HEAT repeats in the Huntington's disease protein

Sir — Huntington's disease (HD) is an autosomal dominant neurogenetic disorder with various progressive abnormalities in the brain[1]. A human gene in which expanded CAG-repeats (runs of glutamine in the gene product) cause the disease[2], as well as homologues in mouse, rat and fugu have been identified (see ref. 1). The gene product (huntingtin) encodes a large, widely expressed protein of 3144 residues[2] located in the cytoplasm[3–5], but no precise function or homology to any other protein has yet been described. In the course of our systematic analyses of multidomain disease proteins[6,7], we have noted that a considerable fraction of huntingtin contains tandem arrays of a repeat that we call HEAT, after four functionally characterized proteins in which the repeat was detected: huntingtin, elongation factor 3 (EF3), the regulatory A subunit (65 kd) of protein phosphatase 2A (PP2A) and TOR1, a target of rapamycin that seems to be essential for progression of the G1 phase of the cell cycle.

When searching sequence databases with huntingtin, we found some weak similarity to the regulatory A subunit of PP2A[8] using the program Blastp (probability of matching by chance, P=0.042, ref. 9). This PP2A subunit consists entirely of 15 repeats of about 40 amino acids[8]; three are homologous to a region in the protein kinase VPS15 (ref. 10). Reciprocal database searches with the 65 kd regulatory A subunit of PP2A as a query indicated an even more tempting similarity to huntingtin (P = 0.0023). Curiously, PP2A matches significantly some other cytoplasmic regulatory proteins (Blastp P-values in the order of 10^{-6}). Further Blastp iterations together with motif and

Table 1 Summary of proteins containing HEAT repeats

Protein	Species[a]	Function[b] domains	Location[c]	HEAT repeats[d]	Positions	Protein size	Database[e]
Huntingtin	A	vesicle trafficking[4]	cyt	10	205–329 745–942 534–1710	3190	P42858
EF3	F	protein biosynthesis/C-terminal ABC transporter domain[23]	cyt	8	9–322	1043	P16521
A subunit PP2A	A,F,P	regulation of protein phosphorylation[8]	cyt	15	39–635	635	P31383
TOR (FRAP)	A,F	G1 phase progression/C-terminal PI3 kinase domain[24]	cyt	20	71–522 628–1147	2400	P35169
GCN1	F	transport of tRNA substrates?[25]	cyt ?	36	1033–2588	2672	P33892
VP15	F	vesicle mediated protein transport/ N-terminal protein kinase[26]; C-terminal WD40 repeats (data not shown)	cyt	7	418–730	1453	P22219
Importins	A	nuclear protein import/export pathway[27]	cyt	11	122–482 600–725	875	L38644
Yscl8300_13	F	importin homologue in yeast	cyt ?	11	319–533 592–815	861	U19028
PSE1	F	involved in protein secretion[28]	cyt ?	12	6–298 364–615	1089	P32337
YBA4	F	?	?	4	1480–1667	2493	P35194
YBM7	F	?	?	10	118–355 417–621	918	P38217
YEU0	F	?	?	12	35–247 395–721	1113	P40069
Sc8021x_14	F	?	?	20	163–947	970	Z49704
Ced2045_1	A	?	?	15	787–1015 1072–1286 1310–1521	1792	Z35369
Humkg1bb	A	?	?	14	276–472 856–1050 1242–1399	2032	D43498

[a]Species range (A, animals; F, fungi; P, plants); [b]PP2A, TOR and EF3 might also be involved in transport processes; PP2A acts during the transport-intensive cell cycle events[8] as does TOR[24], EF3 contains an ABC transporter domain[23] that is usually associated with transport processes; [c]cyt, cytoplasm; cyt ?, likely cytoplasm; [d]These correspond to rather conservative cut-offs as at least three consecutive repeats with scores above 2.5 (PROFILEGAP[29]) were required. Thus, more divergent repeats within the identified proteins might exist; [e]database accession numbers (Swissprot numbers start with P, all others are from EMBL).

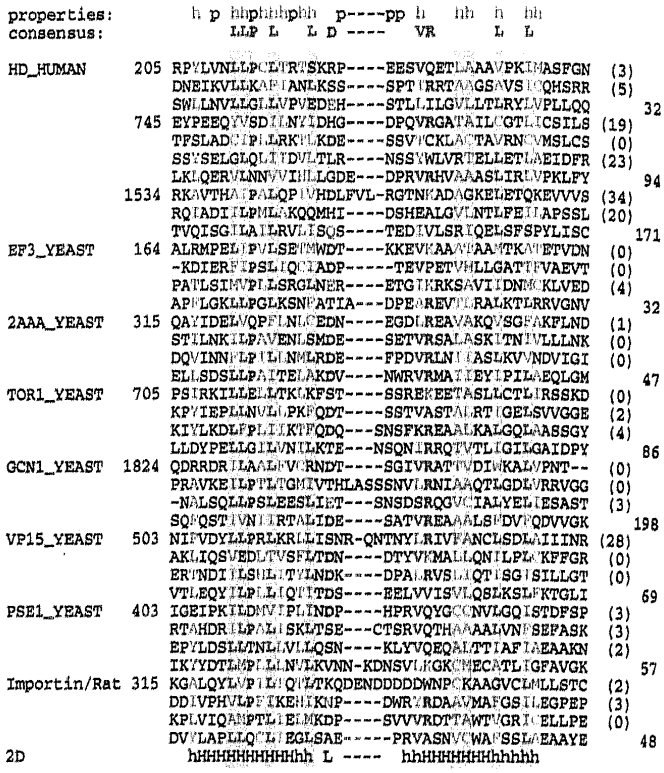

Fig.1 Alignment of HEAT repeats in huntingtin with selected consecutive repeats of the functionally characterised members of this new superfamily. SwissProt database codes are listed in the first column. The residue numbers at the beginning and end of the consecutive repeats flank the sequences. The number of residues between consecutive repeats is shown in brackets. The top 'property line' indicates hydrophobic (h, green) and non-hydrophobic positions (p) conserved in more than 80% of the sequences; the consensus line shows residues that are conserved in more than 40% of the sequences (bold, coloured). Strictly hydrophobic positions are boxed. The bottom line gives the secondary structure prediction (H/h - helix; L/dot - coil, with 82%/72% expected accuracy using the PhD program[19]) as averaged over the 14 sequence families predicted to contain HEAT repeat. All families were independently predicted to contain a helix-coil-helix arrangement in the HEAT regions with only slight variations in the helix borders. Despite the high helical contents, there is no coiled coil potential in these regions as verified by the COILS program[20]. Each protein was independently subjected to a screen for internal repeats using the REPEATS program (M. Vingron, GMD, Bonn) and most of them, including huntingtin, gave highly significant results (6 standard deviations above the mean calculated for random sequences).

Most of the proteins have significant Blastp *P*-values ($< 10^{-6}$) to at least one other member of the HEAT repeat family. In addition, the repeats shown here (excluding those in huntingtin) were used for iterative motif and profile searches (starting with PP2A) to verify the reciprocity of our findings regarding huntingtin (see ref. 21). Not only could all proteins shown in Table 1 be detected, but the HEAT repeats in huntingtin could also be identified unambiguously. For example, the third iteration of the ProfileSearch program[22] with three consecutive repeats gave total scores above 11.89 for all proteins shown in Table 1. The first probable false positive scored with 10.27. A few proteins including adaptins (located at the cytoplasmic face of coated vesicles as part of a clathrin-associated complex) scored below the proteins discussed but above the first false positive, and are additional candidates that might contain HEAT repeats, but have not been considered here due to our conservative cut-off.

profile searches (see Fig. 1) reveal a total of 14 distinct eukaryotic proteins (not counting species redundancies) including huntingtin that contain multiple HEAT repeats (Table 1).

The divergent HEAT repeats vary in length between 37 and 43 amino acids, occur in at least 3 consecutive repeats in every protein (Table 1) and appear to consist of two α helices (Fig. 1). The helical count would be between the nebulin[11] and spectrin repeats[12] with one and three helices, respectively. The rather hydrophobic nature of the repeats suggests a tight packing against each other, but might also contribute to the interaction with other proteins. This is supported by experimental data on HEAT repeat-containing A subunits of PP2A which form rod-like helical structures and bind to T antigens of several viruses as well as to the PP2A B subunit[13]. Other characterised, mainly cytoplasmic repeats also appear to be involved in protein–protein interactions such as leucine-rich repeats that form independent β/α superstructures required for protein-binding[1] as well as ankyrin[15], TPR[16] and WD40 repeats[17] which all seem to contain α-helices.

In addition to the sequence similarity detected by independent methods (Fig.1), the functionally characterised proteins of the HEAT family share a number of features that support our findings. i) The homologous regions are all predicted to adopt an α-helical topology (Fig. 1). ii) Although the HEAT repeats themselves are rather divergent, they always occur as consecutive units multiple times within each protein and thus strengthen our predictions. iii) All proteins of the HEAT family seem to be very large (Table1) and most of them are known to be part of protein complexes. iv) The functionally characterised proteins containing HEAT repeats are eukaryotic regulatory cytoplasmic proteins; most of them seem to be involved in cytoplasmic transport processes (Table 1).

The presence of the HEAT repeats in huntingtin is thus consistent with a recent report that proposes a role in vesicle trafficking[4]. The HEAT repeats in huntingtin succeed the glutamine runs (separated by a proline-rich linker), the extension of which leads to the disease, perhaps via artificial protein agglomeration[18] and/or altering neighbouring domains in the native protein[3].

In conclusion, several HEAT repeats appear to be required to form a rod that provides binding sites for the interaction with other proteins (as shown for PP2A[13]); they might have a general role in cytoplasmic transport processes. The identification of the HEAT repeats allows a first glimpse into the modular architecture of a large group of cytoplasmic regulatory proteins. It might guide ligand-binding studies as well as the determination of the three-dimensional structure of HEAT repeat-containing domains.

Miguel A. Andrade[1]
Peer Bork[1,2]
[1]EMBL, Meyerhofstr. 1, 69012 Heidelberg, Germany
[2]Max-Delbrück-Center for Molecular Medicine, 13122 Berlin-Buch, Germany

1. Albin, R.L. & Tagle, D.A. *Trends Neurosci.* 18, 11–14 (1995).
2. Huntington's disease collaborative research group. *Cell* **72**, 971–983 (1993).
3. Trossier, Y. *et al. Nature Genet.* **10**, 104–110 (1995).
4. DiFiglia, M. *et al. Neuron* **14**, 1075–1081 (1995).
5. Sharp, A.H. *et al. Neuron* **14**, 1065–1074 (1995).
6. The international PKD1 consortium. *Cell* **81**, 289–298 (1995).
7. Bork, P. & Sudol, M. *Trends biochem. Sci* . **19**, 531–533 (1994).
8. Hemmings, B.A. *et al. Biochem.* **29**, 3166–3173 (1990).
9. Altschul, S.F. *et al. Nature Genet.* **6**, 119–129 (1994).
10. Herman, P.K. *et al. Trends Cell Biol.* **2**, 363–368 (1992).
11. Pfuhl, M. *et al. EMBO J.* **13**, 1782–1789 (1994).
12. Hartwig, J.H. *Protein Profile* **1**, 711–749 (1994).
13. Ruediger, R. *et al. J. Virol* . **68**, 123–129 (1994).
14. Kobe, B. & Deisenhofer, J. *Nature* **374**, 183–186 (1995).
15. Bork, P. *Proteins* **17**, 363–374 (1993).
16. Lamb, J.R. *et al. Trends biochem. Sci.* **20**, 257–259 (1995).
17. Neer, E.J. *et al. Nature* **371**, 297–300 (1994).
18. Perutz, M.F. *et al. Proc. natn. Acad. Sci. U.S.A.* **91**, 5355–5358 (1994).
19. Rost, B. & Sander, C. *Proteins* **19**, 55–72 (1994).
20. Lupas, A. *et al. Science* **252**, 1162–1164 (1991).
21. Bork, B. & Gibson, T. *Meth. Enzym.* (in the press).
22. Gribskov, M. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 4355–4358 (1987).
23. Qin, S. *et al. J. biol. Chem.* **265**, 1903–1912 (1990).
24. Sabatini, D.M. *et al. Cell* **78**, 35–43 (1994).
25. Marton, M.J. *et al. Molec. cell. Biol.* **13**, (1993).
26. Stack, J.H. *et al. J. Cell. Biol.* **129**, 321–334 (1995).
27. Radu, A. *et al. Proc. natn. Acad. Sci. U.S.A.* **92**, 1769–1773 (1995).
28. Chow, T. Y.-K. *et al. J. Cell Sci.* **101**, 709–719 (1992).