PROTEOMICS

Supporting Information for Proteomics

DOI 10.1002/pmic.200700741

Martin Irmler, Daniela Hartl, Thorsten Schmidt, Johannes Schuchhardt, Christiane Lach, Helmut E. Meyer, Martin Hrabé de Angelis, Joachim Klose and Johannes Beckers

An approach to handling and interpretation of ambiguous data in transcriptome and proteome comparisons

© 2008 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Supplementary data

Table S1. Statistical significantly regulated protein spots. Only proteins detected in several spots are shown.

Gene symbol	Spot ID	Linear ratio (E13.5 vs e9.5)	opposing ratios	Protein name	gI accession
Apoa1	SID1434	7.14		Apolipoprotein A-I [Precursor]	gi:15126646
Apoa1	SID1427	2.04		apolipoprotein A-I precursor - mouse	gi|74146433
Dpysl2	SID415	5.26		Dihydropyrimidinase-related protein 2	gi|40254595
Dpysl2	SID2789	2.56		Dihydropyrimidinase-related protein 2	gi|40254595
Dpysl2	SID457	2.33		Dihydropyrimidinase-related protein 2	gi|40254595
Eef1b2	SID1217	0.3		Elongation factor 1-beta	gi|5902663
Eef1b2	SID1213	0.4		Elongation factor 1-beta	gi|5902663
Eef1d	SID1070	0.49		eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein	gi|54287684
Eef1d	SID1126	0.5		Elongation factor 1-delta	gi|12856949
Eef1d	SID1069	0.47		eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein	gi|12856949
Etfa	BID964	0.59	*	Electron transferring flavoprotein, alpha polypeptide	gi|66911229
Etfa	BID968	1.2	*	Electron transfer flavoprotein subunit alpha, mitochondrial [Precursor]	gi|13097375
Hnrpa2b1	BID867	1.39		Heterogeneous nuclear ribonucleoproteins A2/B1	gi|23266713
Hnrpa2b1	BID866	1.79		Heterogeneous nuclear ribonucleoproteins A2/B1	GI:124028629
Hnrpa2b1	BID791	1.12		Heterogeneous nuclear ribonucleoproteins A2/B1	gi|23266713
Hspa4	SID200	0.52		Heat shock 70 kDa protein 4	gi|2495342
Hspa4	SID201	0.47		Heat shock protein 4	gi|13277753
Hspa4	SID3997	0.65		Heat shock protein 4	gi|31982275
Hspd1	SID571	0.46	*	heat shock protein, 60 kDa	gi|31981679
Hspd1	SID654	0.3	*	heat shock protein, 60 kDa	gi|51452
Hspd1	SID576	1.2	*	heat shock protein, 60 kDa	gi|31981679
Hspd1	SID619	0.69	*	60 kDa heat shock protein, mitochondrial [Precursor]	gi|51452
Marcks	SID319	1.33		Myristoylated alanine-rich C-kinase substrate	gi|6678768
Marcks	SID306	2.17		Myristoylated alanine-rich C-kinase substrate	gi|6678768
Naca	SID966	0.43		Nascent polypeptide-associated complex subunit alpha	gi|41350312
Naca	SID968	0.47		Nascent polypeptide-associated complex subunit alpha	gi|71151989
Naca	SID951	0.45		Nascent polypeptide-associated complex subunit alpha	gi|41350312
Nmral1	SID3938	1.67		NmrA-like family domain containing 1	gi|13378302
Nmral1	BID1008	1.22		NmrA-like family domain containing 1	gi|24431937
Prdx2	SID1495	1.61		Prdx2 protein	gi|3603241
Prdx2	SID1483	1.52		Prdx2 protein	gi|3603241
Stmn1	SID1654	2.5	*	Stathmin	gi|14625464
Stmn1	SID1632	2.7	*	Stathmin	gi|14625464
Stmn1	SID1646	2.5	*	Stathmin	gi|9789995
Stmn1	SID1644	2.44	*	Stathmin	gi|14625464
Stmn1	SID1639	0.42	*	Stathmin	gi|9789995
Stmn1	SID1631	2.94	*	Stathmin	gi|9789995
Tpt1	SID1387	0.43		Translationally-controlled tumor protein	gi|16758734
Tpt1	SID1386	0.47		Translationally-controlled tumor protein	gi|16758734
Tubb5	SID999	0.33	*	Tubulin beta-5 chain	gi|7106439
Tubb5	SID1999	2.13	*	Tubulin beta-5 chain	gi|27465535
Uchl1	SID3172	3.03		Ubiquitin carboxyl-terminal hydrolase isozyme L1	gi|6755929
Uchl1	SID1379	3.13		Ubiquitin carboxyl-terminal hydrolase isozyme L1	gi|25058057

Table S2. ChipInspector (Genomatix) analysis of genes represented by two statistical significant transcripts. Statistical analysis was done using default settings and the transcript of each gene with the highest coverage (number of significant oligos) was selected. If opposing transcripts were detected, the two transcripts with the highest coverage were selected (up- and down-regulated, one of each).

GeneName	ID transcript 1	Transcript 1 linear ratio E13.5 vs E9.5	ID transcript 2	Transcript 2 linear ratio E13.5 vs E9.5	Difference between transcripts
Elavl1	AK028769	1.98	AK078333	0.39	3'- UTR
Hspa4	AK054211	1.77	AK133772	0.25	alternative splicing
Mtpn	AK146305	0.54	AK146305	1.65	3'- UTR
Dbi	AK148753	0.58	AK148753	2.07	3'- UTR+ alternative splicing
Cotl1	AK148854	0.54	AK148854	2.4	3'- UTR
Vim	AK131618	0.53	AK151196	1.64	3'- UTR
Marcks	AK017779	1.84	AK151646	0.52	alternative splicing
Sh3bgrl	AK151501	1.75	AK151871	0.51	3'- UTR
Idh3a	AK032787	0.55	AK160054	1.96	3'- UTR
Ywhag	AK164696	0.45	AK164696	1.71	3'- UTR
Pafah1b2	AK012164	1.52	NM_008775	0.57	3'- UTR
Ptbp1	AK172712	0.45	NM_008956	1.88	3'- UTR
Nudt5	AK018345	1.68	NM_016918	0.41	alternative splicing
Ssbp1	AK013425	1.94	NM_212468	0.5	alternative splicing

Table S3. Enrichment of transport-related GO-terms among up- or down-regulated proteins analysed with A) BiblioSphere (Genomatix [1]) or B) FunCat [2]. A) BiblioSphere was used to search the GO database for enriched terms in the sets of up-or down-regulated proteins. The two sets were analysed separately and z-scores>2 indicate significant enrichment. B) For FunCat, the sets of up- and down-regulated proteins were compared directly and the number (No) of proteins in each analysis is provided. p-values<0.1 may indicate statistical significant enrichment.

A. Bibliosphere analysis	GO Term	GO ID	Observed	Expected	Z-score
up-regulated proteins	transport	GO:0006810	13	7.75	2.03
down-regulated proteins	transport	GO:0006810	6	6.12	-0.05

B. FunCat analysis	FunCat term	GO mapping	No. in up-regulated / No total	No in down-regulated / No total	p-value
up- versus down-regulated proteins	FC:20 (cellular transport, transport facilitation/routes)	GO:0006886	13/57	3/42	0.055

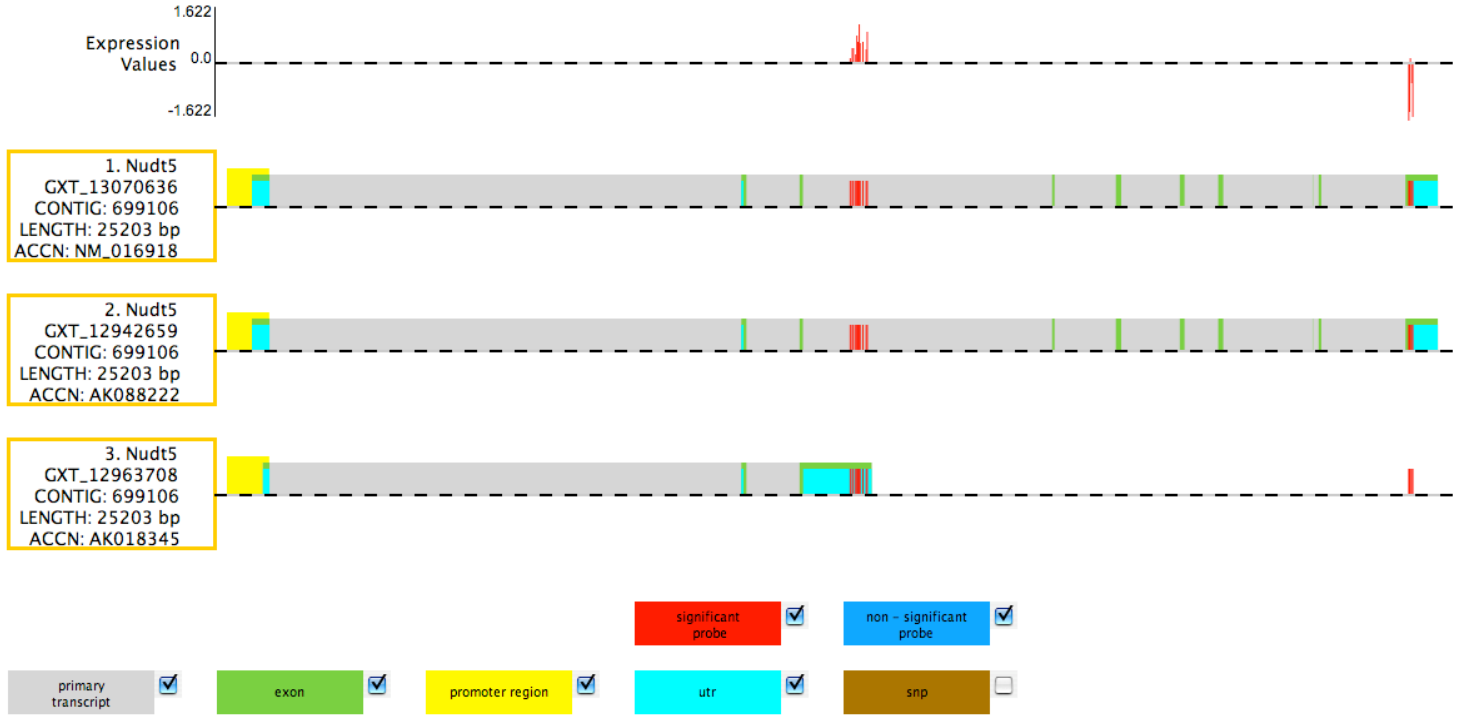


Fig. S1. ChipInspector (Genomatix) diagram showing significant probes detected for the Nudt5 gene. The top line indicates up- and down-regulation (log2 values) of individual oligos, represented by red bars. Expression of a shorter transcript is increased whereas the full-length transcript is down-regulated.

Supplementary Materials and Methods

Embryonic tissue collection

Pregnant C57Bl/6 mice were sacrificed according to the German Animal Protection Act to obtain embryonic mouse brain tissue. Heads of embryos were cut following a straight line delineated by the following two positions: *post otocyst* until just past the *mandible anlage*. Embryonic head tissue was shock frozen in liquid nitrogen and stored at -80°C until protein or mRNA extraction was performed. Embryos were staged according to somite number and morphological criteria. For proteome analysis, two pooled samples containing 12 embryo heads each were prepared for E9.5 and E13.5 respectively.

RNA isolation

Total RNA was isolated employing the Lipid Rich Tissue Mini kit (Qiagen). Briefly, Qiazol reagent was added to frozen tissue, followed immediately by Ultraturrax homogenization. After purification the RNA was incubated with DNase (turbo DNA-free, Ambion) and RNA quality was assessed using the RNA 600 Pico Assay (Agilent).

Expression profiling

High quality total RNA (1 µg) was amplified with the MessageAmp II-Biotin Enhanced Single Round aRNA Amplification Kit (Ambion) according to manufacturer's instructions. 10 µg of amplified aRNA were hybridized on Affymetrix MOE430 2.0 GeneChips® containing about 45,000 probe sets. Staining and scanning was done according to the Affymetrix expression protocol. In total, 16 embryo heads distributed over three developmental stages (6 heads at E9.5 and E13.5) were analyzed. Our array data were submitted to GEO (GSE8091).

Statistical transcriptome analysis

Statistical analysis of the microarrays was performed by utilizing the statistical programming environment R (R Development Core Team [3]), using packages available from http://cran.r-project.org and mips.gsf.de/mips/staff/mader/software. Briefly, probe set summaries were calculated according to RMA. A log-scale transformation (log, basis 2) using normalization against mean of all E9.5 samples employing the loess smoother [4] operating on M-A scale was carried out. Genewise testing for differential expression with Welch's *t*-test and Benjamini-Hochberg multiple testing correction (FDR < 0.05), and application of non-stringent filters for basal expression and fold-change (mean ≥ 100, ratios of means ≥ 1.5, both on linear scale) was used.

Statistical analysis with the ChipInspector software (Genomatix) was used to analyse ambiguous transcript expression (Table S2, Fig. S1), since it allows the exact mapping of probes to exons or specific regions in the 3'-UTR. This analysis would not have been possible with conventional probe set-based tools, since probe sets can span larger transcript regions or be distributed across several exons. ChipInspector can not be used for analyses of absolute expression levels (Fig. 1), since it does not provide absolute expression values. Standard settings for analysis were used (FDR<0.05) and the transcript of each gene with the highest coverage (number of significant oligos) was selected. If transcripts with opposing regulation were detected, the two transcripts with the highest coverage were selected (up- and down-regulated, one of each).

Protein Extraction and Labeling

Protein extracts were prepared from individual embryo heads and pooled in two samples. Due to small sample size, E9.5 tissue material was pooled before protein extraction. Our published protein extraction procedure [5] was slightly altered. Briefly,

tissue was ground to fine powder in a sample tube pre-cooled in liquid nitrogen and only one part (v/w) of sample buffer (50 mM TRIZMA Base (Sigma-Aldrich, Steinheim, Germany), 50 mM KCl and 20% (w/v) glycerol at pH 7.5) was added. Samples were labeled by Cy3 minimal dye (GE Healthcare, Munich, Germany). An E11.5 pool sample aliquot was used as internal standard and labeled by Cy5 minimal dye (GE Healthcare). Labeling was carried out according to manufacturer's instructions adding 400 pmol fluorescent dye per 50 µg of protein. Each Cy3-labeled sample was mixed with the same amount of internal standard. The protein extracts were mixed with 70 mM dithiothreitol (Biorad, Munich, Germany), 2% (v/w) of ampholyte mixture Servalyte pH 2-4 (Serva, Heidelberg, Germany) and stored at -80 °C. For protein spot counting in 2-D gels, saturation dye labeled E9.5 and E11.5 samples were utilized (Cy3, GE Healthcare).

Two-Dimensional Gel Electrophoresis (2-DE)

Protein samples were separated by the large-gel 2-DE technique developed in our laboratory as described previously [6]. The gel format was 40 cm (isoelectric focusing) x 30 cm (SDS-PAGE) x 1.0 mm (gel width). For visualization of protein spots, gels were scanned at a resolution of 100 µm using a 532 nm and 633 nm laser for Cy3 and Cy5 labeled samples respectively (Typhoon 9400, GE Healthcare).

Spot evaluation procedure

Protein spot patterns were evaluated by *Delta2D* imaging software version 3.4 (Decodon, Greifswald, Germany). Briefly, spot patterns of in-gel standards were matched using *Delta2D* "exact" mode matching protocol. Using "union" mode a fusion image was generated, including the visible spots from all 2D gels within the project. The fusion image was used for spot detection (settings: local background region 100, average spot size 1, sensitivity 100%) followed by manual spot editing. A spot pattern containing 2004 protein spots was transferred from the fusion image to

all other 2-D gel images to make sure that spot identification number for each spot is the same for every gel in the project. Relative spot volume intensities (fractions of 100%) were used for quantitative protein expression analysis. After background extraction, normalized values were exported in spreadsheet format from *Delta2D* for statistical analysis. Data sets were analyzed applying a t-test analysis, including Benjamini-Hochberg FDR correction ($p<0.1$). Sample size comprised n = 6 for E9.5 and n = 5 for E13.5.

Protein Identification

For protein identification by mass spectrometry, 40 μL extract was separated by 2-DE and stained using a MS-compatible silver staining protocol [7]. In order to assign corresponding protein spots from analytical fluorescent to quantitative silver stained 2-D gels, spot patterns were matched using *Delta2D*. Protein spots of interest were excised from 2-D gels and subjected to in-gel tryptic digestion. Peptides were analyzed by a Reflex 4 MALDI-TOF mass spectrometer (Bruker Daltonics, Bremen Germany) as described previously [7]. Alternatively, ESI-MS and -MS/MS on a LCQ Deca XP ion trap (Thermo Finnigan, Waltham, MA) was applied. Mass spectra were analyzed using our in-house MASCOT software package license version 2.1 automatically searching NCBI databases. MS/MS ion search was performed with this set of parameters: (i) taxonomy: *Mus musculus*, (ii) proteolytic enzyme: trypsin, (iii) maximum of accepted missed cleavages: 1, (iv) mass value: monoisotopic, (v) peptide mass tolerance 0.8 Da, (vi) fragment mass tolerance: 0.8 Da, and (vii) variable modifications: oxidation of methionine and acrylamide adducts (propionamide) on cysteine. Only proteins with scores corresponding to $p<0.05$, with at least two peptides identified in two independent identifications were considered. Furthermore, theoretical and practical molecular weight and pI for each protein identified by database search were compared. Our results are summarized in Sup. Table 6.

Analysis of correlation and biological functions

Official gene symbols (MGI, http://www.informatics.jax.org/) were used to match transcriptome and proteome datasets. To match proteins and transcripts at the level of expression signals (for Fig. 1), all protein spot values belonging to the same gene were summed up and for transcripts the probe set with the highest expression value was used. To match proteins and transcripts at the level of expression ratios, protein spots belonging to the same gene were summed up and ratios were calculated; for transcripts, ratios of probe set belonging to the same gene were averaged (Fig. 2). For ChipInspector analysis, all selected transcripts (see section above for statistics) were matched to their corresponding protein using official gene symbols (MGI). Enrichment of transport-related GO-terms among up- or down-regulated proteins was analysed with BiblioSphere (Genomatix [1]) or FunCat [2].

PROMPT features

Functional roles of gene products were described in terms of the manually curated hierarchical functional catalogue (FunCat) developed at MIPS [2]. In this catalogue each of the 16 main classes (e.g., metabolism, energy) may contain up to six subclasses. An essential feature of FUNCAT is its multidimensionality, meaning that any protein can be assigned to multiple categories. Carefully verified FunCat assignment as well as enzyme classifications, InterPro domains [9, 10] and genomic (e.g. number of exons) as well as sequence information was taken from the Mouse Functional Genome Database (MfunGD) [11]. Gene Ontology (GO) functional annotations [12], Mammalian Phenotype (MP) classification [13] and the currently known number of single nucleotide polymorphisms (SNPs) were obtained from The Mouse Genome Database (MGD) [14] as of June 2007.

Disorder, low complexity, probability to be non-globular or adopt a coiled-coiled structure, similarity to clusters of orthologous groups (COGs) [15, 16] and protein

structural information were taken from our PEDANT database *Mus_musculus_new* [17-19]. Disorder attributes were calculated with the software GlobPlot [20]. GlobProt utilizes the statistics of proteins known to have unstructured regions [21, 22]. Low complexity and non-globular regions were computed with the original SEG algorithm [23]. The probability that a protein structure adopts a coiled-coiled structure was calculated with the program COILS program [24]. Protein structures were further analyzed with regard of their secondary structure content i.e. percentage of residues that are part of alpha helices, beta sheets or coils and their three dimensional structure folds using the SCOP classification schema. Like FunCat the SCOP database [25, 26] provides a hierarchical classification of protein structural domains. Data retrieval from PEDANT was accomplished using the protein mapping and comparison tool PROMPT version 0.9.2 [27]. SCOP class and COG assignments to *M. musculus* proteins were based on BLAST [28] hits with an E-value of 10^{-4} (PROMPT default thresholds).

The number of alternating hydrophobic/hydrophilic stretches was computed as described [29] using the implementation of PROMPT. The residues A, C, F, G, I, L, M, P, V, W and Y were considered to be hydrophobic and H, Q, N, S, T, K, R, D, E were considered hydrophilic in this thereby. The hydrophobicity of a protein was defined as $\frac{\sum_{i=1}^{n} H_i}{n}$, with H_i denoting the hydrophobicity value of the amino acid at position *i* of a protein of *n* amino acids. Hydrophobicity values were calculated using the Kyte-Doolittle scale [30]. Molecular protein mass, theoretical isoelectric point (pI) as well the averaged hydrophobicity were calculated with the PROMPT software from the amino acid sequences.

PROMPT statistics

All statistical tests and most figures were done using the R software package version 2.5.0 (www.r-project.org) and PROMPT [27]. To compare the distributions of two

unpaired samples with non-Gaussian or unknown distributions, the rank-sum Mann-Whitney (MW) test and the two sample Kolmogorov-Smirnov (KS) were applied using the significance threshold α=0.1. The null hypothesis of the Mann-Whitney test is that the abundance means are equal. The null hypothesis of the Kolmogorov-Smirnov test is that the values of the two samples are drawn from the same continuous distributions. Both tests have the advantage that they make almost no assumptions about the distribution of data.

References for supplementary data

[1] Deng, X., Stachlewitz, R. F., Liguori, M. J., Blomme, E. A., *et al.*, Modest inflammation enhances diclofenac hepatotoxicity in rats: role of neutrophils and bacterial translocation. *J Pharmacol Exp Ther* 2006, *319*, 1191-1199.
[2] Ruepp, A., Zollner, A., Maier, D., Albermann, K., *et al.*, The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 2004, *32*, 5539-5545.
[3] Team, R. D. C., R: A language and environment for statistical computing. 2005.
[4] Cleveland, W. S., Grosse, E., and Shy, W. M., *Statistical Models in S, chapter Local regression models. Chapter 8*, 1992.
[5] Zabel, C., Klose, J., Protein extraction for 2-D electrophoresis. *Methods Mol Biol* 2007, *in press*.
[6] Zabel, C., Klose, J., High resolution large gel 2-D electrophoresis. *Methods Mol Biol* 2007, *in press*.
[7] Nebrich, G., Herrmann, M., Sagi, D., Klose, J., Giavalisco, P., High MS-compatibility of silver nitrate-stained protein spots from 2-DE gels using ZipPlates and AnchorChips for successful protein identification. *Electrophoresis* 2007, *28*, 1607-1614.
[8] Rainer, J., Sanchez-Cabo, F., Stocker, G., Sturn, A., Trajanoski, Z., CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res* 2006, *34*, W498-503.
[9] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., *et al.*, InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform* 2002, *3*, 225-235.
[10] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., *et al.*, New developments in the InterPro database. *Nucleic Acids Res* 2007, *35*, D224-228.
[11] Ruepp, A., Doudieu, O. N., van den Oever, J., Brauner, B., *et al.*, The Mouse Functional Genome Database (MfunGD): functional annotation of proteins in the light of their cellular context. *Nucleic Acids Res* 2006, *34*, D568-571.
[12] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, *25*, 25-29.
[13] Smith, C. L., Goldsmith, C. A., Eppig, J. T., The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* 2005, *6*, R7.

[14] Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., *et al.*, The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res* 2005, *33*, D471-475.
[15] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., *et al.*, The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003, *4*, 41.
[16] Tatusov, R. L., Koonin, E. V., Lipman, D. J., A genomic perspective on protein families. *Science* 1997, *278*, 631-637.
[17] Frishman, D., Mokrejs, M., Kosykh, D., Kastenmuller, G., *et al.*, The PEDANT genome database. *Nucleic Acids Res* 2003, *31*, 207-211.
[18] Riley, M. L., Schmidt, T., Artamonova, II, Wagner, C., *et al.*, PEDANT genome database: 10 years online. *Nucleic Acids Res* 2007, *35*, D354-357.
[19] Riley, M. L., Schmidt, T., Wagner, C., Mewes, H. W., Frishman, D., The PEDANT genome database in 2005. *Nucleic Acids Res* 2005, *33*, D308-310.
[20] Linding, R., Russell, R. B., Neduva, V., Gibson, T. J., GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 2003, *31*, 3701-3708.
[21] Tompa, P., Intrinsically unstructured proteins. *Trends Biochem Sci* 2002, *27*, 527-533.
[22] Wright, P. E., Dyson, H. J., Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999, *293*, 321-331.
[23] Wootton, J. C., Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 1994, *18*, 269-285.
[24] Lupas, A., Van Dyke, M., Stock, J., Predicting coiled coils from protein sequences. *Science* 1991, *252*, 1162-1164.
[25] Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., *et al.*, SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 2004, *32*, D226-229.
[26] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995, *247*, 536-540.
[27] Schmidt, T., Frishman, D., PROMPT: a protein mapping and comparison tool. *BMC Bioinformatics* 2006, *7*, 331.
[28] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., Basic local alignment search tool. *J Mol Biol* 1990, *215*, 403-410.
[29] Wong, P., Fritz, A., Frishman, D., Designability, aggregation propensity and duplication of disease-associated proteins. *Protein Eng Des Sel* 2005, *18*, 503-508.
[30] Kyte, J., Doolittle, R. F., A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, *157*, 105-132.