup>d</sup>
Novel	STAG1	Homologs of yeast subunit of the cohesin complex 3 (Scc3/IRR1)	4	0	0	0	u	STAG2 STAG3
	SERAC1	serine active site containing 1	3	0	0	0	u	
	C8orf73		5	0	0	0	u	
	C17orf66		6	0	0	0	u	
	KIAA0423	LOC23116	11	0	0	5	u	
	PSMD1	proteasome 26S subunit, non-ATPase family	5	0	0	0	u	PSMD2 PSMD5
No PFAM/SMART	MRO	Maestro	3	0	0	0	[30]	NP_775760.2 Q8NDA8_HUMAN Q8ND95_HUMAN
	NIPBL	Nipped-B homolog ( <i>Drosophila</i> )	7	0	0	5	[31]	
	FRAP1	FRAP1/mTOR	16	0	0	11	[3]	
Domain redefinition	RRP12	Ribosomal RNA processing 12 homolog (yeast)	9	1	1	0	[28]	
	CLASP1	CLASP family	10	4	4	7	[38]	CLASP2
	CKAP5	CKAP5	18	3	3	10	[31]	
Not reported in literature	KIAA1468		6	0	3	0	u	
	HEATR2		9	6	6	11	u	
	HEATR4		9	0	3	0	u	
	HEATR6		7	2	2	4	u	
	TMCO7		7	0	2	0	u	
	STK36	Serine/threonine kinase 36, fused homolog ( <i>Drosophila</i> )	7	3	3	0	u	
	INTS4	integrator complex subunit 4	8	5	5	5	u	Q96LV5_HUMAN
	RTDR1	Rhabdoid tumor deletion region protein 1	6	0	1*	0	u	
	LOC165186		7	1	1	0	u	
	C1orf175		6	0	1	0	u	

<sup>a</sup>Hits not included: reported in the literature (SF3B1, MMS19, huntingtin, PSME4, NCAPD3, NCAPG2, TBCD, BTA1F1, KOG1, PDS5B); armadillo repeats (JUP, RAP1GDS1); likely false positives (OBSCN, P2RY8, PACS2); other genes homologous to 3D structures discussed before.

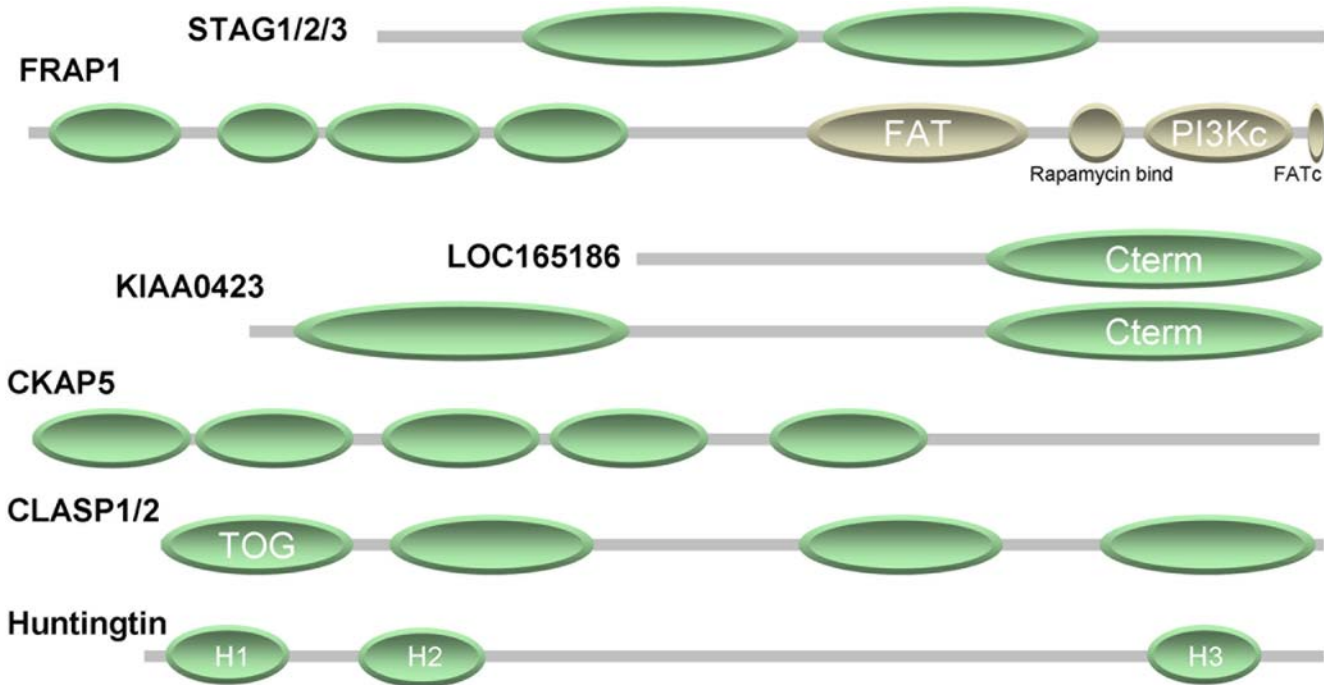
<sup>b</sup>Predicted number of repeats by ARD (A), SMART (S), PFAM (P), REP (R).

<sup>c</sup>Reference column. "u" indicates unknown from the point of view of the literature.

<sup>d</sup>Close homologous genes are indicated.

\*Armadillo repeat.

doi:10.1371/journal.pcbi.1000304.t002



**Figure 2. Selected human protein families with alpha-rod repeats.** The cartoon summarizes the findings for seven human proteins. The green ellipses represent regions of alpha-rod repeats as deduced by a combination of our method, analysis of homologs, and iterative sequence analysis. Further details for each case, including an overview of repeat predictions and regions with amino acid bias overlaid to the multiple sequence alignment of the family using an update of the BiasViz software [71] are available as supplementary Figure S3 in Text S1. doi:10.1371/journal.pcbi.1000304.g002

to worms, has an extra N-terminal domain of alpha-rod repeats connected to the C-terminal repeat domain by a middle linker that is enlarged in the chordate sequences.

Human CKAP5/TOG (cytoskeleton associated protein 5), a component of the centrosome that is required for spindle pole assembly [35], has similar-length homologs in mammals, frog, and fly. Analysis of the family identifies five alpha-rods of six repeats each in these sequences and a C-terminal non-repeat containing domain (Figure 2 and Figure S3D in Text S1). The worm homologs are shorter since they have only three of the repeat domains. The structure of one of those domains in *Caenorhabditis elegans* zyg9 was solved and confirmed the presence of an alpha-rod of six repeats [36].

The CLASP family proteins are microtubule-associated proteins, conserved in animals, fungi, and plants [37]. In humans, there are two homologs, hCLASP1 and hCLASP2, which, similar to CKAP5, associate with the ends of growing microtubules to participate in mitotic spindle formation [38]. Their multiple sequence alignment with homologs suggests that they are formed by four alpha-rods (Figure 2 and Figure S3E in Text S1), also noted in [38].

Other genes previously identified in the literature and by SMART/PFAM are: TBCD (tubulin folding cofactor D) reported by [31]; PSME4/PA200, identified as containing 18 HEAT-like repeats in [39]; BTAF1 (RNA polymerase II, B-TFIID transcription factor-associated, 170 kDa) whose homolog in yeast, Mot1, was noted by [31]; MMS19, involved in nucleotide excision repair and transcription, noted by [40]; huntingtin [3]; both subunits of non-SMC condensin II complex D3 and G2, noted by [31]; and PDS5B/APRIN, a chromatin regulator in hormonal differentiation [41], whose homolog Spo76 in *Sordaria macrospore* was noted by [31].

The existence of two cases where the evidence of repeats originates from low resolution electron microscopy images deserves special mention. SF3B1 (splicing factor 3b, subunit 1) is

proposed to have 22 repeats according to the structure obtained by single-particle electron cryomicroscopy at a resolution of less than 10 angstroms of its complex with splicing factor 3a (SF3B14/P14) where it is shown to coil around SF3B14 [42]. The low resolution electron microscopy structure of the yeast complex of mTOR with KOG1 suggests that KOG1 has a middle alpha-rod domain [41]. We can confirm through ARD analysis that both SF3B1 and KOG1 have alpha-rods in the regions suggested.

As noted in the section on analysis of PDB, armadillo repeats are not well detected by ARD and generally PFAM and SMART are as good or better than ARD in recognizing them (for example, for JUP and ARMC8). However, two genes are detected by ARD that are covered by one single PFAM armadillo match and no SMART matches: these are HSPBP1 (hsp70-interacting protein) whose solved 3D structure indicates four armadillo repeats [43] and newly identified RTRD1, for which we detect 3 and 6 repeats, respectively.

Finally, of all 52 protein families with human genes we recognized just three false positives: PACS2 (phosphofurin acidic cluster sorting protein 2), OBSCN (obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF), and P2RY9 (purinergic receptor P2Y, G-protein coupled, 8). This was determined by lack of further evidence (no homology to regions with repeats in other families, incompatible secondary structure predictions) combined with a small number of hits in the human sequence, in homologs in other species, or by the overlap of those hits with other domains.

### Short Repeats Highly Identical within Protein Sequences

In the results of fungal and prokaryotic sequences, we noted a number of cases where the repeats identified for the sequences selected were so similar that it was possible to align most of the repeats

by hand in stark contrast to the very divergent examples noted above. We illustrate these with 8 examples, which are not related by homology (see Table S3 in Text S1). Their high percentage of inter-repeat sequence identity is indicative of very recent events of duplication occurring independently in these eight examples. Secondary structure prediction suggests that the structure of the repeat is composed of two helices of  $\sim 10$  residues, with a middle loop of three, and an outer loop of  $\sim 10$  residues, for a total length of 31–35 aa.

Although most of the repeats were identified by SMART and PFAM (EZ\_HEAT and HEAT\_PBS profiles, respectively), not all repeat instances were marked and some were detected with the alternative HEAT profile. In contrast, ARD identified all obvious repetitions and some additional borderline ones.

Orthologs of these eight examples were identified in related taxa (Table S3 in Text S1). The puzzling question remains of why or how these eight apparently unrelated families arose and converged to these short alpha-rod repeats. Whether there are common mechanisms for the duplication and selection of these repeats and for their functions is, at the moment, unclear.

### Dissecting Huntingtin

The human protein huntingtin is involved in Huntington's disease. Its function remains unclear [44]. In 1995 we described that huntingtin contains HEAT repeats [3] but their identification was restricted to 10 units covering  $\sim 400$  scattered amino acids out of a total sequence length of 3144 amino acids. Since then, no other characteristic structural features have been described for this protein, which complicates its description in terms of separate domains with independent folds and functions. As a result no 3D structure of any fragment of this protein has been yet solved, and although interacting partners of this protein have been found they are mostly restricted to the N-terminal 500 amino acids of the protein [45]. Here, we applied the methodology described above to define alpha-rods in huntingtin and subsequently tested the validity of our predictions experimentally.

Initially, we produced an alignment of human huntingtin with a representative set of homologous sequences from the database (provided as supplementary Dataset S2). For this we used not only sequences from protein databases but also sequences derived from ESTs and from genomic fragments. We identified for the first time the existence of huntingtin homologs in worms (nematoda genus *Caenorhabditis*, and annelida *Capitella sp.*), amoebae (*Naegleria fowleri* and *Dictyostelium discoideum*), sea anemone *Nematostella vectensis*, and choanoflagellate *Monosiga brevicollis*, notably expanding the scope of this family. We did not find homologs of huntingtin in fungi.

The analysis of human huntingtin by ARD suggests six matches but other low scoring hits are consistently present in homologs. Comparison to biased regions sharply defines two N-terminal domains of six and seven repeats (H1 from amino acid 114 to 413 and H2 from 672 to 969) and suggests the existence of a C-terminal domain of seven repeats (H3 from 2667 to 2938) (Figure 2 and Figure S3F in Text S1). Iterative sequence searches using PSIBLAST with these regions indicated homology to HEAT repeats in otherwise unrelated proteins in the 2<sup>nd</sup> or 3<sup>rd</sup> iterations. Consistently, sequence analysis suggested a HEAT-repeat fold (using SVMfold [46]), and threading suggested that those regions adopt a HEAT-repeat fold with high likelihood (using GenTHREADER [47]). The comparative protein structure modeling tool TASSER-Lite [48] produced an alpha-rod for H1 and H2, but an alpha-beta barrel for H3 (incompatible with the predicted secondary structure of the region using JPRED3 [26]). Given secondary structure predictions and scattered matches it is tempting to speculate that other alpha-rods exist outside of the H1, H2, and H3 domains. However, we were

unable to obtain consistent results using PSIBLAST or threading for fragments outside these regions.

To test our predictions, we produced huntingtin fragments spanning the complete sequence of the protein but separating the predicted alpha-rods into different fragments (Figure 3A) in order to study intra-molecular domain interactions in huntingtin by yeast two hybrid (Y2H) assays (see Methods). Our rationale is that only well defined domains will fold and produce interactions, whereas wrongly defined domains will either not interact or produce nonspecific interactions.

We found that the huntingtin fragment Htt507-1230 with the H2 domain self-associates in the Y2H assays. In addition, interactions between Htt507-1230 and Htt1-506Q23 (H1 domain) as well as with the fragment Htt2721-3144 (H3 domain) were observed (Figure 3B). No other interactions were observed.

The results obtained with the Y2H assays were also confirmed in mammalian cells using a modified version of the LUMIER method (luminescence-based mammalian interactome mapping technology, [49]). Protein A (PA)-Renilla luciferase- and Firefly-V5 luciferase (Luc)-tagged huntingtin fusion proteins were co-expressed in HEK293 cells and were assessed for the expression of the fusion proteins by immunoblotting and luciferase assays (Figure 3C and 3D). The PA-Renilla-tagged fusion protein is then immunoprecipitated from the soluble cell extracts with IgG coated Dynal magnetic beads. After washing, binding of the Firefly-V5 Luc-tagged fusion protein is quantified by measuring the firefly luciferase activity in a luminescence plate reader. As shown in Figure 3D, interactions between the huntingtin fragments Htt1-506Q23 and Htt507-1230, Htt507-1230 and Htt507-1230, Htt507-1230 and Htt2721-3144 were observed with the assays.

Taken together, these experimental results give the first evidence of domains in huntingtin that mediate potential intras as well as inter-molecular huntingtin interactions. One of many plausible structural assemblies of huntingtin's domains that are consistent with our results and with those in the literature is discussed in Figure 4.

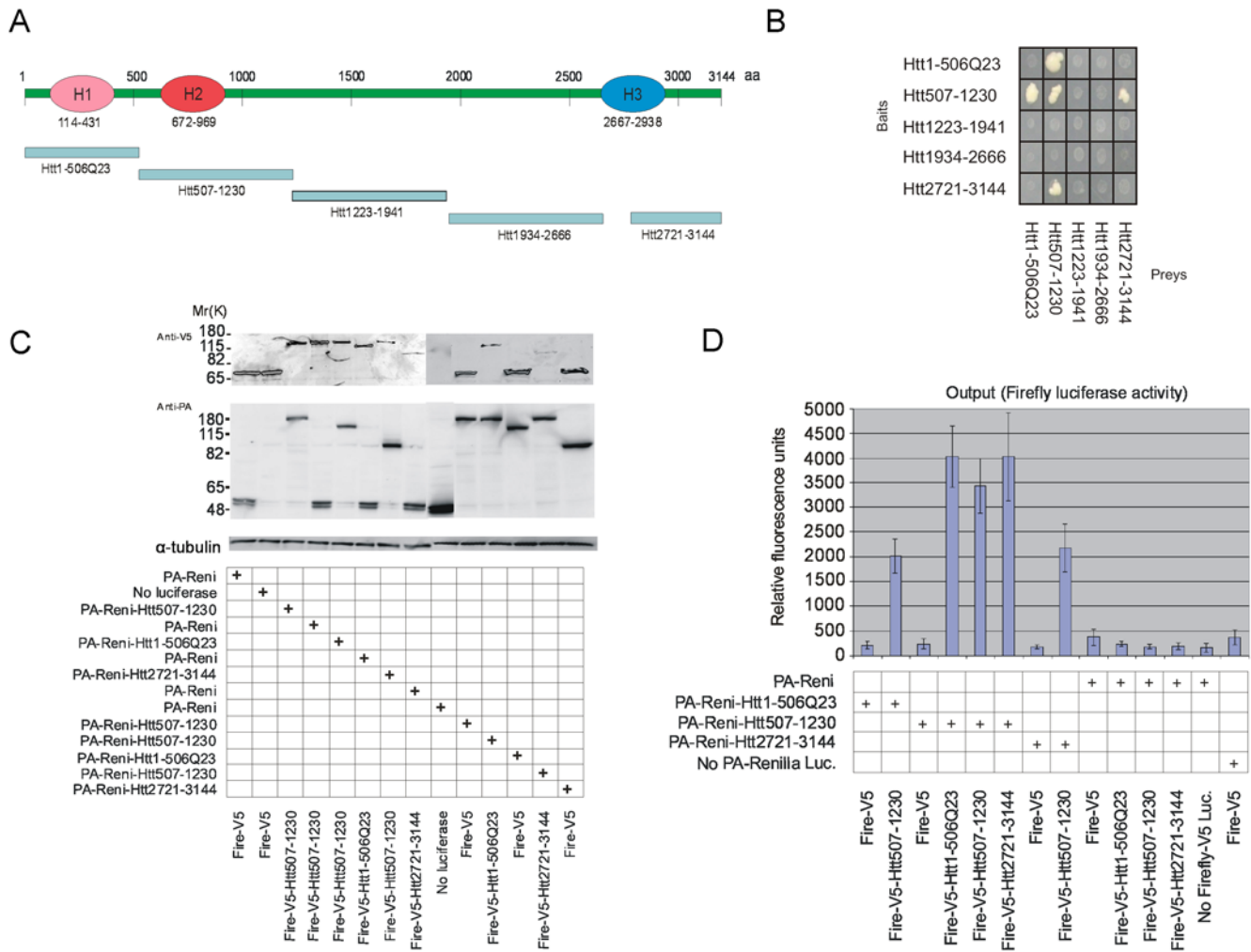
## Discussion

### Performance of the Method

We have developed and applied a neural network for the prediction of alpha-rod repeats. Analysis of the results suggests that it discovers more repeat-containing proteins and repeats per protein than sequence similarity based methods using manually curated profiles, which were previously the best method to detect these repeats. We estimate a level of false positives below 10%: 1 in 12 families in the analysis of PDB (approximately 8%), 3 in 52 families in the analysis of human genes (below 6%). The level of false negatives could be eventually reduced by expanding the training set after new structures of sequences with alpha-rod repeats are solved, but one must be cautious about this to avoid over-prediction. Here, we preferred to train the neural network with a conservative set of known structures to demonstrate that they allow detection of recently identified cases.

We consider it very encouraging that the network learned from a small number of examples and generalized to recognize repeats not used in the training, e.g. the shorter PBS lyase repeats, or those found for the first time in six human protein families. Most of the repeats detected correspond to HEAT, PBS, and Armadillo.

Whereas the network effectively detected a number of unrelated alpha-rod repeat types, it failed to detect the HAT repeats [6]. Although their length is similar, their structural arrangement in highly parallel helices [50] and the conservation of aromatic residues



**Figure 3. Study of interactions between fragments of huntingtin.** (A) Schematic overview of huntingtin fragments used in Y2H and LUMIER experiments. (B) The results obtained with the Y2H assays. (C) The expression of different fusion pairs was analyzed by Western blot using antibodies against V5-epitope (Invitrogen, 1:5000, monoclonal antibody) and Protein-A (Sigma 1:2000, polyclonal antibody); 15  $\mu$ l from 100  $\mu$ l of each cell extract was loaded onto SDS-PAGE gel. Detection with anti-tubulin antibodies was used as a loading control. (D) Firefly luciferase activities of immunopurified protein complexes in relative fluorescence units (RFU). doi:10.1371/journal.pcbi.1000304.g003

[51] make them significantly different from HEAT and Armadillo repeats explaining why they cannot be detected by our method.

The performance of PFAM, SMART and ARD in predicting each type of alpha-rod repeats in sequences deposited in the PDB database is summarized in Table 3. ARD outperforms PFAM and SMART in the detection of HEAT and PBS repeats but underperforms in the detection of Armadillo repeats (although it identifies some proteins with Armadillo repeats that escape detection by both PFAM and SMART, see Table S2 in Text S1). The proteins in PDB that are currently annotated with HAT repeat regions are detected exclusively by SMART.

### Evolutionary and Structural Implications

The lack of a common evolutionary origin for all repeats forming alpha-rods indicates that some specific constraints drive convergent evolution to repeatedly rediscover these repeats as a common solution to a general functional need: protein-protein interactions. Structures of alpha-rods suggest that they are extremely flexible and this allows the ensemble to coil around their target as a boa constrictor would do with its prey. A good example is given by the

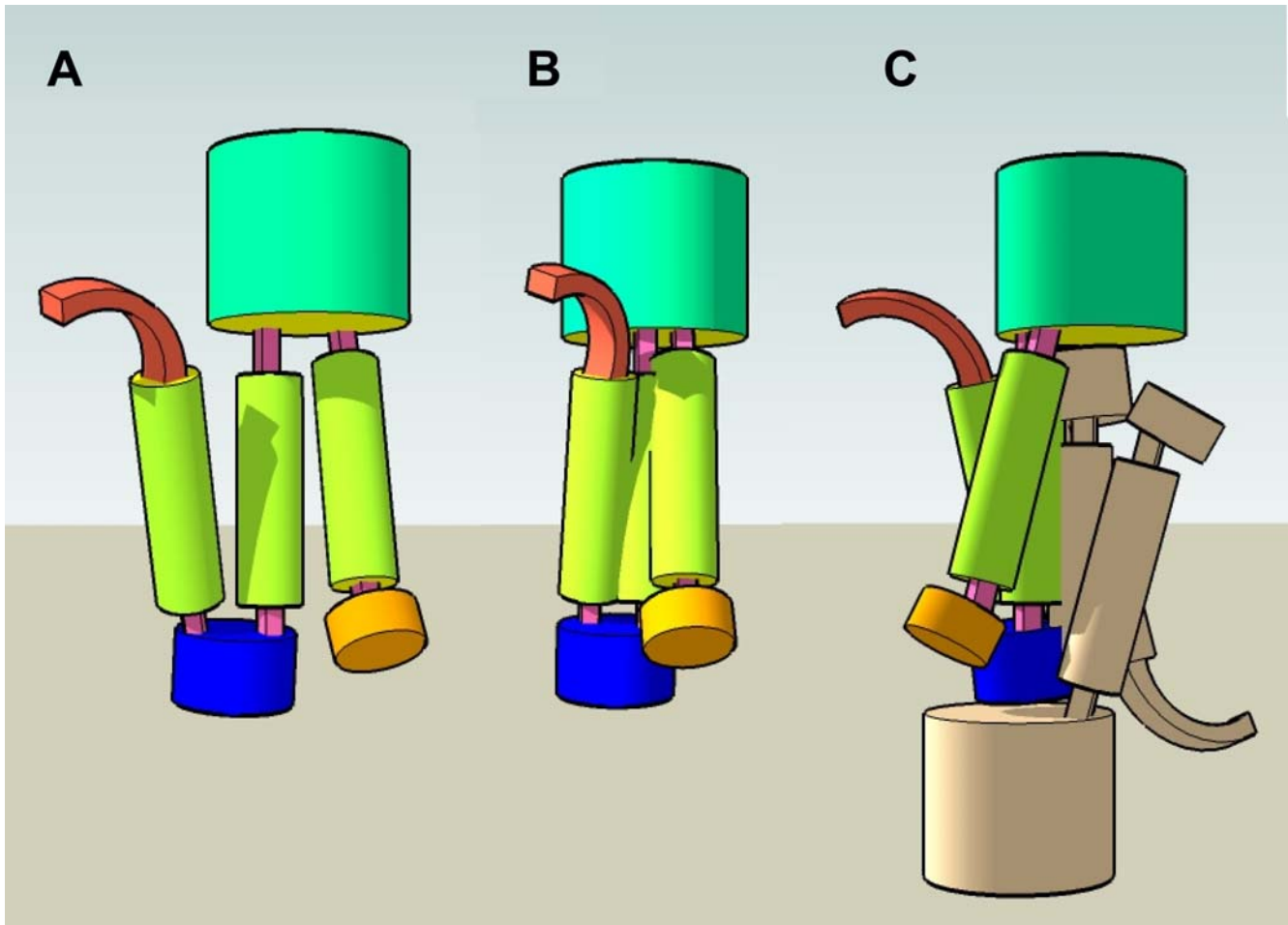
structure of Exportin Cse1p in complex with Kap60p and RanGTP, where both Cse1p and Kap60p are alpha-rods which wrap around each other, and Cse1p wraps around RanGTP [52].

The necessity to coil around proteins possibly explains why the length of these repeats varies between 30 and 45 amino acids. Shorter repeats might not produce enough interactions between the units to form the rod; consequently the rod would not be stable enough and would unfold too easily. Longer repeats might not produce a rod flexible enough to coil around typical protein targets of diameters in the range of 30 to 50 angstroms.

The current data from protein structures and the predictions of protein domains for proteins with alpha-rods (See Table S2 in Text S1) does not suggest the co-occurrence of alpha-rods with other protein domains. We think that this constitutes further evidence that alpha-rods can be used pretty much to bind any protein as needed.

### Functions of Proteins with Alpha-Rods

Neuwald and Hirano identified in [31] several novel HEAT-repeat containing proteins with functions related to chromosomal



**Figure 4. Hypothetical 3D structure of huntingtin.** The cartoon represents a hypothetical model of huntingtin interactions consistent with our results. (a) The N-terminus with the poly-Q tail (red arch) is followed by the H1 alpha-rod domain (residues 114 to 431, yellow cylinder), a small domain (432 to 671, blue), the H2 alpha-rod domain (672 to 969, yellow), a large domain (970 to 2666, green), the H3 alpha-rod domain (2667 to 2938), and a small C-terminal domain (2939–3144). (b) The three rods could assemble by coiling anti-parallel to each other with H2 in the middle: that would explain the interactions between H1 and H2, and between H2 and H3. (c) Formation of a huntingtin homodimer [66] with a second molecule of huntingtin (gray) could happen through their H2 domains. The N-terminal poly-Q tail and the H1 domain remain exposed and can interact with other proteins, as previously reported [45]. The figure was produced with Google SketchUp.  
doi:10.1371/journal.pcbi.1000304.g004

organization and microtubule interaction. In agreement with this, here we have identified many alpha-rod repeat containing sequences with related functions, notably direct tubulin binding.

**Table 3.** Evaluation of the predictions of PFAM, SMART and ARD, for all proteins in the PDB with four types of alpha-rod repeats.

	PFAM <sup>1</sup>	%	SMART <sup>2</sup>	%	ARD	%	Total
HEAT	21	36	0	0	58	100	58
PBS	1	25	1	25	3	75	4
Armadillo	44	80	50	90	28	50	55
HAT	0	0	7	100	0	0	7

<sup>1</sup>PFAM profiles used were PF02984 (HEAT), PF03130 (PBS), PF00514 (Armadillo) and PF02184 (HAT).

<sup>2</sup>SMART profiles used were SM00567 (PBS), SM00185 (Armadillo) and SM00386 (HAT).

doi:10.1371/journal.pcbi.1000304.t003

A well characterized example is the TOG domain (an alpha-rod of HEAT repeats), which binds tubulin heterodimers to assist addition of tubulin to the plus-end of microtubules [53]; the crystal structure of the TOG domain in *Caenorhabditis elegans* Zyg9 suggests how this interaction may happen through intra-repeat turns [36]. There is evidence of other microtubule-interacting sequences with alpha-rod repeats: yeast Stu2p binds tubulin [36], clathrin-coated vesicles are assembled along microtubules [54], the protein phosphatase 2A (PP2A) binds to microtubules [55], armadillo-repeat containing sperm antigen 6 (Spag6) colocalizes with microtubules [56] (its homolog in *Chlamydomonas reinhardtii* is PF16, involved in protein-protein interactions required for microtubule stability and flagellar motility [57]), huntingtin association with microtubules was initially found in vitro [58] and then with the beta subunit of tubulin in vivo [59].

A particular case is the plant specific family Tortifolia1/TOR1/SPR2, first characterized in *Arabidopsis thaliana* as microtubule-associated protein and containing HEAT repeats [60]. Its N-terminal HEAT repeat domain has been proven to bind to tubulin [61]. Our analysis suggests that this domain possibly contains seven repeats and is distantly related to the CLASP family (data



not shown). Several non-plant protozoan sequences (in amoeba *Dictyostelium discoideum*, and in ciliates *Paramecium tetraurelia* strain *d4-2* and *Tetrahymena thermophila* *SB210*) are more similar to the plant family than to distantly related metazoan members hinting at a complex evolution for this family, possibly involving horizontal transfer events between plants and protozoa (data not shown).

Other proteins with alpha-rod repeats not known to be directly involved in interaction with microtubules or tubulin have broadly associated functions: excess importin-beta blocks kinetochore-associated microtubule formation and enhances centrosome-associated microtubule formation [62], STAG/Scs3 localizes to the spindle poles during mitosis and interacts with NuMA, a spindle pole-associated factor required for mitotic spindle organization [60].

This evidence further confirms a general function of eukaryotic alpha-rods in the organization of cellular structure, chromosome segregation, vesicular transport, and control of cell division by protein-protein interactions that tend to involve the microtubules if not tubulin subunits directly.

### Study of Huntingtin

We demonstrated how to combine information from homologous proteins and secondary structure predictions for a better definition of domains of repeats. We used this approach to define three domains of alpha-rod repeats in human huntingtin: H1 between positions 114–413, H2 between 672–969, and H3 between 2667–2938 (Figure 3A). The definition of these three domains correlates well with previous definitions of cleavage sites in huntingtin. In striatum of brains from patients of Huntington's disease a 40–50 kDa N-terminal and a C-terminal 30–50 kDa fragment are observed [63], which would include H1 and H3, respectively. In addition, several caspase cleavage sites have been verified for huntingtin in positions 513, 552 and 586 [64], which fall in between predicted H1 and H2 alpha-rods.

Using our predictions, we verified for the first time interactions between domains of human huntingtin. These involve three domains of HEAT-repeats. Interactions between domains composed of HEAT-repeats are known. For example, several of the subunits of the AP1 clathrin adaptor core are an alpha-rod of HEAT-repeats and interact with each other [65]. We observed the self-association of one of the huntingtin fragments containing a HEAT-repeat domain. This suggests the possibility that huntingtin homodimerizes through inter-molecular association of this domain, in agreement with previous reports [66]. Homodimerization through interaction of domains with HEAT repeats has been suggested for the DNA-PKc/Ku70/Ku80 complex [67].

The interaction of these domains implies their folding in functional units that correspond to the boundaries we have defined. These results are the first demonstration of domains in huntingtin. This opens avenues for further research into the structure and function of this large protein, which had been hampered until now by its lack of definition in terms of structural units. It is now possible to study the interaction of huntingtin with other proteins on a per domain basis.

### Conclusion

We have provided a way forward for the description of these elusive repeats that will facilitate the characterization of domains, structures, and eventually functions of a large number of proteins, possibly up to 0.5% of the proteomes of eukaryotic organisms. Further work is needed to expand the scope of the method, for example to detect HAT repeats and conceivably other as-yet undiscovered alpha-rod repeats. To facilitate the use of the method we have made it available at <http://www.ogic.ca/>

projects/ard. Results of the analysis of protein families can be studied together using ARD in combination with secondary structure predictions via an updated version of our BiasViz multiple sequence alignment viewer (<http://biasviz.sourceforge.net>).

## Methods

### Neural Network

We used a neural network of feed-forward type with three layers of neurons [14]. Inputs were obtained by scanning the sequence with a 39 amino acid window. The encoding procedure converts the sequence into a binary string where each amino acid is codified by the binary pattern. The length of the entry layer is 39 times 20, where 20 is the number of possible amino acids. One hidden layer with three neurons is used for connecting the inputs with the output layer containing one neuron predicting whether the window is on a repeat or not (e.g. takes real values from 0.1 to 0.9 where the larger values indicates the larger probability of the repeat detection). This architecture was found to be optimal in terms of recall and precision on the training set and computation time required for training and evaluation. Further details of algorithm and training procedure are available in the supplementary Text S1.

### Cloning of Huntingtin Fragments

DNA fragments coding for huntingtin fragments separating predicted domains of alpha-rod repeats were generated by PCR amplification using pAC1-HD plasmid as template. PCR reactions contained, in a 50  $\mu$ l volume, ~50 ng plasmid DNA, 15 pmol primer oligonucleotides, 20 mM TRIS-HCl pH 8.8, 2.5 mM MgCl<sub>2</sub>, 50 mM KCl, 10 mM 2-mercaptoethanol and 2.5 U Pwo DNA polymerase (Sigma). Fragments were amplified in 30 cycles with the following profile: 60 s denaturation at 94°C followed by 120 s annealing at 45–65°C and 120 s extension at 72°C. Amplified DNA products were isolated from 1.2% agarose gel and recombined into GATEWAY compatible pDONR221 plasmid (Invitrogen), thus creating the desired entry DNA plasmids. The identity of all PCR products was verified by DNA sequencing. The sequences of the oligonucleotide primers used to generate huntingtin fragments are available at the supplementary Text S1.

Recombination of entry vectors with pACT-DM and pBTM116\_D9 plasmids was used to create prey and bait plasmid constructs for Y2H interaction mating, respectively. Recombination of different DNA fragments was checked by BsrGI restriction.

### Y2H Analysis of Huntingtin Fragments

DNA sequences encoding the huntingtin fragments Htt1-506Q23, Htt507-1230, Htt1223-1941, Htt1934-2666, Htt2536-3144 and Htt2721-3144 were sub-cloned into DNA binding domain (baits) and activation domain (preys) Y2H plasmids using GATEWAY technology (Invitrogen) and a matrix of individual MATa and MATalpha yeast strains was generated for systematic interaction mating [68]. Then, yeast strains expressing bait and prey proteins were mixed in 96-well microtiter plates and diploid yeast strains were formed on YPD agar plates. Y2H interactions were scored by the frequency of appearance on the SDIV agar plates and  $\beta$ -galactosidase activity in SDII and SDIV nylon membranes, respectively. Growth in SDII-agar was monitored as a mating control.

### Cell Line, Cell Culture and Western Blot

Human embryonic kidney HEK293 cells were seeded in 96-well plates and cultured in Dulbecco's modified Eagle's medium

supplemented with 10% fetal bovine serum at 37°C and 5% CO<sub>2</sub>. Co-transfection of plasmids was done using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocol. The analyses were performed after 48 hours of transfection. For immunoblotting and LUMIER assay, cells were lysed at 4°C for 40 min in 100 µl lysis buffer containing 50 mM HEPES-KOH pH = 7.4, 150 mM NaCl, 0.1% NP40, 1.5 mM MgCl<sub>2</sub>, 1 mM EDTA, 1 mM DTT, 75 Unit/ml Benzodase (Merck) in the presence of protease inhibitor cocktail (Roche Diagnostic). The expression of the constructs was analyzed by Western blot using antibodies against V5-epitope (Invitrogen) and Protein-A (Sigma), while equal protein loading with anti-tubulin antibodies (Figure 3C).

### LUMIER Assay

For LUMIER assay two vectors were generated based on pCDNA3.1(+) (Clontech). For the pPAREni-DM the following cassette was cloned between the BamHI and XbaI sites: Kozak sequence, a double protein A epitope, Renilla Luciferase and the ccdB cassette with flanking R1 and R2 att-sites. For the pFireV5-DM vector the following cassette was cloned between the BamHI and XbaI sites: firefly Luciferase, V5 epitope and the ccdB cassette with flanking R1 and R2 att-sites. (Sequences of cloned inserts are in Supplementary Table S4 in Text S1).

Pairs of PA-Renilla and firefly-V5-tagged huntingtin-fragment fusion proteins were co-expressed in HEK293 cells. Cell extracts were prepared and assessed for the expression of the fusion proteins by immunoblotting and luciferase assays. Protein complexes were isolated from 70 µl cell extracts using 5 µl IgG-coated Dynal magnetic beads (Dynabeads M-280 Sheep anti-Rabbit IgG), subsequently washed with 100 µl PBS, and the binding of the firefly-V5-tagged fusion huntingtin fragment (Co-

IP) to the PA-Renilla-tagged fusion huntingtin fragment protein was quantified by measuring the firefly luciferase activity in a luminescence plate reader (TECAN Infinite M200). Renilla activity was also measured as a control for PA-Renilla constructs expression and binding (IP, data not shown). Luciferase activity was measured using the Dual-Glo Luciferase Assay System (Promega) and a luminescence plate reader (TECAN Infinite M200). Each experiment was performed as triplicate transfection.

### Supporting Information

**Dataset S1** Annotated sequences used for the training set  
Found at: doi:10.1371/journal.pcbi.1000304.s001 (0.14 MB TDS)

**Dataset S2** Full length multiple sequence alignment of human huntingtin and representative homologs  
Found at: doi:10.1371/journal.pcbi.1000304.s002 (0.19 MB TDS)

**Text S1** Supplementary text and supporting figures  
Found at: doi:10.1371/journal.pcbi.1000304.s003 (0.63 MB DOC)

### Acknowledgments

The authors wish to thank Dr. Maciej Lalowski (University of Helsinki) for critical reading of the manuscript and useful comments.

### Author Contributions

Conceived and designed the experiments: EEW MAAN. Performed the experiments: GAP TR US AA RF PP. Analyzed the data: GAP TR US LSP. Contributed reagents/materials/analysis tools: GAP SS MRH. Wrote the paper: MAAN.

### References

- Andrade MA, Ponting CP, Gibson TJ, Bork P (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol* 298: 521–537.
- Andrade MA, Perez-Iratxeta C, Ponting CP (2001) Protein repeats: structures, functions, and evolution. *J Struct Biol* 134: 117–131.
- Andrade MA, Bork P (1995) HEAT repeats in the Huntington's disease protein. *Nat Genet* 11: 115–116.
- Andrade MA, Petosa C, O'Donoghue SI, Muller CW, Bork P (2001) Comparison of ARM and HEAT protein repeats. *J Mol Biol* 309: 1–18.
- Peifer M, Berg S, Reynolds AB (1994) A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell* 76: 789–791.
- Preker PJ, Keller W (1998) The HAT helix, a repetitive motif implicated in RNA processing. *Trends Biochem Sci* 23: 15–16.
- Boguski MS, Murray AW, Powers S (1992) Novel repetitive sequence motifs in the alpha and beta subunits of prenyl-protein transferases and homology of the alpha subunit to the MAD2 gene product of yeast. *New Biol* 4: 408–411.
- Mao Y, Nickitenko A, Duan X, Lloyd TE, Wu MN, et al. (2000) Crystal structure of the VHS and FYVE tandem domains of Hrs, a protein involved in membrane trafficking and signal transduction. *Cell* 100: 447–456.
- Sagermann M, Stevens TH, Matthews BW (2001) Crystal structure of the regulatory subunit H of the V-type ATPase of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 98: 7134–7139.
- Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34: D247–D251.
- Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, et al. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34: D257–D260.
- Gruber M, Soding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. *J Struct Biol* 155: 140–145.
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305: 567–580.
- Rosenblatt F (1962) *Principles of Neurodynamics*. Washington, DC: Spartan Books.
- Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232: 584–599.
- Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40: 502–511.
- Rost B, Casadio R, Fariselli P, Sander C (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci* 4: 521–533.
- Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins* 20: 216–226.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
- McGuinness BE, Hirota T, Kudo NR, Peters JM, Nasmyth K (2005) Shugoshin prevents dissociation of cohesin from centromeres during mitosis in vertebrate cells. *PLoS Biol* 3: e86. doi:10.1371/journal.pbio.0030086.
- Hauf S, Roitinger E, Koch B, Dittlich CM, Mechtler K, et al. (2005) Dissociation of cohesin from chromosome arms and loss of arm cohesion during early mitosis depends on phosphorylation of SA2. *PLoS Biol* 3: e69. doi:10.1371/journal.pbio.0030069.
- Losada A, Yokochi T, Kobayashi R, Hirano T (2000) Identification and characterization of SA/Sec3p subunits in the *Xenopus* and human cohesin complexes. *J Cell Biol* 150: 405–416.
- Chelysheva L, Diallo S, Vezon D, Gendrot G, Vrielynck N, et al. (2005) AtREC8 and AtSCC3 are essential to the monopolar orientation of the kinetochores during meiosis. *J Cell Sci* 118: 4621–4632.
- Cena A, Orłowski J, Machula K, Fronk J, Kurlandzka A (2007) Substitution F659G in the *Irr1p/Sec3p* cohesin influences the cell wall of *Saccharomyces cerevisiae*. *Cell Struct Funct* 32: 1–7.
- Lupas A, Baumeister W, Hofmann K (1997) A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex). *Trends Biochem Sci* 22: 195–196.
- Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* 36: W197–W201.
- Kajava AV (2002) What curves  $\alpha$ -solenoids? Evidence for an  $\alpha$ -helical toroid structure of Rpn1 and Rpn2 proteins of the 26 S proteasome. *J Biol Chem* 277: 49791–49798.
- Oeffinger M, Dlakic M, Tollervy D (2004) A pre-ribosome-associated HEAT-repeat protein is required for export of both ribosomal subunits. *Genes Dev* 18: 196–209.
- Staub E, Fizev P, Rosenthal A, Hinzmann B (2004) Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire. *Bioessays* 26: 567–581.
- Smith L, Van Hateren N, Willan J, Romero R, Blanco G, et al. (2003) Candidate testis-determining gene, *Maestro* (Mro), encodes a novel HEAT repeat protein. *Dev Dyn* 227: 600–607.
- Neuwald AF, Hirano T (2000) HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. *Genome Res* 10: 1445–1452.

32. Turner NC, Lord CJ, Iorns E, Brough R, Swift S, et al. (2008) A synthetic lethal siRNA screen identifying genes mediating sensitivity to a PARP inhibitor. *EMBO J* 27: 1368–1377.
33. Baillat D, Hakimi MA, Naar AM, Shilatifard A, Cooch N, et al. (2005) Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell* 123: 265–276.
34. Nakayama M, Kikuno R, Ohara O (2002) Protein–protein interactions between large proteins: two-hybrid screening using a functionally classified library composed of long cDNAs. *Genome Res* 12: 1773–1784.
35. Cassimeris L, Morabito J (2004) TOGp, the human homolog of XMAP215/Dis1, is required for centrosome integrity, spindle pole organization, and bipolar spindle assembly. *Mol Biol Cell* 15: 1580–1590.
36. Al-Bassam J, Larsen NA, Hyman AA, Harrison SC (2007) Crystal structure of a TOG domain: conserved features of XMAP215/Dis1-family TOG domains and implications for tubulin binding. *Structure* 15: 355–362.
37. Mimori-Kiyosue Y, Grigoriev I, Lansbergen G, Sasaki H, Matsui C, et al. (2005) CLASP1 and CLASP2 bind to EB1 and regulate microtubule plus-end dynamics at the cell cortex. *J Cell Biol* 168: 141–153.
38. Akhmanova A, Steinmetz MO (2008) Tracking the ends: a dynamic protein network controls the fate of microtubule tips. *Nat Rev Mol Cell Biol* 9: 309–322.
39. Kajava AV, Gorbea C, Ortega J, Rechsteiner M, Steven AC (2004) New HEAT-like repeat motifs in proteins regulating proteasome structure and function. *J Struct Biol* 146: 425–430.
40. Hatfield MD, Reis AM, Obeso D, Cook JR, Thompson DM, et al. (2006) Identification of MMS19 domains with distinct functions in NER and transcription. *DNA Repair (Amst)* 5: 914–924.
41. Adami A, Garcia-Alvarez B, Arias-Palomo E, Barford D, Llorca O (2007) Structure of TOR and its complex with KOG1. *Mol Cell* 27: 509–516.
42. Golas MM, Sander B, Will CL, Luhrmann R, Stark H (2003) Molecular architecture of the multiprotein splicing factor SF3b. *Science* 300: 980–984.
43. Shomura Y, Dragovic Z, Chang HC, Tzvetkov N, Young JC, et al. (2005) Regulation of Hsp70 function by HspBP1: structural analysis reveals an alternate mechanism for Hsp70 nucleotide exchange. *Mol Cell* 17: 367–379.
44. Nasir J, Floresco SB, O’Kusky JR, Diewert VM, Richman JM, et al. (1995) Targeted disruption of the Huntington’s disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell* 81: 811–823.
45. Harjes P, Wanker EE (2003) The hunt for huntingtin function: interaction partners tell many different stories. *Trends Biochem Sci* 28: 425–433.
46. Melvin I, Ie E, Kuang R, Weston J, Stafford WN, et al. (2007) SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. *BMC Bioinformatics* 8(Suppl 4): S2.
47. McGuffin LJ, Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* 19: 874–881.
48. Lee SY, Skolnick J (2008) Benchmarking of TASSER\_2.0: an improved protein structure prediction algorithm with more accurate predicted contact restraints. *Biophys J* 95: 1956–1964.
49. Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, et al. (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* 307: 1621–1625.
50. Bai Y, Auperin TC, Chou CY, Chang GG, Manley JL, et al. (2007) Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. *Mol Cell* 25: 863–875.
51. Legrand P, Pinaud N, Minvielle-Sebastia L, Fribourg S (2007) The structure of the CstF-77 homodimer provides insights into CstF assembly. *Nucleic Acids Res* 35: 4515–4522.
52. Matsuura Y, Stewart M (2004) Structural basis for the assembly of a nuclear export complex. *Nature* 432: 872–877.
53. Al-Bassam J, van Bruegel M, Harrison SC, Hyman A (2006) Stu2p binds tubulin and undergoes an open-to-closed conformational change. *J Cell Biol* 172: 1009–1022.
54. Falkowska-Hansen B, Falkowski M, Metharom P, Kronic D, Goerdts S (2007) Clathrin-coated vesicles form a unique net-like structure in liver sinusoidal endothelial cells by assembling along undisrupted microtubules. *Exp Cell Res* 313: 1745–1757.
55. Sontag E, Numbhakdi-Craig V, Bloom GS, Mumby MC (1995) A novel pool of protein phosphatase 2A is associated with microtubules and is regulated during the cell cycle. *J Cell Biol* 128: 1131–1144.
56. Sapiro R, Tarantino LM, Velazquez F, Kiriakidou M, Hecht NB, et al. (2000) Sperm antigen 6 is the murine homologue of the *Chlamydomonas reinhardtii* central apparatus protein encoded by the PF16 locus. *Biol Reprod* 62: 511–518.
57. Smith EF, Lefebvre PA (2000) Defining functional domains within PF16: a central apparatus component required for flagellar motility. *Cell Motil Cytoskeleton* 46: 157–165.
58. Tukamoto T, Nukina N, Ide K, Kanazawa I (1997) Huntington’s disease gene product, huntingtin, associates with microtubules in vitro. *Brain Res Mol Brain Res* 51: 8–14.
59. Hoffner G, Kahlem P, Djian P (2002) Perinuclear localization of huntingtin as a consequence of its binding to microtubules through an interaction with beta-tubulin: relevance to Huntington’s disease. *J Cell Sci* 115: 941–948.
60. Gregson HC, Schmiesing JA, Kim JS, Kobayashi T, Zhou S, et al. (2001) A potential role for human cohesin in mitotic spindle aster assembly. *J Biol Chem* 276: 47575–47582.
61. Yao M, Wakamatsu Y, Itoh TJ, Shoji T, Hashimoto T (2008) Arabidopsis SPIRAL2 promotes uninterrupted microtubule growth by suppressing the pause state of microtubule dynamics. *J Cell Sci* 121: 2372–2381.
62. Tulu US, Fagerstrom C, Ferenz NP, Wadsworth P (2006) Molecular requirements for kinetochore-associated microtubule formation in mammalian cells. *Curr Biol* 16: 536–541.
63. Mende-Mueller LM, Toneff T, Hwang SR, Chesselet MF, Hook VY (2001) Tissue-specific proteolysis of Huntingtin (htt) in human brain: evidence of enhanced levels of N- and C-terminal htt fragments in Huntington’s disease striatum. *J Neurosci* 21: 1830–1837.
64. Wellington CL, Ellerby LM, Gutekunst CA, Rogers D, Warby S, et al. (2002) Caspase cleavage of mutant huntingtin precedes neurodegeneration in Huntington’s disease. *J Neurosci* 22: 7862–7872.
65. Heldwein EE, Macia E, Wang J, Yin HL, Kirchhausen T, et al. (2004) Crystal structure of the clathrin adaptor protein 1 core. *Proc Natl Acad Sci U S A* 101: 14108–14113.
66. Li W, Serpell LC, Carter WJ, Rubinsztein DC, Huntington JA (2006) Expression and characterization of full-length human huntingtin, an elongated HEAT repeat protein. *J Biol Chem* 281: 15916–15922.
67. Spagnolo L, Rivera-Calzada A, Pearl LH, Llorca O (2006) Three-dimensional structure of the human DNA-PKcs/Ku70/Ku80 complex assembled on DNA and its implications for DNA DSB repair. *Mol Cell* 22: 511–519.
68. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell* 122: 957–968.
69. Collins BM, McCoy AJ, Kent HM, Evans PR, Owen DJ (2002) Molecular architecture and functional model of the endocytic AP2 complex. *Cell* 109: 523–535.
70. Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem Sci* 25: 300–302.
71. Huska MR, Buschmann H, Andrade-Navarro MA (2007) BiasViz: visualization of amino acid biased regions in protein alignments. *Bioinformatics* 23: 3093–3094.