Published online 1 November 2009 Nucleic Acids Research, 2010, Vol. 38, Database issue **D497–D501**
doi:10.1093/nar/gkp914

CORUM: the comprehensive resource of mammalian protein complexes—2009

Andreas Ruepp[1,*], Brigitte Waegele[1,2], Martin Lechner[1], Barbara Brauner[1], Irmtraud Dunger-Kaltenbach[1], Gisela Fobo[1], Goar Frishman[1], Corinna Montrone[1] and H.-Werner Mewes[1,2]

[1]Institute for Bioinformatics and Systems Biology (IBIS), Helmholtz Zentrum München—German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, D-85764 Neuherberg and [2]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Received September 10, 2009; Accepted October 7, 2009

ABSTRACT

CORUM is a database that provides a manually curated repository of experimentally characterized protein complexes from mammalian organisms, mainly human (64%), mouse (16%) and rat (12%). Protein complexes are key molecular entities that integrate multiple gene products to perform cellular functions. The new CORUM 2.0 release encompasses 2837 protein complexes offering the largest and most comprehensive publicly available dataset of mammalian protein complexes. The CORUM dataset is built from 3198 different genes, representing ∼16% of the protein coding genes in humans. Each protein complex is described by a protein complex name, subunit composition, function as well as the literature reference that characterizes the respective protein complex. Recent developments include mapping of functional annotation to Gene Ontology terms as well as cross-references to Entrez Gene identifiers. In addition, a 'Phylogenetic Conservation' analysis tool was implemented that analyses the potential occurrence of orthologous protein complex subunits in mammals and other selected groups of organisms. This allows one to predict the occurrence of protein complexes in different phylogenetic groups. CORUM is freely accessible at (http://mips.helmholtz-muenchen.de/genre/proj/corum/index.html).

INTRODUCTION

Major cellular processes like cell cycle, protein folding and protein degradation depend on the activity of protein complexes (1). To date there are no reliable estimates about the total number of protein complexes in cells (complexome), but data from single cell organisms provide evidence, that more than half of the gene products are involved in the formation of protein complexes (2). In the advent of protein network analyses, topological properties of protein complexes resulted in paraphrases such as 'party hubs' (3) or 'multi-interface hubs' (4). Bioinformatics analysis of protein–protein interaction (PPI) datasets revealed that protein complex subunits are stronger evolutionary conserved and show a higher essentiality than proteins from other interactions (4).

As the most comprehensive PPI and protein complex data are available for *Saccharomyces cerevisiae*, most of these discoveries were obtained using data from yeast. In addition to a manually curated dataset of protein complexes (5), tag-based high-throughput approaches were performed in order to define the yeast complexome (6,7). The importance of manually curated gold-standards was demonstrated by analyses of results from high-throughput experiments. In an assessment of different high-throughput technologies for the analysis of PPIs it was shown, that each method, depending on its physiochemical constraints, captures interactions for different subsets of proteins (8). Thus, none of the existing methods is able to detect all interactions and it was also shown that even the combined dataset of five different methods missed ∼40% of experimentally validated, manually curated interactions (9).

For mammals no comprehensive high-throughput dataset of protein complexes is publicly available. Bioinformatics analyses of the mammalian complexome can be performed either by using artificially constructed protein complexes (10) or data from manually curated datasets (11,12). In 2008, the CORUM database was introduced as the most comprehensive catalogue of mammalian protein complexes. All data are manually curated

[*]To whom correspondence should be addressed. Tel: +49 3187 3189; Fax: +49 3187 3585; Email: andreas.ruepp@helmholtz-muenchen.de

including information of protein complex subunits and methods of purification as well as additional information such as functional annotation using the Functional Catalogue (FunCat) annotation scheme (13), stoichiometry of the subunits and information about association with diseases (14). Analyses of the CORUM dataset have shown (i) that mammalian protein complexes are most frequently composed of 3 or 4 different subunits and (ii) that proteins tend to be reused in up to 53 protein complexes (15).

The CORUM dataset has been used for a number of bioinformatics analyses like tissue-specific expression of proteins (16), functional interpretation of high-throughput data (17–19) or to predict interactions of protein regions (20). In addition, the dataset contributes to web-based applications like the DICS database of functional modules (21) or the COFECO tool for composite function annotation (22).

The CORUM Release 2.0 presents a significantly extended dataset that now consists of 2837 mammalian protein complexes. In addition to existing cross-references the dataset was mapped to Entrez Gene identifiers and functional annotation of Gene Ontology (GO) terms. In order to enable more specific search results in comments, the content is now distributed into the three sections ‘Disease Comment’, ‘Functional Comment’ and ‘Subunit Comment’. Finally, an analysis tool was implemented that allows one to predict the occurrence of orthologous protein complex subunits in other mammals and other groups of organisms. The ‘Phylogenetic Conservation’ tool provides a probability whether or not a protein complex is likely to occur in the analysed model organisms. CORUM is freely accessible at http://mips.helmholtz-muenchen.de/genre/proj/corum/index.html.

NEW DEVELOPMENTS

Dataset and cross-references

In 2008 the CORUM dataset consisted of 1750 mammalian protein complexes, mainly characterized in human (60%), mouse (14%) and rat (14%) (14). While the relative abundance of the related organisms remained stable in the meantime, the number of protein complexes has grown to 2837 in September 2009. Thus, CORUM is the largest set of mammalian protein complexes publicly available.

However, compared to data from single-cell organisms only a minor fraction of the mammalian complexome has been discovered so far. Data from yeast have shown that at least 45% of the gene complement function as subunits in protein complexes (14). Considering that there is no comprehensive mammalian high-throughput dataset available to date, the fraction of genes that are involved in protein complex formation is comparably low. These estimates are based on the number of different complex subunit genes divided by a given number of 20 488 genes in human (14). Compared to the first CORUM release, this fraction increased moderately from 12% (2400 genes) to 16% (3198 genes). The slow increase of novel protein complex subunits presumably results from the reuse of subunits (Figure 1) in different protein complexes or protein complex variants (15). Data from the CORUM ‘Core Set’ (see below) show that proteins like ‘integrin beta-1’, ‘histone deacetylase 1’ and ‘histone deacetylase 2’ appear in 54, 51 and 38 different human protein complexes. Multiple reutilization of protein complex subunits is particularly found in large protein complex families like SNARE complexes and ubiquitin E3 ligases. The ubiquitin E3 ligase subunit ring-box 1 (Rbx1), for example, was identified in 35 complexes.

In addition to the complete dataset, CORUM now offers a reduced ‘Core Dataset’ for download and searches that avoids redundancies of data. Thoroughly investigated protein complexes like ‘SNARE complex (Vamp2, Snap25, Stx1a, Cplx1)’, ‘succinyl-CoA synthetase, ADP-forming’ and ‘cytochrome bc1-complex (EC 1.10.2.2), mitochondrial’ are characterized in more than one mammalian organism. Due to the close phylogenetic relationship between mammals it can be assumed that the majority of protein complexes are conserved in

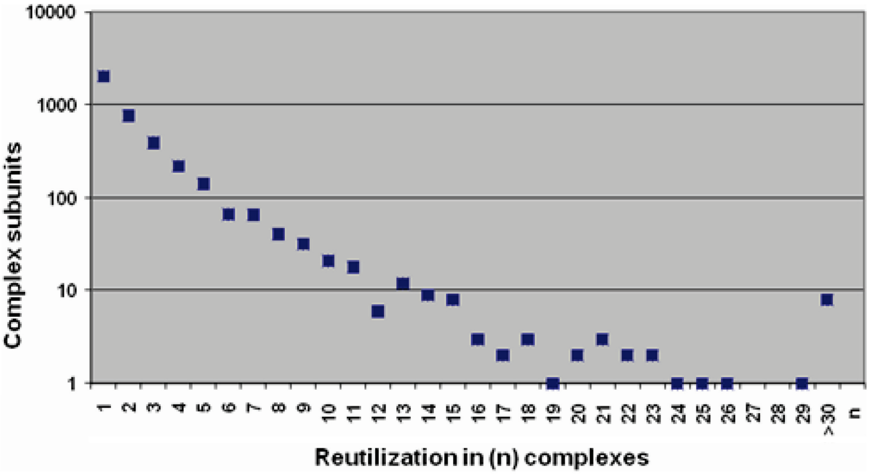


Figure 1. Reutilization of protein complex subunits. The plot shows that most proteins (2038) are found in only one protein complex and only eight proteins are subunits of at least 30 protein complexes. Data for the analysis are based on the CORUM ‘Core Set’.

mammals. However, as the aim of CORUM is to provide a comprehensive dataset, also evolutionary conserved protein complexes from different organisms (interologous protein complexes) are annotated in CORUM. To some extent this introduces redundancies, but on the other hand proves that the same protein complex in fact exists in different organisms.

Results from several laboratories that investigated the same protein complex but characterized the molecule with a different composition are another source of dataset expansion. These may stem from different experimental conditions that result in different complex compositions depending on the stringency of the experimental procedures or from different biomaterial that was used for the characterization. Bioinformatics applications like machine learning require non-redundant datasets. For these users we offer the 'Core Set' of 2084 distinct protein complexes. For the set only one representative of each interologous group of protein complexes or from protein complex variants was selected. We chose protein complexes which were thoroughly characterized and preferably from Homo sapiens.

Annotation of protein complex subunits in CORUM is performed with UniProt identifiers. Since some users prefer identifiers from Entrez Gene, we mapped the UniProt identifiers to the corresponding Entrez Gene identifiers. This was realized in a semi-automatic procedure using the CRONOS tool (23). CRONOS allows the mapping of identifiers, gene names and protein names from various resources like UniProt, RefSeq and Ensembl. In total, 4310 out of 4336 distinct subunits (98%) could be mapped to corresponding Entrez Gene identifiers. For 26 gene products like MRPS15 from *Bos taurus* or SPCS1 from *Canis familiaris* no respective identifier was available in Entrez.

CORUM is the only resource of protein complexes that includes functional annotation of the molecules. We use the FunCat annotation scheme for protein and protein complex function characterization (13). The FunCat has been used for genome annotation and was also frequently used for the analysis of protein networks and high-throughput experiments (13). The hierarchical structure of the FunCat allows browsing for protein complexes with particular cellular functions or localizations. In recent years, GO has become a widely used tool for the annotation of eukaryotic genomes (24). In contrast to the FunCat annotation scheme, the GO is constructed as a set of acyclic graphs, allowing more than one parent class per child (24). In order to enable bioinformatics analyses of protein complexes based on GO terms, the new CORUM release provides a mapping from FunCat to GO. The mapping was performed using the table that is available for download at http://www.geneontology.org/external2go/mips2go. As a result 840 FunCat categories could be mapped to 896 GO terms. Manual inspection of 100 randomly chosen protein complexes revealed that FunCat categories and GO terms are in agreement.

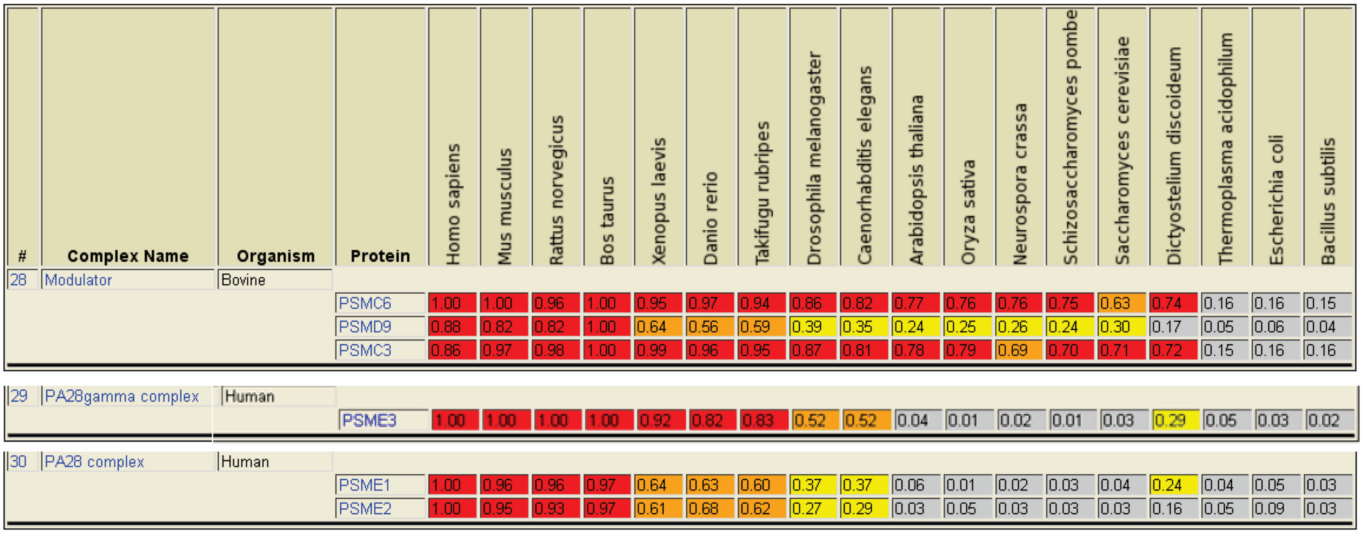
Some valuable information concerning protein complexes cannot be covered by systematic annotation schemes but is represented as free text comment in CORUM. This information includes protein complex composition (e.g. additional subunits of unknown identity), association of protein complexes with diseases or particular functional properties. In the first CORUM release this additional information was collected in a single comment field. In CORUM release 2.0 this content is now distributed among the three comment fields 'Functional Comment', 'Disease Comment' and 'Subunit Comment'. This separation allows to search in a particular type of information or using a wild card '_' for instance to retrieve all 223 protein complexes with information about disease association.

Phylogenetic analysis of protein complexes

Protein complex subunits from protein complexes like ribosomes and chaperonins are highly conserved in evolution. Beside ribosomal RNAs, subunits from complexes such as RNA polymerases (25) and F1-ATPases (26) were used for phylogenetic analyses in the early days of sequence-based phylogenetic analyses. Based on data from 191 sequenced genomes, 2 years ago a novel endeavor was started to investigate highly conserved proteins for phylogenetic analysis (27). Analysis revealed 31 highly conserved proteins that allow a new reconstruction of the tree of life and 28 of these proteins are known to be protein complex subunits (23 ribosomal proteins). To enable scientists to obtain some insight into the phylogenetic conservation of subunits, the 'Phylogenetic Conservation' tool has been developed for comparative proteome analysis. The 'Phylogenetic Conservation' tool is based on sequence similarity data that are obtained from the SIMAP database (28). The Similarity Matrix of Proteins (SIMAP) database provides a comprehensive and up-to-date dataset of the pre-calculated sequence similarity matrix and sequence-based features like InterPro domains for all proteins contained in the major public sequence databases.

The 'Phylogenetic Conservation' tool in CORUM presents the similarity of the protein complex subunits to proteins from other organisms as tables (Figure 2). As default comparison to 18 organisms are shown, four mammals (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Bos taurus*), three other vertebrates (*Xenopus laevis*, *Danio rerio* and *Takifugu rubripes*), two invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*), two plants (*Arabidopsis thaliana* and *Oryza sativa*), three fungi (*Neurospora crassa*, *Schizosaccharomyces pombae* and *S. cerevisiae*), one slime mold (*Dictyostelium discoideum*) and three prokaryotes (*Thermoplasma acidophilum*, *Escherichia coli* and *Bacillus subtilis*). In addition to the numerical values, the degree of protein sequence similarity is colour coded.

The conservation of protein complexes appears to be conserved among all phylogenetic related organisms and separates organisms of distant phylogenetic relation, depending on the respective complex. This can be illustrated with the proteasome and three proteasome activatory complexes. Two subunits of the 'Modulator (PA700-dependent proteasome activator)' are highly conserved (red colour) within all eukaryotes, whereas the 'PA28 gamma complex' is only highly conserved



#	Complex Name	Organism	Protein	Homo sapiens	Mus musculus	Rattus norvegicus	Bos taurus	Xenopus laevis	Danio rerio	Takifugu rubripes	Drosophila melanogaster	Caenorhabditis elegans	Arabidopsis thaliana	Oryza sativa	Neurospora crassa	Schizosaccharomyces pombe	Saccharomyces cerevisiae	Dictyostelium discoideum	Thermoplasma acidophilum	Escherichia coli	Bacillus subtilis
28	Modulator	Bovine																			
			PSMC6	1.00	1.00	0.96	1.00	0.95	0.97	0.94	0.86	0.82	0.77	0.76	0.76	0.75	0.63	0.74	0.16	0.16	0.15
			PSMD9	0.88	0.82	0.82	1.00	0.64	0.56	0.59	0.39	0.35	0.24	0.25	0.26	0.24	0.30	0.17	0.05	0.06	0.04
			PSMC3	0.86	0.97	0.98	1.00	0.99	0.96	0.95	0.87	0.81	0.78	0.79	0.69	0.70	0.71	0.72	0.15	0.16	0.16
29	PA28gamma complex	Human																			
			PSME3	1.00	1.00	1.00	1.00	0.92	0.82	0.83	0.52	0.52	0.04	0.01	0.02	0.01	0.03	0.29	0.05	0.03	0.02
30	PA28 complex	Human																			
			PSME1	1.00	0.96	0.96	0.97	0.64	0.63	0.60	0.37	0.37	0.06	0.01	0.02	0.03	0.04	0.24	0.04	0.05	0.03
			PSME2	1.00	0.95	0.93	0.97	0.61	0.68	0.62	0.27	0.29	0.03	0.05	0.03	0.03	0.03	0.16	0.05	0.09	0.03

Figure 2. Phylogenetic conservation of proteasome regulatory protein complexes. Results of the phylogenetic conservation tool from CORUM for the three proteasome regulators 'Modulator', 'PA28 gamma complex' and 'PA28 complex' are shown. Similarity of protein complex subunits to proteins from other organisms are represented color coded as well as opt. score/self score ratios. The data are obtained from the SIMAP database.

within vertebrates (Figure 2). Finally, high conservation of the '11 S REG complex' is restricted to the four mammalian proteomes. The 20 S proteasome complex is a high-molecular-weight protease that is essential for protein degradation in mammals. Results of the 'Phylogenetic Conservation' tool reveal weak similarity for proteins in the archaeon *T. acidophilum* (Supplementary Figure S1). In fact, an archetype of proteasomes, consisting of only two different subunits is frequently found in archaea (29). On the other hand, sophisticated proteasome architectures like the 26 S proteasome or the availability of several proteasome activatory complexes are not found in *Thermoplasma* or other prokaryotes. In agreement with this observation, the three above mentioned proteasome activators show no similarity to proteins from *Thermoplasma* (Figure 2). Results of the 'Phylogenetic Conservation' tool can be retrieved for single protein complexes or for multiple complexes that were found by one of the search options in CORUM.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Thomas Rattei for providing data from SIMAP.

FUNDING

ERA-NET PathoGenoMics 'Pathomics' grant (BMBF) (to B.W.). Funding to open access charge: Helmholtz Center Munich (Helmholtz Zentrum München).

Conflict of interest statement. None declared.

REFERENCES

1. Alberts,B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.
2. Guldener,U., Munsterkotter,M., Kastenmuller,G., Strack,N., Van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J., Perez-Ortin,J.E. *et al.* (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.
3. Han,J.D., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L.V., Dupuy,D., Walhout,A.J., Cusick,M.E., Roth,F.P. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.
4. Kim,P.M., Lu,L.J., Xia,Y. and Gerstein,M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938–1941.
5. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
6. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dumpelfeld,B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
7. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N., Tikuisis,A.P. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.
8. Jensen,L.J. and Bork,P. (2008) Biochemistry. Not comparable, but complementary. *Science*, **322**, 56–57.
9. Braun,P., Tasan,M., Dreze,M., Barrios-Rodiles,M., Lemmens,I., Yu,H., Sahalie,J.M., Murray,R.R., Roncari,L., de Smet,A.S. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods*, **6**, 91–97.
10. Lage,K., Karlberg,E.O., Storling,Z.M., Olason,P.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tumer,Z., Pociot,F., Tommerup,N. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.
11. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
12. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M.

et al. (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.

13. Ruepp,A., Zollner,A., Maier,D., Albermann,K., Hani,J., Mokrejs,M., Tetko,I., Guldener,U., Mannhaupt,G., Munsterkotter,M. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
14. Ruepp,A., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Stransky,M., Waegele,B., Schmidt,T., Doudieu,O.N., Stumpflen,V. *et al.* (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, D646–D650.
15. Wong,P., Althammer,S., Hildebrand,A., Kirschner,A., Pagel,P., Geissler,B., Smialowski,P., Blochl,F., Oesterheld,M., Schmidt,T. *et al.* (2008) An evolutionary and structural characterization of mammalian protein complex organization. *BMC Genomics*, **9**, 629.
16. Bossi,A. and Lehner,B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.*, **5**, 260.
17. Friedel,C.C., Dolken,L., Ruzsics,Z., Koszinowski,H. and Zimmer,R. (2009) Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Res.*, **37**, e115.
18. Meyer,E., Aglyamova,G.V., Wang,S., Buchanan-Carter,J., Abrego,D., Colbourne,J.K., Willis,B.L. and Matz,M.V. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219.
19. Zampieri,M., Soranzo,N. and Altafini,C. (2008) Discerning static and causal interactions in genome-wide reverse engineering problems. *Bioinformatics*, **24**, 1510–1515.
20. Schelhorn,S.E., Lengauer,T. and Albrecht,M. (2008) An integrative approach for predicting interactions of protein regions. *Bioinformatics*, **24**, i35–i41.
21. Dietmann,S., Georgii,E., Antonov,A., Tsuda,K. and Mewes,H.W. (2009) The DICS repository: module-assisted analysis of disease-related gene lists. *Bioinformatics*, **25**, 830–831.
22. Sun,C.H., Kim,M.S., Han,Y. and Yi,G.S. (2009) COFECO: composite function annotation enriched by protein complex data. *Nucleic Acids Res.*, **37**, W350–W355.
23. Waegele,B., Dunger-Kaltenbach,I., Fobo,G., Montrone,C., Mewes,H.W. and Ruepp,A. (2009) CRONOS: the cross-reference navigation server. *Bioinformatics*, **25**, 141–143.
24. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.
25. Puhler,G., Leffers,H., Gropp,F., Palm,P., Klenk,H.P., Lottspeich,F., Garrett,R.A. and Zillig,W. (1989) Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc. Natl Acad. Sci. USA*, **86**, 4569–4573.
26. Iwabe,N., Kuma,K., Hasegawa,M., Osawa,S. and Miyata,T. (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad Sci. USA*, **86**, 9355–9359.
27. Ciccarelli,F.D., Doerks,T., von,M.C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
28. Rattei,T., Arnold,R., Tischler,P., Lindner,D., Stumpflen,V. and Mewes,H.W. (2006) SIMAP: the similarity matrix of proteins. *Nucleic Acids Res.*, **34**, D252–D256.
29. Lupas,A., Zuhl,F., Tamura,T., Wolf,S., Nagy,I., De,M.R. and Baumeister,W. (1997) Eubacterial proteasomes. *Mol. Biol. Rep.*, **24**, 125–131.