D272–D276 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 8 November 2010*
doi:10.1093/nar/gkq1100

PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins

Amy L. Robertson[1], Mark A. Bate[1], Steve G. Androulakis[2], Stephen P. Bottomley[1,*] and Ashley M. Buckle[1,*]

[1]The Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Faculty of Medicine, Nursing and Health Sciences and [2]Monash eResearch Centre, Monash University, Clayton, Victoria 3800, Australia

Received August 14, 2010; Revised and Accepted October 18, 2010

ABSTRACT

The polyglutamine diseases are caused in part by a gain-of-function mechanism of neuronal toxicity involving protein conformational changes that result in the formation and deposition of β-sheet rich aggregates. Recent evidence suggests that the misfolding mechanism is context-dependent, and that properties of the host protein, including the domain architecture and location of the repeat tract, can modulate aggregation. In order to allow the bioinformatic investigation of the context of polyglutamines, we have constructed a database, PolyQ (http://pxgrid.med.monash.edu.au/polyq). We have collected the sequences of all human proteins containing runs of seven or more glutamine residues and annotated their sequences with domain information. PolyQ can be interrogated such that the sequence context of polyglutamine repeats in disease and non-disease associated proteins can be investigated.

INTRODUCTION

Polyglutamine (PolyQ) repeats are implicated in several neurodegenerative diseases, including Huntington's disease and several spinocerebellar ataxia's. It is commonly thought that a toxic gain-of-function mechanism is triggered by the presence of a polyQ tract, involving a conformational change within the protein and the formation and deposition of β-sheet rich amyloid-like fibrils (1–3).

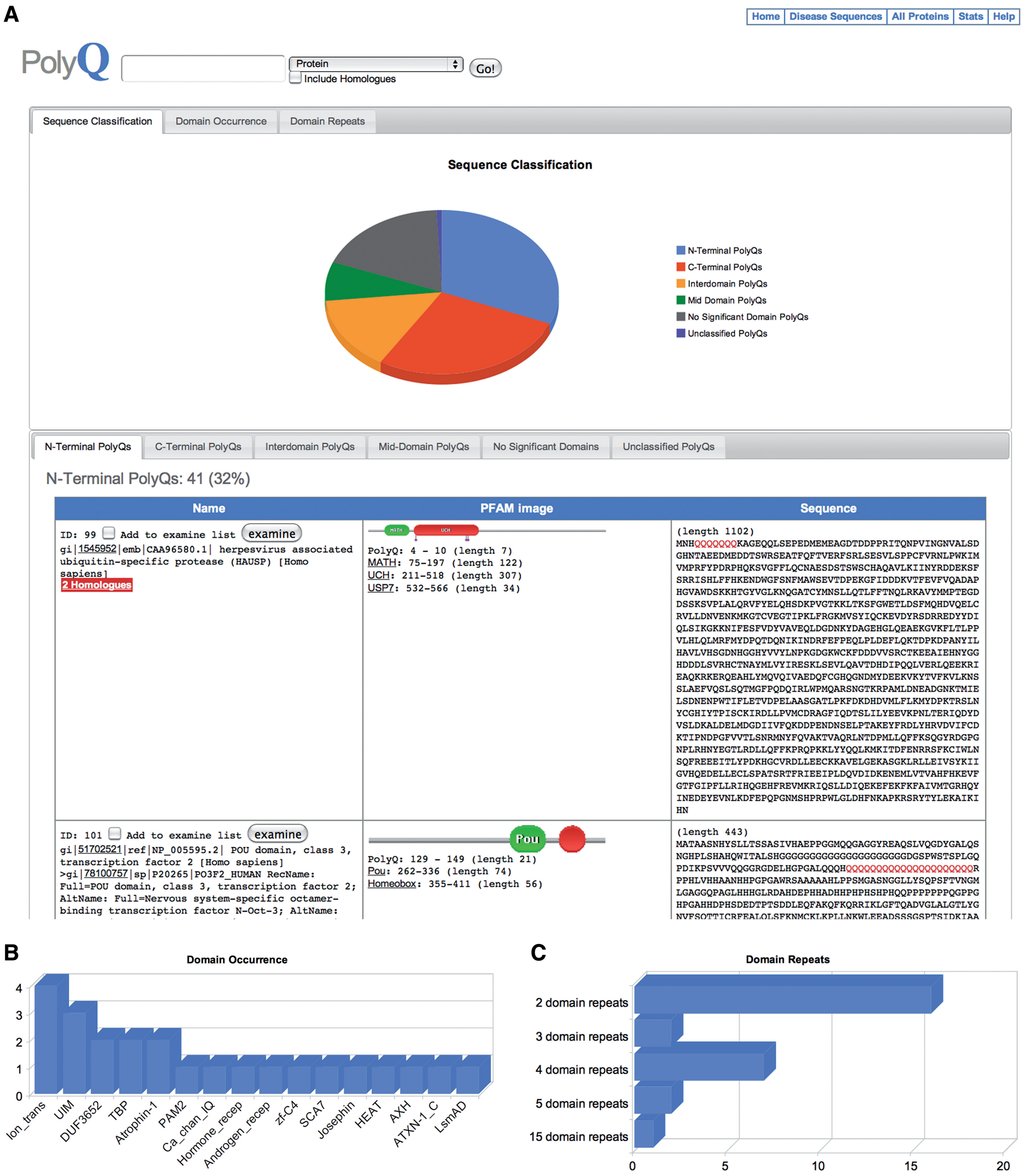
The length of the polyQ repeat is critical to pathogenesis; however, there is evidence that other protein factors, including the location, type and number of flanking domains can modulate pathogenesis (4–10). Although there are many human polyQ-containing proteins (11), only nine polyQ-containing proteins are implicated in pathogenesis, and the precise repeat threshold to pathogenesis varies within the disease subset, for example, a 37 glutamine repeat is sufficient to lead to Huntington's disease, while SCA3 results only when the polyQ repeat expands to 45 or greater (12–14).

Many other human, non-disease related proteins contain polyQ repeats, which are intrinsically prone to expansion at the genetic level (11,15,16). In fact, a 40 glutamine repeat is the normal allele present in forkhead box P2 transcription factor; a protein that has not been found to be associated with a polyQ disease (17,18). This evidence has led to the hypothesis that protein characteristics modulate the propensity of polyQ-containing proteins to aggregate and cause disease. To investigate the variable characteristics of polyQ proteins we have performed a bioinformatics investigation of the protein context of polyglutamine repeats, and constructed a web-accessible database of all human proteins containing a polyQ repeat greater than seven glutamines in length, termed 'PolyQ'. The PolyQ database provides a tool to compare the polyQ repeat location, the occurrence/type of domains and the number of domain repeats present across disease and non-disease proteins.

PolyQ DESCRIPTION AND USE

PolyQ was created using open-source MySQL relational database server software, version 5.0.82 (http://www.mysql.com), running on an Apple 8-core 3.0 GHz Xeon/OS X Server (version 10.5.8). The database consists of three tables. A web-based query interface to the database was developed using the PHP5 programming language, hosted via Apache 2.2.14. The user interface was developed with the utilisation of the JQuery Javascript library and JQuery widgets. Charts and graphs are constructed on the fly using the Google Visualization API.

*To whom correspondence should be addressed. Tel: +613 9902 9313; Fax: +613 9902 9500; Email: ashley.buckle@monash.edu
Correspondence may also be addressed to Stephen P. Bottomley. Tel: +613 9902 9313; Fax: +613 9902 9500; Email: steve.bottomley@monash.edu

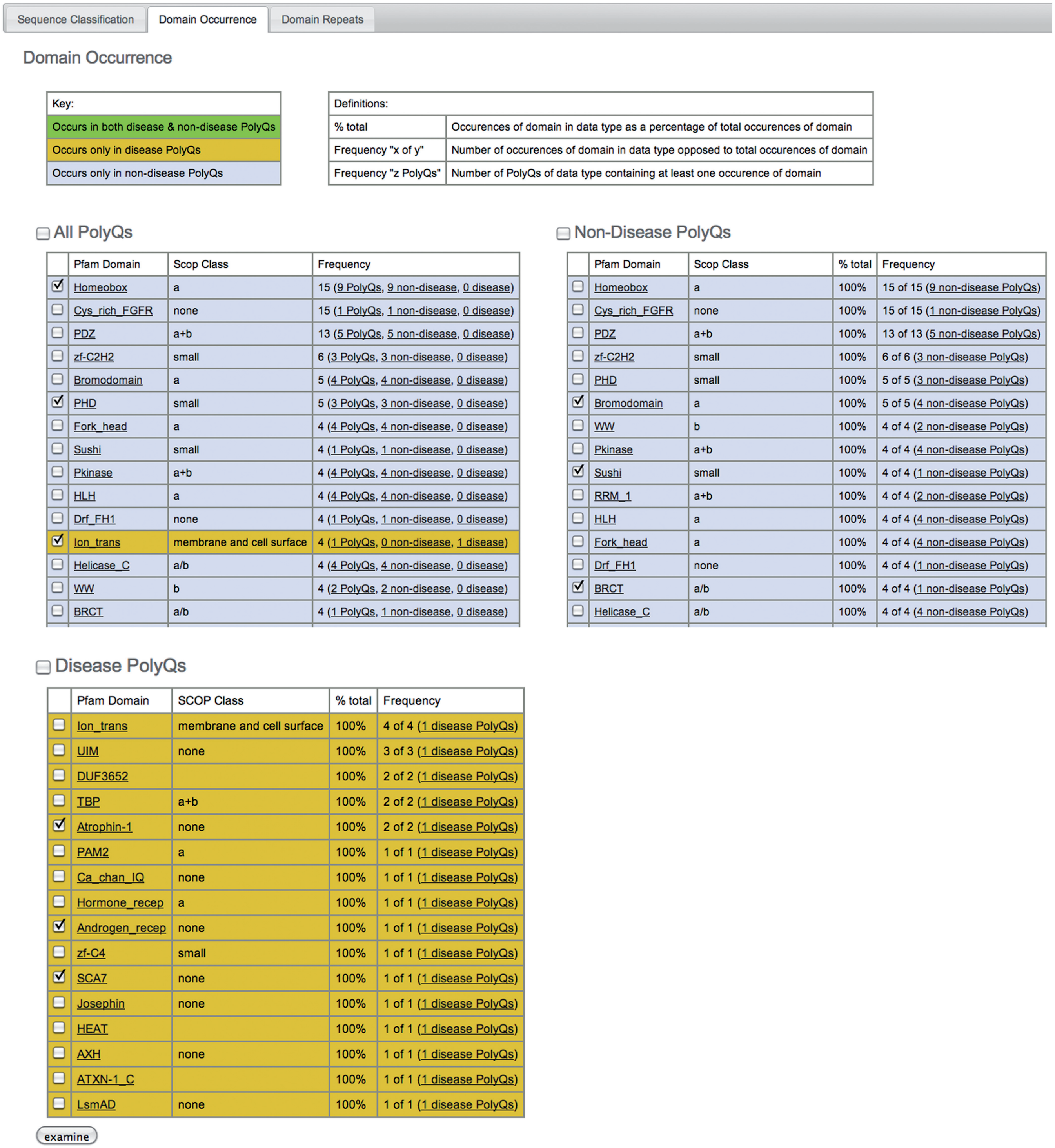


N-Terminal PolyQs: 41 (32%)

Name	PFAM image	Sequence
ID: 99 □ Add to examine list examine gi|1545952|emb|CAA96580.1| herpesvirus associated ubiquitin-specific protease (HAUSP) [Homo sapiens] 2 Homologues	PolyQ: 4 - 10 (length 7) MATH: 75-197 (length 122) UCH: 211-518 (length 307) USP7: 532-566 (length 34)	(length 1102) MNHQQQQQQQKAGEQQLSEPEDMEMEAGDTDDPPRITQNPVINGNVALSD GHNTAEEDMEDDTSWRSEATFQFTVERFSRLSESVLSPPCFVRNLPWKIM VMPRFYPDRPHQKSVGFFLQCNAESDSTSWSCHAQAVLKIINYRDDEKSF SRRISHLFFHKENDWGFSNFMAWSEVTDPEKGFIDDDKVTFEVFVQADAP HGVAWDSKKHTGYVGLKNQGATCYMNSLLQTLFFTNQLRKAVYMMPTEGD DSSKSVPLALQRVFYELQHSDKPVGTKKLTKSFGWETLDSFMQHDVQELC RVLLDNVENKMKGTCVEGTIPKLFRGKMVSYIQCKEVDYRSDRREDYYDI QLSIKGKKNIFESFVDYVAVEQLDGDNKYDAGEHGLQEAEKGVKFLTLPP VLHLQLMRFMYDPQTDQNIKINDRFEFPEQLPLDEFLQKTDPKDPANYIL HAVLVHSGDNHGGHYVVYLNPKGDGKWCKFDDDVVSRCTKEEAIEHNYGG HDDDLSVRHCTNAYMLVYIRESKLSEVLQAVTDHDIPQQLVERLQEEKRI EAQKRKERQEAHLYMQVQIVAEDQFCGHQGNDMYDEEKVKYTVFKVLKNS SLAEFVQSLSQTMGFPQDQIRLWPMQARSNGTKRPAMLDNEADGNKTMIE LSDNENPWTIFLETVDPELAASGATLPKFDKDHDVMLFLKMYDPKTRSLN YCGHIYTPISCKIRDLLPVMCDRAGFIQDTSLILYEEVKPNLTERIQDYD VSLDKALDELMDGDIIVFQKDDPENDNSELPTAKEYFRDLYHRVDVIFCD KTIPNDPGFVVTLSNRMNYFQVAKTVAQRLNTDPMLLQFFKSQGYRDGPG NPLRHNYEGTLRDLLQFFKPRQPKKLYYQQLKMKITDFENRRSFKCIWLN SQFREEEITLYPDKHGCVRDLLEECKKAVELGEKASGKLRLLEIVSYKII GVHQEDELLECLSPATSRTFRIEEIPLDQVDIDKENEMLVTVAHFHKEVF GTFGIPFLLRIHQGEHFREVMKRIQSLLDIQEKEFEKFKFAIVMTGRHQY INEDEYEVNLKDFEPQPGNMSHPRPWLGLDHFNKAPKRSRYTYLEKAIKI HN
ID: 101 □ Add to examine list examine gi|51702521|ref|NP_005595.2| POU domain, class 3, transcription factor 2 [Homo sapiens] >gi|78100757|sp|P20265|PO3F2_HUMAN RecName: Full=POU domain, class 3, transcription factor 2; AltName: Full=Nervous system-specific octamer-binding transcription factor N-Oct-3; AltName:	PolyQ: 129 - 149 (length 21) Pou: 262-336 (length 74) Homeobox: 355-411 (length 56)	(length 443) MATAASNHYSLLTSSASIVHAEPPGGMQQGAGGYREAQSLVQGDYGALQS NGHPLSHAHQWITALSHGGGGGGGGGGGGGGGGGGGDGSPWSTSPLGQ PDIKPSVVVQQGGRGDELHGPGALQQQHQQQQQQQQQQQQQQQQQQQQQR PPHLVHHAANHHPGPGAWRSAAAAAHLPPSMGASNGGLLYSQPSFTVNGM LGAGGQPAGLHHHGLRDAHDEPHHADHHPHPHSHPHQQPPPPPPPPQGPPG HPGAHHDPHSDEDTPTSDDLEQFAKQFKQRRIKLGFTQADVGLALGTLYG NVFSQTTICRFEALQLSFKNMCKLKPLLNKWLEEADSSSGSPTSIDKIAA

Figure 1. (**A**) Typical results of a simple search (blank in this instance), showing graphical breakdown according to sequence classification; Results shown graphically according to domain occurrence (**B**) and domain repeats (**C**) using the tabs at top of page [as seen in (A)].

The PolyQ database was populated by extracting all human sequences from the NCBI non-redundant (NR) database that contained at least seven consecutive glutamine residues. We then performed a Pfam (19) domain search to find protein domains within this subset of sequences. The NCBI NR contains many versions of the same protein, which created bias in the statistical analysis of PolyQ location data. We simplified the



Domain Occurrence

Key:
Occurs in both disease & non-disease PolyQs
Occurs only in disease PolyQs
Occurs only in non-disease PolyQs

Definitions:	
% total	Occurences of domain in data type as a percentage of total occurences of domain
Frequency "x of y"	Number of occurences of domain in data type opposed to total occurences of domain
Frequency "z PolyQs"	Number of PolyQs of data type containing at least one occurence of domain

☐ All PolyQs

	Pfam Domain	Scop Class	Frequency
☑	Homeobox	a	15 (9 PolyQs, 9 non-disease, 0 disease)
☐	Cys_rich_FGFR	none	15 (1 PolyQs, 1 non-disease, 0 disease)
☐	PDZ	a+b	13 (5 PolyQs, 5 non-disease, 0 disease)
☐	zf-C2H2	small	6 (3 PolyQs, 3 non-disease, 0 disease)
☐	Bromodomain	a	5 (4 PolyQs, 4 non-disease, 0 disease)
☑	PHD	small	5 (3 PolyQs, 3 non-disease, 0 disease)
☐	Fork_head	a	4 (4 PolyQs, 4 non-disease, 0 disease)
☐	Sushi	small	4 (1 PolyQs, 1 non-disease, 0 disease)
☐	Pkinase	a+b	4 (4 PolyQs, 4 non-disease, 0 disease)
☐	HLH	a	4 (4 PolyQs, 4 non-disease, 0 disease)
☐	Drf_FH1	none	4 (1 PolyQs, 1 non-disease, 0 disease)
☑	Ion_trans	membrane and cell surface	4 (1 PolyQs, 0 non-disease, 1 disease)
☐	Helicase_C	a/b	4 (4 PolyQs, 4 non-disease, 0 disease)
☐	WW	b	4 (2 PolyQs, 2 non-disease, 0 disease)
☐	BRCT	a/b	4 (1 PolyQs, 1 non-disease, 0 disease)

☐ Non-Disease PolyQs

	Pfam Domain	Scop Class	% total	Frequency
☐	Homeobox	a	100%	15 of 15 (9 non-disease PolyQs)
☐	Cys_rich_FGFR	none	100%	15 of 15 (1 non-disease PolyQs)
☐	PDZ	a+b	100%	13 of 13 (5 non-disease PolyQs)
☐	zf-C2H2	small	100%	6 of 6 (3 non-disease PolyQs)
☐	PHD	small	100%	5 of 5 (3 non-disease PolyQs)
☑	Bromodomain	a	100%	5 of 5 (4 non-disease PolyQs)
☐	WW	b	100%	4 of 4 (2 non-disease PolyQs)
☐	Pkinase	a+b	100%	4 of 4 (4 non-disease PolyQs)
☑	Sushi	small	100%	4 of 4 (1 non-disease PolyQs)
☐	RRM_1	a+b	100%	4 of 4 (2 non-disease PolyQs)
☐	HLH	a	100%	4 of 4 (4 non-disease PolyQs)
☐	Fork_head	a	100%	4 of 4 (4 non-disease PolyQs)
☐	Drf_FH1	none	100%	4 of 4 (1 non-disease PolyQs)
☑	BRCT	a/b	100%	4 of 4 (1 non-disease PolyQs)
☐	Helicase_C	a/b	100%	4 of 4 (4 non-disease PolyQs)

☐ Disease PolyQs

	Pfam Domain	SCOP Class	% total	Frequency
☐	Ion_trans	membrane and cell surface	100%	4 of 4 (1 disease PolyQs)
☐	UIM	none	100%	3 of 3 (1 disease PolyQs)
☐	DUF3652		100%	2 of 2 (1 disease PolyQs)
☐	TBP	a+b	100%	2 of 2 (1 disease PolyQs)
☑	Atrophin-1	none	100%	2 of 2 (1 disease PolyQs)
☐	PAM2	a	100%	1 of 1 (1 disease PolyQs)
☐	Ca_chan_IQ	none	100%	1 of 1 (1 disease PolyQs)
☐	Hormone_recep	a	100%	1 of 1 (1 disease PolyQs)
☑	Androgen_recep	none	100%	1 of 1 (1 disease PolyQs)
☐	zf-C4	small	100%	1 of 1 (1 disease PolyQs)
☑	SCA7	none	100%	1 of 1 (1 disease PolyQs)
☐	Josephin	none	100%	1 of 1 (1 disease PolyQs)
☐	HEAT		100%	1 of 1 (1 disease PolyQs)
☐	AXH	none	100%	1 of 1 (1 disease PolyQs)
☐	ATXN-1_C		100%	1 of 1 (1 disease PolyQs)
☐	LsmAD	none	100%	1 of 1 (1 disease PolyQs)

Figure 2. Selecting the 'Stats' menu text shows the entire database contents to be grouped into non-disease and disease causing proteins. To aid analysis specific entries can be selected (indicated by a tickbox), using the 'examine' button and grouped together.

analysis by indentifying protein variants/isoforms and using only the longest protein isoforms (which we termed 'master sequences'), therefore eliminating splice variants/protein fragments. Multiple variants/isoforms of each protein were crudely identified by comparing the protein sequence following the PolyQ chains. The original sequences were then subjected to the BLASTClust (20), FORCE (21), MCL (22) and HomoClust algorithms (23), and the variants/isoforms were adjusted as necessary. The crude identification used the 10 amino acids immediately after the PolyQ chain as a 'search string'; any sequence that had the 10 amino acids immediately following its own polyQ chain was presumed to have homology with that sequence. The homology groups were confirmed by analyzing the data using the above algorithms. This yielded a total of 128 master sequences, from an original data set of >700 polyQ-containing human protein sequences.

The database can be searched according to protein name, Pfam domain or sequence. The results of a typical search, shown in Figure 1A, show both a graphical summary (Figure 1A, top) and textual details (Figure 1A, bottom) according to sequence classification (see below). The graphical summary shows pie chart and bar chart representations of the results according to sequence classification (Figure 1A, top), Pfam domain occurrence (Figure 1B) and Pfam domain repetition (Figure 1C). Retrieved database entries are listed in table format with one row per protein, and three columns containing protein name (with links to the GenBank entry), Pfam domains, and protein sequence (with the polyQ region annotated), respectively. Homologs in the database can be included or excluded from the search. From this view, the domain and sequence context of the polyQ sequence can be identified and further interrogated. To aid analysis specific entries can be selected from the results (using the 'examine' button) and grouped together.

Sequence classification

The data are sorted and annotated according to the following sequence classifications: *N-Terminal PolyQs*—sequences where the first polyQ chain appears before all Pfam domains; *C-Terminal PolyQs*—sequences where the last polyQ chain appears after all Pfam domains; *Interdomain PolyQs*—sequences where the polyQ chains appear between the first Pfam domain and the last Pfam domain; *Mid Domain PolyQs*—sequences in which the polyQ chain appears in the middle of a Pfam domain, or overlaps a Pfam domain; *No Significant Domain PolyQs*—sequences that do not contain any significant Pfam domains; *Unclassified PolyQs*—sequences that did not fit into any of the above classifications. Each group is readily accessed using the tabs in the web page (Figure 1A). We have also further reduced the redundancy in the data by clustering sequence homologs, and have also tagged known disease proteins.

Domain occurrence, repeats and disease statistics

The website features pre-constructed pages that show the database entries sorted according to non-disease and disease-causing proteins respectively. This distinction is applied to the sequence classifications above, the domain occurrence (e.g. listing all domains, Figure 1B), and domain repeats (Figure 1C). This allows database entries to be grouped and examined according to whether the polyQ tracts are found in non-disease or disease-causing proteins (Figure 2).

CONCLUSIONS AND FUTURE DIRECTIONS

PolyQ is a valuable resource for theoreticians and experimentalists looking for insights into the context of PolyQ repeats in proteins and relationships with disease. Although the query tool allows searching across much of the database, we are developing a custom interface that will allow user-configurable queries against the whole data set as well as user customization of how the results are displayed. We are also adding the structural information [e.g. from the SCOP (24), CATH (25) and PDB databases (26)] to the resources such that the structural context of polyQ repeats can be investigated.

FUNDING

This work is supported by National Health and Medical Research Council and the Australian Research Council. S.P.B. and A.M.B. are NHMRC Senior Research Fellows. Funding for open access charge: National Health and Medical Research Council (Australia).

Conflict of interest statement. None declared.

REFERENCES

1. Perutz,M.F., Johnson,T., Suzuki,M. and Finch,J.T. (1994) Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl Acad. Sci. USA*, **91**, 5355–5358.
2. Chen,S., Berthelier,V., Hamilton,J.B., O'Nuallain,B. and Wetzel,R. (2002) Amyloid-like features of polyglutamine aggregates and their assembly kinetics. *Biochemistry*, **41**, 7391–7399.
3. Robertson,A.L., Horne,J., Ellisdon,A.M., Thomas,B., Scanlon,M.J. and Bottomley,S.P. (2008) The structural impact of a polyglutamine tract is location-dependent. *Biophys. J.*, **95**, 5922–5930.
4. Stefani,M. and Dobson,C.M. (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.*, **81**, 678–699.
5. DiFiglia,M., Sapp,E., Chase,K.O., Davies,S.W., Bates,G.P., Vonsattel,J.P. and Aronin,N. (1997) Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, **277**, 1990–1993.
6. Wellington,C.L., Ellerby,L.M., Hackam,A.S., Margolis,R.L., Trifiro,M.A., Singaraja,R., McCutcheon,K., Salvesen,G.S., Propp,S.S., Bromm,M. *et al.* (1998) Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *J. Biol. Chem.*, **273**, 9158–9167.
7. Ellerby,L.M., Andrusiak,R.L., Wellington,C.L., Hackam,A.S., Propp,S.S., Wood,J.D., Sharp,A.H., Margolis,R.L., Ross,C.A., Salvesen,G.S. *et al.* (1999) Cleavage of atrophin-1 at caspase site aspartic acid 109 modulates cytotoxicity. *J. Biol. Chem.*, **274**, 8730–8736.
8. Ellisdon,A.M., Thomas,B. and Bottomley,S.P. (2006) The two-stage pathway of ataxin-3 fibrillogenesis involves a polyglutamine-independent step. *J. Biol. Chem.*, **281**, 16888–16896.
9. Saunders,H.M. and Bottomley,S.P. (2009) Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins. *Protein Eng. Des. Sel.*, **22**, 447–451.
10. Robertson,A.L. and Bottomley,S.P. (2010) Towards the treatment of polyglutamine diseases: the modulatory role of protein context. *Curr. Med. Chem.*, **17**, 3058–3068.
11. Faux,N.G., Bottomley,S.P., Lesk,A.M., Irving,J.A., Morrison,J.R., de la Banda,M.G. and Whisstock,J.C. (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.*, **15**, 537–551.
12. Goto,J., Watanabe,M., Ichikawa,Y., Yee,S.B., Ihara,N., Endo,K., Igarashi,S., Takiyama,Y., Gaspar,C., Maciel,P. *et al.* (1997) Machado-Joseph disease gene products carrying different carboxyl termini. *Neurosci. Res.*, **28**, 373–377.

13. Padiath,Q.S., Srivastava,A.K., Roy,S., Jain,S. and Brahmachari,S.K. (2005) Identification of a novel 45 repeat unstable allele associated with a disease phenotype at the MJD1/SCA3 locus. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **133B**, 124–126.
14. Li,W., Serpell,L.C., Carter,W.J., Rubinsztein,D.C. and Huntington,J.A. (2006) Expression and characterization of full-length human huntingtin, an elongated HEAT repeat protein. *J. Biol. Chem.*, **281**, 15916–15922.
15. Ohshima,K., Kang,S. and Wells,R.D. (1996) CTG triplet repeats from human hereditary diseases are dominant genetic expansion products in Escherichia coli. *J. Biol. Chem.*, **271**, 1853–1856.
16. Sarkar,P.S., Chang,H.C., Boudi,F.B. and Reddy,S. (1998) CTG repeats show bimodal amplification in E. coli. *Cell*, **95**, 531–540.
17. Margolis,R.L., Abraham,M.R., Gatchell,S.B., Li,S.H., Kidwai,A.S., Breschel,T.S., Stine,O.C., Callahan,C., McInnis,M.G. and Ross,C.A. (1997) cDNAs with long CAG trinucleotide repeats from human brain. *Hum. Genet.*, **100**, 114–122.
18. Mizutani,A., Matsuzaki,A., Momoi,M.Y., Fujita,E., Tanabe,Y. and Momoi,T. (2007) Intracellular distribution of a speech/language disorder associated FOXP2 mutant. *Biochem. Biophys. Res. Commun.*, **353**, 869–874.
19. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
20. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
21. Wittkop,T., Baumbach,J., Lobo,F.P. and Rahmann,S. (2007) Large scale clustering of protein sequences with FORCE -a layout based heuristic for weighted cluster editing. *BMC Bioinformatics*, **8**, 396.
22. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
23. Chen,C.-Y., Chung,W.-C. and Su,C.-T. (2006) Exploiting homogeneity in protein sequence clusters for construction of protein family hierarchies. *Pattern Recogn.*, **39**, 2356–2369.
24. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
25. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
26. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.