Published online 10 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D305–D309***
doi:10.1093/nar/gkn869

AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse

Masafumi Shionyu[1], Akihiro Yamaguchi[1], Kazuki Shinoda[1], Ken-ichi Takahashi[1] and Mitiko Go[1,2,3,*]

[1]Department of Bioscience, Faculty of Bioscience, Nagahama Institute of Bio-Science and Technology, 1266 Tamura-cho, Nagahama, Shiga 526-0829, [2]Ochanomizu University, 2-1-1 Ohtsuka, Bunkyo-ku, Tokyo 112-8610 and [3]Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

Received August 11, 2008; Revised September 27, 2008; Accepted October 18, 2008

ABSTRACT

We have constructed a database, AS-ALPS (alternative splicing-induced alteration of protein structure), which provides information that would be useful for analyzing the effects of alternative splicing (AS) on protein structure, interactions with other biomolecules and protein interaction networks in human and mouse. Several AS events have been revealed to contribute to the diversification of protein structure, which results in diversification of interaction partners or affinities, which in turn contributes to regulation of bio-molecular networks. Most AS variants, however, are only known at the sequence level. It is important to determine the effects of AS on protein structure and interaction, and to provide candidates for experimental targets that are relevant to network regulation by AS. For this purpose, the three-dimensional (3D) structures of proteins are valuable sources of information; however, these have not been fully exploited in any other AS-related databases. AS-ALPS is the only AS-related database that describes the spatial relationships between protein regions altered by AS ('AS regions') and both the proteins' hydrophobic cores and sites of inter-molecular interactions. This information makes it possible to infer whether protein structural stability and/or protein interaction are affected by each AS event. AS-ALPS can be freely accessed at http://as-alps.nagahama-i-bio.ac.jp and http://genomenetwork.nig.ac.jp/as-alps/.

INTRODUCTION

An increasing number of experiments have now demonstrated the biological significance of the diversification of protein structure generated by alternative splicing (AS), which results in diversification of interactions with other bio-molecules (1,2). For example, Dscam, a surface receptor required for both neuronal wiring and immune responses, is subjected to AS, producing thousands of isoforms. X-ray structural analyses of, and binding assays on, the amino-terminal four ecto-domains of two distinct Dscam isoforms have revealed that a sequence-variable region generated by mutually exclusive AS between the two isoforms is directly involved in isoform-specific homophilic binding (2). There are, however, huge numbers of AS variants that remain to be structurally and functionally characterized. To empower such experimental analyses, it is important to link information associated with protein structure, interaction and network to each AS variant, and also to predict the effect of each AS event (3). For such purposes, we have constructed a database, AS-ALPS (alternative splicing-induced alteration of protein structure).

There already exist a number of databases dealing with AS, many of which contain supplementary information regarding functional domains, such as those in InterPro (4). Although such information at domain-level resolution could be useful, it is often insufficient to infer functional change by AS, because amino acid sequences altered by AS ('AS regions') are often smaller than the typical domain size (3): most AS regions have a comparable size with that of secondary structure, super-secondary structure or module (5–7). Therefore, AS events need to be analyzed in terms of inter-residue and inter-atomic interactions between AS regions and other remaining regions of the protein. The three-dimensional (3D) structures of proteins are valuable sources of information for inferring the effects of AS on protein structure and interaction. Among AS-related databases, only AS-ALPS and ProSAS (8) integrate information about 3D protein structures. ProSAS, however, does not provide any information on protein interaction sites based on 3D structures,

[*]To whom correspondence should be addressed. Tel: +81 3 5978 5100; Fax: +81 3 3941 5990; Email: go.mitiko@ocha.ac.jp

despite the importance of such information in detailed functional analysis of splice isoforms. This omission is probably because the main interest of ProSAS would be protein structure itself. Only AS-ALPS provides locations of interaction sites in the 3D structures of protein complexes with other proteins, nucleic acids or small ligand molecules, and also indicates whether AS regions include the residues that form such interaction sites. From this information, users can infer the effect of AS on protein interactions.

AS-ALPS also provides other unique and useful information. First, it provides the locations of hydrophobic cores, which are critical for protein structural stability, and can determine whether AS regions include residues forming these regions. From this information, users can infer the effect of AS on protein stability, which in turn helps to infer the effect of AS on protein interactions, i.e. to the extent that a stability change might alter such interactions. This information is not provided in any other AS-related databases, including ProSAS. Second, AS-ALPS provides links from AS-related transcripts to the protein network database KEGG (9). Users can then easily check which network and which node in the network can be influenced by AS. Third, AS-ALPS provides information regarding whether each AS variant mRNA can be subjected to nonsense-mediated mRNA decay (NMD), as predicted from the locations of premature termination codons (10). This information is only provided in one other AS-related database, H-DBAS (11). Taken together, these related data help users to select experimental targets from a huge collection of alternative transcripts, and to interpret the molecular mechanisms of AS-related network regulation.

DATABASE CONTENTS

AS regions

AS regions were detected as follows. First, a cluster of AS variant sequences was detected by aligning human (or mouse) full-length cDNA sequences from H-invDB (12) [or FANTOM (13)] and RefSeq (14) with the corresponding genomic DNA sequence stored at NCBI (Figure 1A). One of the AS variants, usually the most similar RefSeq entry to the counterpart in SwissProt (15), was defined as the reference transcript sequence. AS regions were then defined as regions of amino acid sequences that differ between the reference sequence and other AS variant sequences (Figure 1B). It should be noted that AS regions include not only amino acid sequence regions encoded by alternative exons but also those changed by frame shift due to AS. Appropriate comments were added to identify probably artifactual AS regions that were either derived from truncated terminal regions in incomplete transcripts or caused by inaccurate sequence alignment. The detected AS regions were also classified according to their sequence position and the type of change (insertion/deletion or substitution). In the case of substitution, we calculated the identity between the mutually replaced amino acid sequences, when they were comparable, allowing a measurement of the extent of the change.

Positional relationships of AS regions to structural domains

For each AS isoform, protein structural domains were assigned with the hidden Markov models in SUPERFAMILY (16). AS-ALPS provides positional relationships of AS regions to such structural domains (Figure 1B), from which users can determine whether each AS event changes a whole structural domain or just a part.

Positional relationships of AS regions to hydrophobic cores and interaction sites

The detailed procedures for annotating AS regions with 3D structural information were described in our previous paper (3). Briefly, a BLASTP homology search (17) against entries in PDB (18) was performed for each AS isoform. In the 3D structures for the hit PDB entries, hydrophobic cores and interaction sites were detected. Hydrophobic cores were defined as a spatial cluster of at least five amino acid residues meeting two conditions: (i) that solvent accessibility <0.2 and (ii) that >75% of surrounding atoms in contact with their side chain be carbon atoms (3). Interaction sites were defined as amino acid residues within 4 Å of other proteins, nucleic acids or ligands (19). AS-ALPS provides the positional relationships of AS regions to such hydrophobic cores and interaction sites (Figure 1C), from which users can infer the effects of AS on protein structure and interactions.

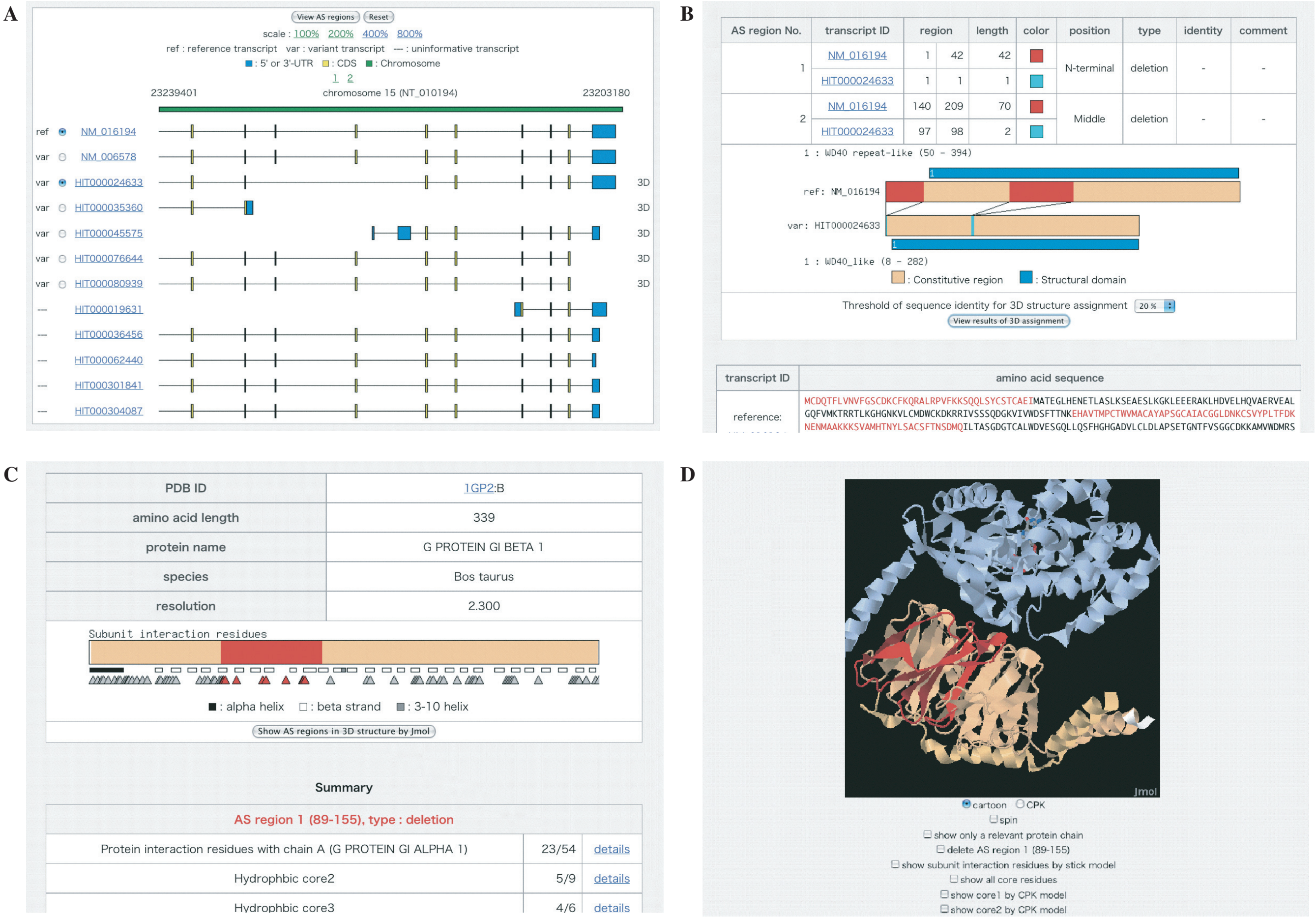
Possibility of NMD

AS-ALPS reports the possibility that each AS variant mRNA is subjected to NMD, as predicted from the locations of premature termination codons by the rule of Nagy and Maquat (10). It should be noted that some mRNAs have been experimentally demonstrated to be subjected to both NMD and translation (20). We have currently included even transcript sequences with a possibility of NMD in the analysis of AS regions.

ACCESS TO DATABASE

Search options

In the simple search form at the top of every page in AS-ALPS, either Transcript ID or Entrez Gene ID can be used as the search key. The advanced search page (Figure 2A) allows users to combine several search keys, such as species, protein name, attributes of AS regions (location, type and availability of 3D structure) and KEGG pathway name. If the 3D structure option is set to ‘available’, users can add search keys specifying whether AS regions include residues forming hydrophobic cores and what types of interaction sites the AS regions include. Search by amino acid sequence is also available, and is executed by the search engine FASTA (21). Users can also search by PDB ID. Search results are tabulated as in Figure 2B, where annotated information is briefly shown for each cluster of AS variants. The full list of


A
View AS regions
Reset
scale : 100% 200% 400% 800%
ref : reference transcript var : variant transcript --- : uninformative transcript
🟦 : 5' or 3'-UTR 🟨 : CDS 🟩 : Chromosome
1 2
23239401
chromosome 15 (NT_010194)
23203180
B
C
D
Jmol

ref	●	NM_016194	
var	○	NM_006578	
var	●	HIT000024633	3D
var	○	HIT000035360	3D
var	○	HIT000045575	3D
var	○	HIT000076644	3D
var	○	HIT000080939	3D
---		HIT000019631	
---		HIT000036456	
---		HIT000062440	
---		HIT000301841	
---		HIT000304087	

AS region No.	transcript ID	region		length	color	position	type	identity	comment
1	NM_016194	1	42	42	🟥	N-terminal	deletion	-	-
	HIT000024633	1	1	1	🟦				
2	NM_016194	140	209	70	🟥	Middle	deletion	-	-
	HIT000024633	97	98	2	🟦				

```
1 : WD40 repeat-like (50 - 394)
ref: NM_016194
var: HIT000024633
1 : WD40_like (8 - 282)
```

🟧 : Constitutive region 🟦 : Structural domain

Threshold of sequence identity for 3D structure assignment 20 %

View results of 3D assignment

transcript ID	amino acid sequence
reference:	MCDQTFLVNVFGSCDKCFKQRALRPVFKKSQQLSYCSTCAEIMATEGLHENETLASLKSEAESLKGKLEEERAKLHDVELHQVAERVEAL GQFVMKTRRTLKGHGNKVLCMDWCKDKRRIVSSSQDGKVIVWDSFTTNKEHAVTMPCTWVMACAYAPSGCAIACGGLDNKCSVYPLTFDK NENMAAKKKSVAMHTNYLSACSFTNSDMQILTASGDGTCALWDVESGQLLQSFHGHGADVLCLDLAPSETGNTFVSGGCDKKAMVWDMRS

PDB ID	1GP2:B
amino acid length	339
protein name	G PROTEIN GI BETA 1
species	Bos taurus
resolution	2.300

```
Subunit interaction residues
```

■ : alpha helix □ : beta strand ▨ : 3-10 helix

Show AS regions in 3D structure by Jmol

Summary

AS region 1 (89-155), type : deletion		
Protein interaction residues with chain A (G PROTEIN GI ALPHA 1)	23/54	details
Hydrophbic core2	5/9	details
Hydrophbic core3	4/6	details

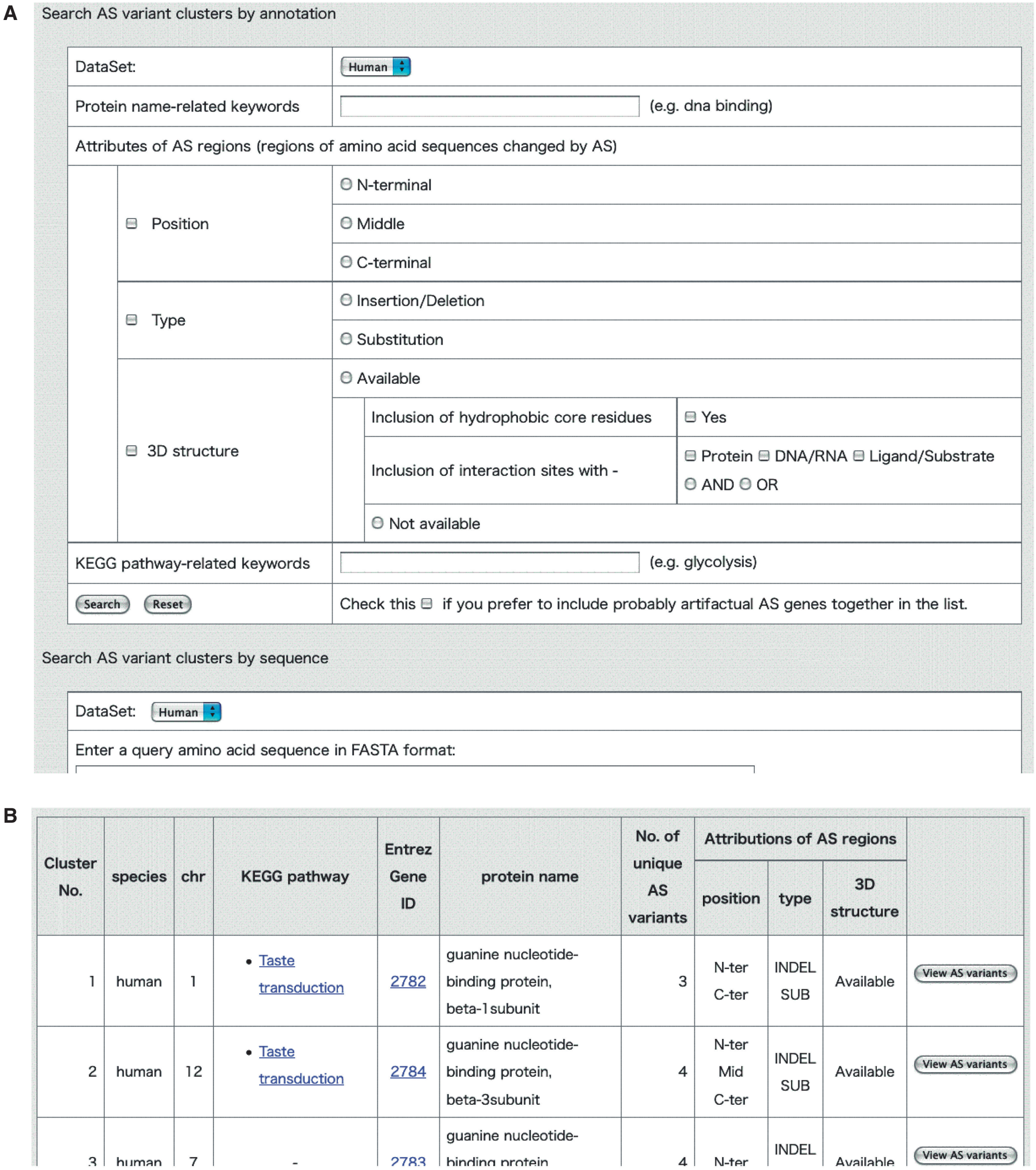
● cartoon ○ CPK
□ spin
□ show only a relevant protein chain
□ delete AS region 1 (89-155)
□ show subunit interaction residues by stick model
□ show all core residues
□ show core1 by CPK model
□ show core2 by CPK model

Figure 1. Data display screens. (**A**) A cluster of AS variant sequences (CDS in yellow and UTR in blue) aligned with a chromosome sequence (green). A '3D' mark in the rightmost column indicates that AS regions derived from the variant sequence in that row have 3D structure-related annotations. (**B**) Features of AS regions and their positional relationships to structural domains (blue). AS regions are colored by the type of change (deletion: red, insertion: cyan and substitution: violet). (**C**) Positional relationships of AS regions to hydrophobic cores and interaction sites detected in a relevant PDB entry. AS regions are shown in the same colors as in (B). Interaction sites are denoted with triangles. The positions of secondary structures are depicted with small open or filled boxes. (**D**) Spatial locations of AS regions in the 3D structure. A relevant protein chain with AS regions is shown in the same colors as in (C). Other interacting molecules are also shown in different colors.

genes subjected to AS can be browsed from the front page in the same format as that of the search result page, and also downloaded in tab-delimited format.

Data display screens

AS-ALPS has several data display screens for each type of information. First, clicking the 'View AS variants' button in the table of search results (Figure 2B) reveals the first screen, which shows a cluster of AS variants (Figure 1A). Next, after selecting one of variant transcripts, clicking the 'View AS regions' button leads to the second screen, which shows information regarding AS regions determined by comparing the amino acid sequences of the reference transcript and the user-selected variant (Figure 1B). AS regions are depicted in different colors depending on the type of change relative to the reference. Structural domains are also schematically represented with blue bars. In this example, one AS region of the deletion type (shown in red) is observed within the structural domain. Next, clicking the 'View results of 3D assignment' button leads to a list of assigned PDB entries accompanied by their amino acid sequence identities, as compared with the reference transcript. At that point, one of the PDB entries can be selected for viewing on the next screen, which shows the positional relationships of AS regions to hydrophobic cores and interaction sites detected in the selected PDB entry (Figure 1C). In this example, the AS region of the deletion type (shown in red) includes a part of the interaction sites (shown with triangles). Detailed information about interaction sites and hydrophobic cores is also provided at the bottom of the screen (Figure 1C). Next clicking the 'Show AS regions in 3D structure' button launches the Jmol 3D viewer (http://www.jmol.org/) (Figure 1D), showing the relevant protein chain in the same color as in Figure 1C, together with other interacting molecules. Spatial locations of hydrophobic cores can be viewed by using check boxes below the 3D viewer. The RasMol script for the 3D view can be downloaded. Branching off from the main route in AS-ALPS, the possibility of NMD can be checked in the page of detailed information on each transcript, which



Search AS variant clusters by annotation

DataSet:	Human		
Protein name-related keywords	(e.g. dna binding)		
Attributes of AS regions (regions of amino acid sequences changed by AS)			
□ Position	○ N-terminal		
	○ Middle		
	○ C-terminal		
□ Type	○ Insertion/Deletion		
	○ Substitution		
□ 3D structure	○ Available		
		Inclusion of hydrophobic core residues	□ Yes
		Inclusion of interaction sites with -	□ Protein □ DNA/RNA □ Ligand/Substrate ○ AND ○ OR
	○ Not available		
KEGG pathway-related keywords	(e.g. glycolysis)		
Search Reset	Check this □ if you prefer to include probably artifactual AS genes together in the list.		

Search AS variant clusters by sequence

DataSet: Human

Enter a query amino acid sequence in FASTA format:

Cluster No.	species	chr	KEGG pathway	Entrez Gene ID	protein name	No. of unique AS variants	Attributions of AS regions			
							position	type	3D structure	
1	human	1	• Taste transduction	2782	guanine nucleotide-binding protein, beta-1subunit	3	N-ter C-ter	INDEL SUB	Available	View AS variants
2	human	12	• Taste transduction	2784	guanine nucleotide-binding protein, beta-3subunit	4	N-ter Mid C-ter	INDEL SUB	Available	View AS variants
3	human	7	-	2783	guanine nucleotide-binding protein	4	N-ter	INDEL	Available	View AS variants

Figure 2. The advanced search page (**A**) and an example of a table summarizing search results (**B**).

appears when the user clicks a transcript ID in the AS variant cluster page (Figure 1A).

Hyperlinks to external databases

AS-ALPS provides hyperlinks to several external databases, to allow easy access to related information. Each cluster of AS variants, which are assumed to be derived from the same gene, is linked to the relevant entry in Entrez Gene (22, 23) and to relevant protein networks in KEGG (9) (Figure 2B). On the linked KEGG pathway map, the query protein and other alternatively spliced proteins belonging to the same pathway are highlighted in yellow and pink, respectively. In the detailed information page of each transcript, transcript and genome sequence IDs are linked to the corresponding entries in their source databases. Ligands are linked to relevant entries in Het-PDB Navi (19).

CURRENT STATISTICS AND FUTURE DEVELOPMENTS

Detailed statistics of the source data and derived AS-ALPS data are reported on the statistics page accessed from the front page of AS-ALPS. Here, we present some statistics derived from the data set excluding probably artifactual AS regions. The current version of AS-ALPS, version 1.0, contains 11 926 genes subjected to AS ('AS genes') and 31 928 unique AS regions in human, and 7666 AS genes and 13 893 unique AS regions in mouse. The 3D structural information is available for

5067 (42.5%) and 2675 (34.9%) of AS genes in human and mouse, respectively, and is available for 10 056 (31.5%) and 3676 (26.5%) of unique AS regions in human and mouse, respectively. The fraction of AS genes with the 3D data in human (42.5%) has got much larger than that in our previous paper (about 7%) (3), although the procedure of analysis is not exactly the same as the previous one. Transcript data in AS-ALPS will be updated to fully recalculate the database contents at least annually if the source data have substantially changed. Structure data from PDB in AS-ALPS will be updated at least quaterly. To be able to follow the frequent updates of PDB, we are planning to automate AS-ALPS data updates. To enhance usability for researchers, we are also planning to add an option that allows user-uploaded transcript data to be included in the process of determining AS regions and annotating them with 3D structure information.

FUNDING

Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). Funding for open access charge: Genome Network Project from MEXT.

Conflict of interest statement. None declared.

REFERENCES

1. Stetefeld,J. and Ruegg,M.A. (2005) Structural and functional diversity generated by alternative mRNA splicing. *Trends Biochem. Sci.*, **30**, 515–521.
2. Meijers,R., Puettmann-Holgado,R., Skiniotis,G., Liu,J.H., Walz,T., Wang,J.H. and Schmucker,D. (2007) Structural basis of Dscam isoform specificity. *Nature*, **449**, 487–491.
3. Yura,K., Shionyu,M., Hagino,K., Hijikata,A., Hirashima,Y., Nakahara,T., Eguchi,T., Shinoda,K., Yamaguchi,A., Takahashi,K. *et al.* (2006) Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene*, **380**, 63–71.
4. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
5. Go,M. (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature*, **291**, 90–92.
6. Noguti,T., Sakakibara,H. and Go,M. (1993) Localization of hydrogen-bonds within modules in barnase. *Proteins*, **16**, 357–363.
7. Shinoda,K., Takahashi,K. and Go,M. (2007) Retention of local conformational compactness in unfolding of barnase: contribution of end-to-end interactions within quasi-modules. *Biophysics*, **3**, 1–12.
8. Birzele,F., Kuffner,R., Meier,F., Oefinger,F., Potthast,C. and Zimmer,R. (2008) ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.*, **36**, D63–D68.
9. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
10. Nagy,E. and Maquat,L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.*, **23**, 198–199.
11. Takeda,J., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, **35**, D104–D109.
12. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
13. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
14. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
15. UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
16. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. 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.
19. Yamaguchi,A., Iida,K., Matsui,N., Tomoda,S., Yura,K. and Go,M. (2004) Het-PDB Navi.: a database for protein-small molecule interactions. *J. Biochem.*, **135**, 79–84.
20. Dreumont,N., Maresca,A., Boisclair-Lachance,J.F., Bergeron,A. and Tanguay,R.M. (2005) A minor alternative transcript of the fumarylacetoacetate hydrolase gene produces a protein despite being likely subjected to nonsense-mediated mRNA decay. *BMC Mol. Biol.*, **6**, 1.
21. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
22. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
23. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.