Protein Cell 2012, 3(8): 602–608
DOI 10.1007/s13238-012-2914-8

Protein & Cell

COMMUNICATION

Functional annotation from the genome sequence of the giant panda

Tong Huo[1,3], Yinjie Zhang[1,3], Jianping Lin[2,3] ✉

[1] College of Life Sciences, Nankai University, Tianjin 300071, China
[2] College of Pharmacy, Nankai University, Tianjin 300071, China
[3] State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, China
✉ Correspondence: jianpinglin@nankai.edu.cn
Received March 15, 2012 Accepted March 22, 2012

ABSTRACT

The giant panda is one of the most critically endangered species due to the fragmentation and loss of its habitat. Studying the functions of proteins in this animal, especially specific trait-related proteins, is therefore necessary to protect the species. In this work, the functions of these proteins were investigated using the genome sequence of the giant panda. Data on 21,001 proteins and their functions were stored in the Giant Panda Protein Database, in which the proteins were divided into two groups: 20,179 proteins whose functions can be predicted by GeneScan formed the *known-function group*, whereas 822 proteins whose functions cannot be predicted by GeneScan comprised the *unknown-function group*. For the known-function group, we further classified the proteins by molecular function, biological process, cellular component, and tissue specificity. For the unknown-function group, we developed a strategy in which the proteins were filtered by cross-Blast to identify panda-specific proteins under the assumption that proteins related to the panda-specific traits in the unknown-function group exist. After this filtering procedure, we identified 32 proteins (2 of which are membrane proteins) specific to the giant panda genome as compared against the dog and horse genomes. Based on their amino acid sequences, these 32 proteins were further analyzed by functional classification using SVM-Prot, motif prediction using MyHits, and interacting protein prediction using the Database of Interacting Proteins. Nineteen proteins were predicted to be zinc-binding proteins, thus affecting the activities of nucleic acids. The 32 panda-specific proteins will be further investigated by structural and functional analysis.

KEYWORDS Giant panda, GPPD, cross-Blast

INTRODUCTION

The giant panda (*Ailuropoda melanoleuca*) is one of the most critically endangered species due to the fragmentation and loss of its habitat. Studies have shown that only approximately 2500–3000 pandas are distributed in several small mountains in Western China (Zhan et al., 2006). The giant panda has several specific characteristics, including a bamboo-only diet, a very low fecundity rate, a distinctive black and white fur pattern, and a controversial phylogenetic position in evolution (Krause et al., 2008). Studying these traits at the genomic and proteomic levels is imperative to ensure the survival of the species.

Previous studies have used genetic methods to extract phylogenetic and heterogeneous information on the giant panda (Pagés et al., 2008; Hama et al., 2009; Hao et al., 2009). Research has shown that the karyotype of the giant panda is similar to that of the bear (Nash et al., 1998), suggesting that the bear and the giant panda have similar genome sizes. In 2010, Li et al. (2010) successfully generated and assembled a draft sequence of the giant panda genome, thus providing an essential tool for detailed understanding of the biological traits of the giant panda. Their study found that, although it is taxonomically classified as a carnivore, the bamboo diet of the giant panda might be more dependent on its gut microbiome than on its own genetic composition.

In this study, the functions of proteins in the giant panda were investigated using its genome sequence. Data on 21,001 proteins were stored in the Giant Panda Protein Database (GPPD) (http://60.28.101.183/home.html). The functions of these proteins were analyzed using GeneScan (Burge et al., 1997) and subjected to functional classification

using SVM-Prot (Cai et al., 2003), motif prediction using My-Hits (Pagni et al., 2007), and interacting protein prediction using the Database of Interacting Proteins (DIP) (Xenarios et al., 2002).

RESULTS

Known-function group vs. unknown-function group

The functional annotation files of proteins in the giant panda were classified by function field. The functional annotations were presented by Gene Ontology using GeneScan (Burge et al., 1997). In the absence of any annotation for a protein, we regarded the protein to be of unknown function; otherwise, the protein was included in the known-function group. After functions were assigned, we obtained 822 unknown-function proteins and 20,179 known-function proteins.

GPPD

The GPPD was developed to store information on the 21,001 proteins in the giant panda and their predicted functions with web access. A snapshot of the interface of the GPPD is shown in Fig. 1.

According to further information predicted from Uniprot (Magrane et al., 2011), the proteins from the known-function group were classified by biological process, cellular component, molecular function, and tissue specificity. GPPD users can type keywords to search the database. For example, a search for “biological process” will lead to 5500 categories, including cell death, protein insertion into membrane, and cell junction assembly, among others. A search for “cell death” results in 81 proteins shown to be involved in the cell death process.

Filtering the unknown-function group

Filtering by cross-Blast

The giant panda is primarily an herbivore, despite its taxonomic classification as a carnivore. Previous research on the genome of the giant panda has shown that it shares the highest genome sequence homology with the dog among carnivorous animals and with the horse among herbivorous animals. We therefore chose the genome sequences of the dog and horse for comparison with that of the giant panda to identify proteins that might be related to the specific traits of this animal.

Eight hundred twenty-two proteins with unknown functions from the peptide text file of the giant panda genome sequence were analyzed (She et al., 2009) against the dog and horse genome sequences using Blast. Fig. 2 shows the cross-Blast procedure, and Table 1 shows the cross-Blast results.

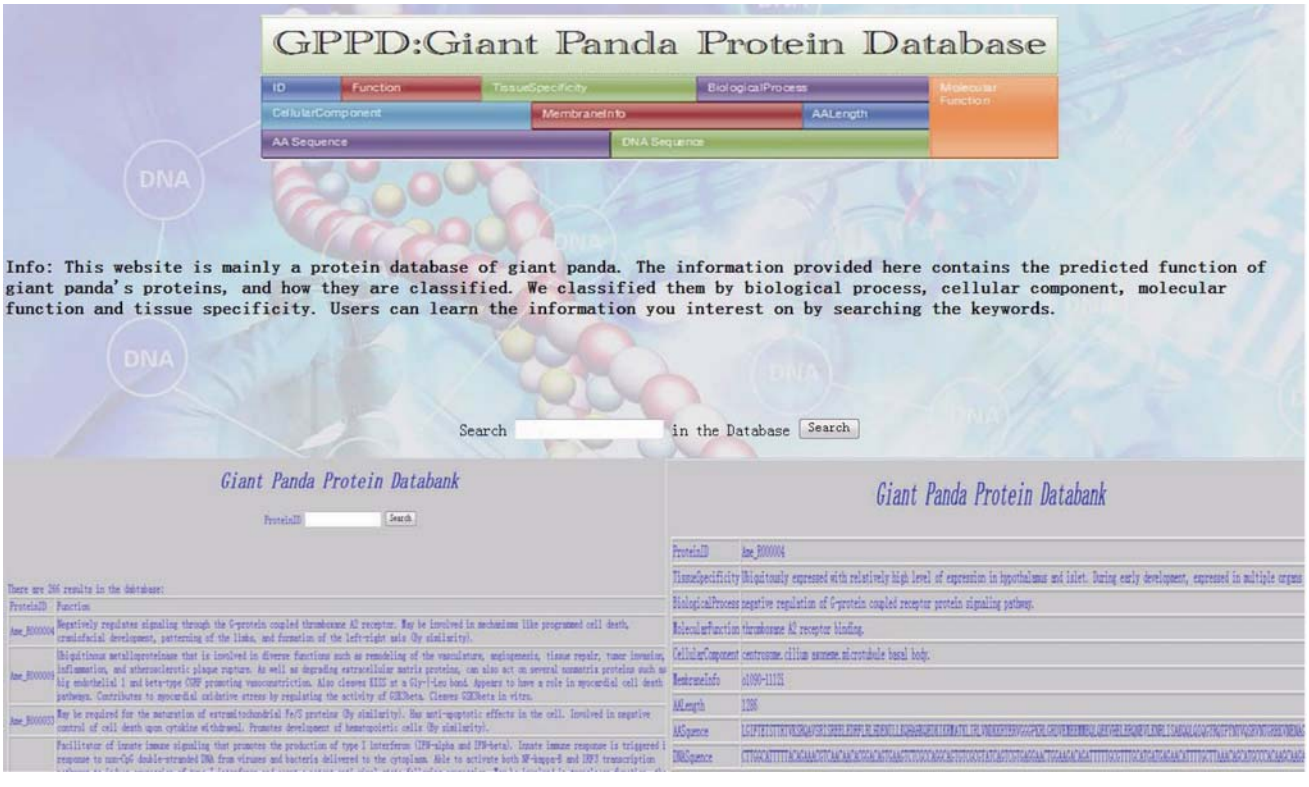


Figure 1. Snapshot of the interface of the GPPD

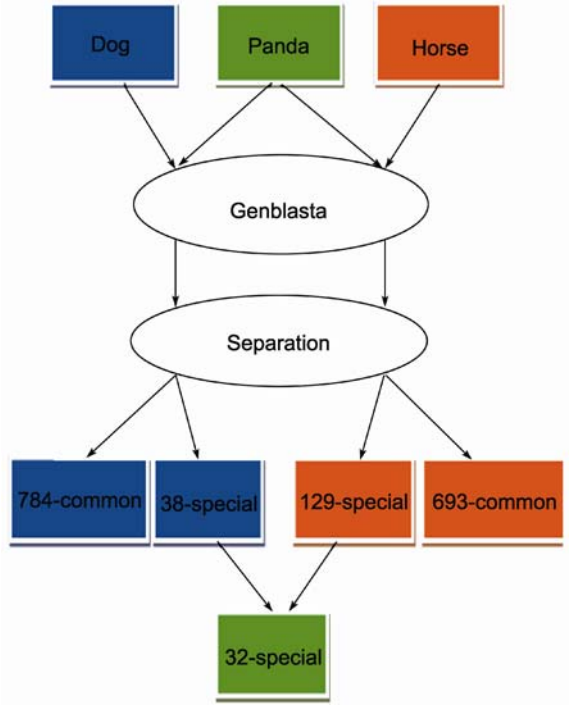


Figure 2. Filtering by the cross-Blast procedure

Table 1 Numbers of proteins from a cross-Blast search of the panda genome against the dog/horse genome

	Common with panda	Specific for panda
Dog	784	38
Horse	693	129
Dog and horse		32

Of the 32 panda-specific proteins, 2 were predicted to be membrane proteins using the TMHMM program. Table S1 shows the TMHMM results. The sequences of these proteins are provided in Supplemental Material.

SVM-Prot

Panels A and B of Fig. 3 show the results of SVM-Prot analysis. Of the 32 proteins, 19 were classified as zinc- binding proteins, whereas the remaining proteins were categorized as lipid-binding/membrane proteins, DNA-binding proteins, nuclear receptors, or transferases, among others.

Motif Scan of MyHits

Table 2 shows the results of Motif Scan analysis. Five proteins with special motifs were found, namely, Ame_R000609 (with a bipartite nuclear localization signal profile), Ame_R013224 (with a glycine-rich region profile and a proline-rich region profile), Ame_R019548 (with a lysine-rich region profile and a bipartite nuclear localization signal profile), Ame_R017570 (with a protamine P1 signature), and Ame_R009902 (with a trefoil domain and glycosyl hydrolases).

Table 2 Motif prediction results using Motif Scan of MyHits

Name	Score	Description
Ame_R000248	?	None
Ame_R000307	?	None
Ame_R000609	!	Bipartite nuclear localization signal profile
Ame_R000622	?	None
Ame_R002839	?	None
Ame_R002945	?	None
Ame_R004369	?	None
Ame_R005446	?	None
Ame_R006158	?	None
Ame_R007364	?	None
Ame_R008064	?	None
Ame_R008184	?	None
Ame_R008820	?	None
Ame_R008980	?	None
Ame_R011061	?	None
Ame_R011984	?	None
Ame_R012522	?	None
Ame_R013038	?	None
Ame_R013090	?	None
Ame_R013224	!	1. Glycine-rich region profile 2. Proline-rich region profile
Ame_R013532	?	None
Ame_R013753	?	None
Ame_R014240	?	None
Ame_R015262	?	None
Ame_R017002	?	None
Ame_R017525	?	None
Ame_R017671	?	None
Ame_R018563	Empty	Empty
Ame_R019548	!	Lysine-rich region profile Bipartite nuclear localization signal profile
Ame_R013532	?	None
Ame_R017570	!	Protamine P1 signature
Ame_R020040	?	None

The subject database includes perox, hamap, pat, freq_pat, prf, pre, pfam_fs, and pfam_ls. A question mark in the score field indicates a questionable or weak match: determining the true- or false-negative status of which requires additional biological evidence. On the other hand, an exclamation point in the score field indicates a strong match: it is very unlikely that this match is false positive.

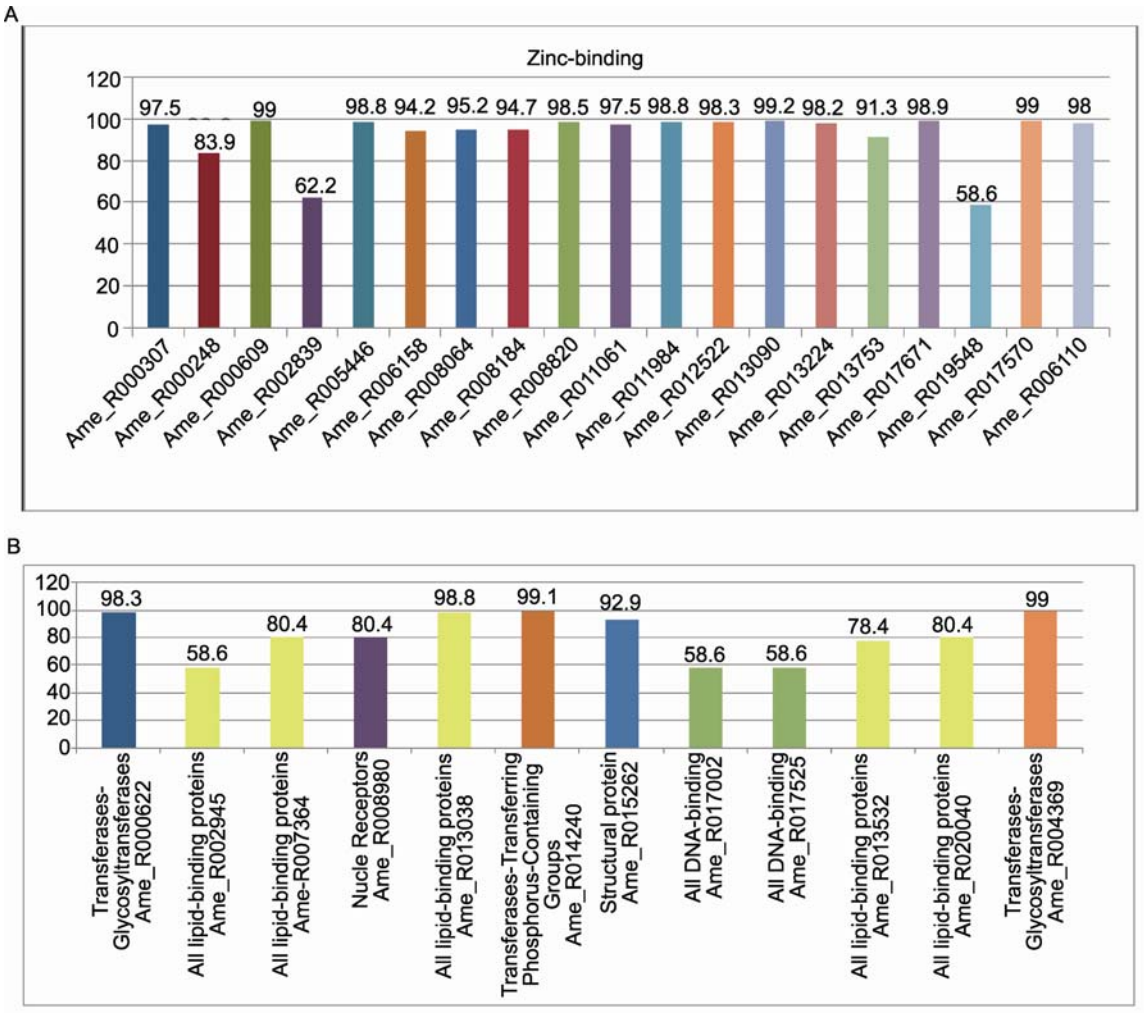


Figure 3. Functional classification of the 32 panda-specific proteins using SVM-Prot. The *y*-coordinate data represent the expected classification accuracy (%) (probability of correct classification). (A) Nineteen proteins were predicted to be zinc-binding proteins. (B) Twelve proteins were predicted to be lipid-binding proteins and DNA-binding proteins, among others. (Ame_R018563 was not included because its amino acid length was too short to be predicted.)

Interacting protein prediction

Table 3 shows the DIP results. Seventeen of the target proteins were predicted to interact with nucleic acids and therefore may be involved in replication/transcription/translation activities. Table 4 shows the results obtained using a combination of the three prediction methods. Two of the proteins were predicted by all three methods to be associated with nucleic acids (see sequences of Ame_R000609 and Ame_R019548 in Supplemental Material), whereas 9 proteins were predicted by two methods, 16 by one method, and 5 by none to be so.

DISCUSSION

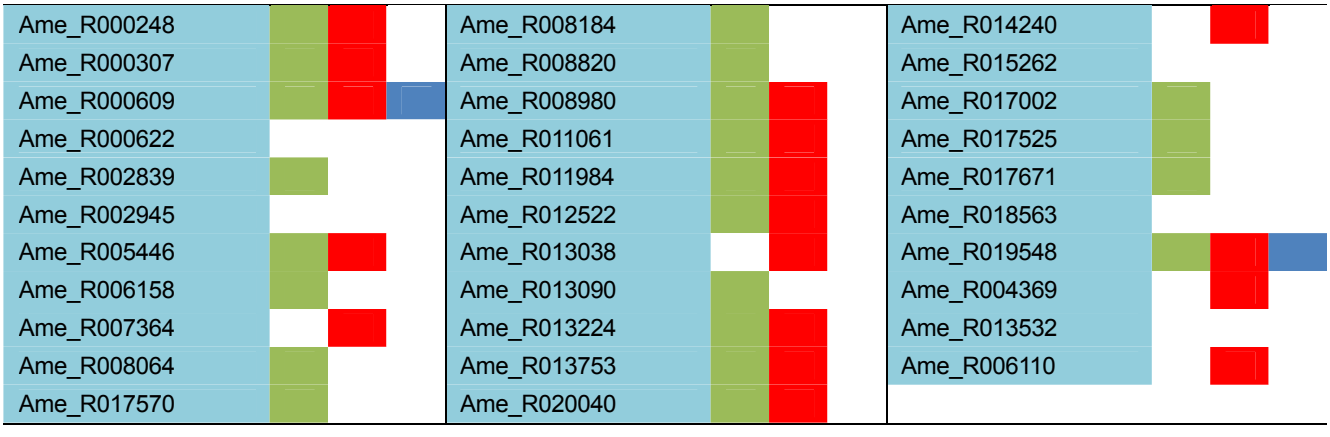
The main structure of the GPPD is illustrated in Fig. 4. The proteins were classified by biological process, cellular component, molecular function, and tissue specificity. One classification can also be specified for different subclasses or pathways according to the specific function of each protein. Biological process contains 5476 pathways, cellular component contains 1421, molecular function contains 3317, and tissue specificity contains 8346.

Using the GPPD, we investigated the special traits of the giant panda, including its digestive and genital systems. Eighty-seven proteins were associated with the digestive function, and 37 were associated with the genital system. We hypothesized that although these functions are already known, they can also play a significant role in determining the giant panda's specific traits. These protein sequences are provided in the supplementary file and can also be obtained from the GPPD. The giant panda genome sequence peptide text file contains 822 proteins with unknown functions. We further hypothesized that several of these proteins are related to the specific traits of the giant panda.

Table 3 Results of interacting protein prediction using the DIP

Name	Interacting protein: name/description
Ame_R000248	DNA-directed RNA polymerase I subunit RPA2
Ame_R000307	RAD2 protein
Ame_R000609	66 kDa U4/U6.U5 small nuclear ribonucleoprotein component
Ame_R000622	CG8378-PA open reading frame
Ame_R002839	Kinesin-like protein 2
Ame_R002945	CG5731-PA open reading frame
Ame_R005446	Guanine nucleotide exchange factor LTE1
Ame_R006158	Keratin, type I cytoskeletal 18
Ame_R007364	Probable pyridine nucleotide-disulfide oxidoreductase ykgC
Ame_R008064	Kekkon-3 (CG4192-PA)
Ame_R008184	Probable membrane protein YPL012w
Ame_R008820	Dynamin-related protein DNM1
Ame_R008980	BUD8 protein
Ame_R011061	DNA-polymerase-delta (CG5949-PA)
Ame_R011984	Probable helicase RAD26
Ame_R012522	Nucleoporin NUP159
Ame_R013038	TBP-associated factor 11
Ame_R013090	CG9791-PA open reading frame
Ame_R013224	Zinc finger, C2H2 type
Ame_R013753	Translation elongation factor eEF-3 homolog YPL226w
Ame_R014240	Combgap (CG8367-PB)
Ame_R015262	Hypothetical protein C17G10.4 (confirmed)
Ame_R017002	Fimbrin
Ame_R017525	CD30 ligand
Ame_R017671	AST2 protein
Ame_R018563	None
Ame_R019548	Finger protein SIG1
Ame_R004369	Polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase)
Ame_R013532	Probable membrane protein YPL032c
Ame_R006110	CG3571-PA open reading frame
Ame_R017570	Probable membrane protein YNL127w
Ame_R020040	Origin recognition complex chain ORC5

Table 4 Prediction of nucleic acid-associated proteins



Cells in green, red, and blue represent proteins predicted using SVM-Prot, the DIP, and MyHits, respectively, to be associated with nucleic acids.

The dog and horse share a high genome sequence homology with the giant panda (Li et al., 2010). The dog is classified as a carnivore, whereas the horse is classified as an herbivore; the giant panda's diet is primarily herbivorous. We therefore chose the dog and horse genome sequences for comparison with the giant panda genome sequence to identify proteins that might be related to the specific traits of the giant panda. Of the 822 proteins with known functions in the giant panda, 38 are unique to the giant panda; 784 proteins are conserved between the dog and panda genome sequences, whereas only 693 are conserved between the horse and panda genome sequences. These findings suggest that the giant panda has a greater similarity to the dog than to the horse. With cross-Blast searches between the dog, horse, and panda genome sequences, 32 proteins were found to be unique to the giant panda, of which 2 were predicted to be membrane proteins. Nineteen of the 32 proteins were further found to be zinc-binding proteins that carry out important biological functions, including serving as transcription factors and participating in substrate activation as well as enzymatic catalysis. The rest of the 32 proteins were predicted to be lipid-binding proteins, DNA-binding proteins, nuclear receptors, or transferases. Motif Scan analysis revealed the following: Ame_R000609 consisted of a bipartite nuclear localization signal sequence that can direct a newly synthesized protein into the nucleus (Wikipedia.org); Ame_R013224 consisted of a glycine-rich region sequence (Bocca et al., 2005) and a proline-rich region sequence (Williamson et al., 1994) that are involved in protein–protein interactions as well as in binding and signal transduction and transcription activation (Williamson et al., 1994; Bocca et al., 2005); Ame_R019548 consisted of a lysine-rich region profile and a bipartite nuclear localization signal profile that can direct a newly synthesized protein into the nucleus and are involved in molecular interactions (Wikipedia.org); Ame_R017570 consisted of a protamine P1 signature that is replaced by histones late in the haploid phase of spermatogenesis and believed to be essential for sperm head condensation and DNA stabilization (Wikipedia.org); and Ame_R009902 consisted of (1) a trefoil (P-type) domain in which a cysteine-rich domain of approximately 45 amino acid residues has been found in some extracellular eukaryotic proteins and (2) a glycosyl hydrolase domain that can catalyze the hydrolysis of the glycosidic linkage to release smaller sugars (Wikipedia.org). The interacting proteins of the 32 proteins were further investigated using a combination of SVM-Prot, MyHits, and the DIP. Of these proteins, 2, 9, 16, and 5 were predicted by all three methods, two methods, one method, and none, respectively, to associate with nucleic acids. All these functions indicate that the 32 proteins might be related to specific traits of the giant panda. The 30 non-membrane proteins will be further investigated by protein crystallography to elucidate their structure and function.

In summary, the GPPD was developed to store information on the 21,001 proteins in the giant panda and their predicted functions. We have screened the genome sequence of the giant panda and identified 32 proteins that may be related to its specific characteristics. The structural and functional data on these proteins provide a starting point for a more detailed understanding of the biological traits of the species.

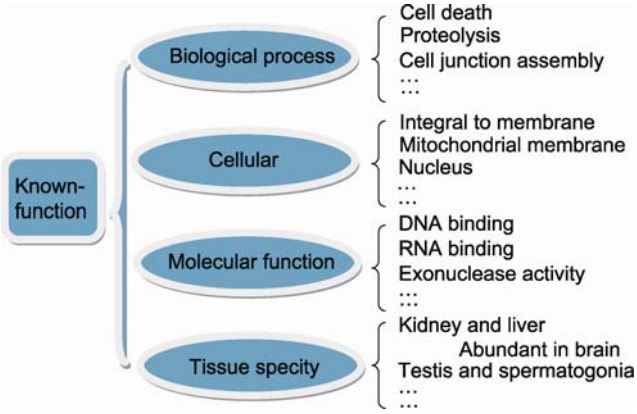


Figure 4. Main structure of the GPPD

MATERIALS AND METHODS

Materials

In 2010, Li et al. successfully generated and assembled a draft sequence of the giant panda genome, thereby providing an essential tool for detailed understanding of the biological traits of the giant panda. The draft giant panda genome sequence was provided by BGI Shenzhen (Li et al., 2010). GenBlastA (She et al., 2009), GeneScan (Burge et al., 1997), Augustus (Stanke et al., 2003), and GeneWise (Birney et al., 2004) were used to generate CDS, peptide, and gff text files as described by Li et al. (2010). The specific proteins were filtered from the total of 21,001 proteins. The dog and horse cDNA files were obtained from Ensembl (http://www.ensembl. org/index.html).

Methods

Cross-Blast

Fig. 2 shows a schematic of the cross-Blast procedure. We first performed a Blast search of the panda sequence against that of the dog or horse sequence and identified the commonly conserved proteins. We then performed a Blast search of the panda sequence against the dog and horse sequences and identified the panda-specific proteins using GenBlastA (She et al., 2009). The panda protein database was used to provide the query sequences, and the dog/horse cDNA database was used for the target sequences. The e-value, minimum percentage of query gene coverage, and minimum score of the HSP group were set to 1e−5, 0.01, and −100, respectively.

SVM-Prot

SVM-Prot (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi) (Cai et al.,

2003) is a web-based support vector machine software that classifies a protein into a functional family from its primary sequence. Protein sequences of interest can be placed in this program, which predicts their properties using the SVM algorithm and gives some values to evaluate the predictions. We further classified the 32 panda-specific proteins identified by cross-Blast using this tool (Cai et al., 2003).

Motif Scan of MyHits

The Motif Scan of MyHits (http://myhits.isb-sib.ch/cgi-bin/ motif_scan) (Pagni et al., 2007) identifies sequence motifs. This tool can use eight motif databases (perox, hamap, pat, freq_pat, prf, pre, pfam_fs, and pfam_ls; Pagni et al., 2007) to perform predictions. We ran Motif Scan on the 32 panda-specific proteins.

DIP

The DIP (http://dip.doe-mbi.ucla.edu/dip/Search.cgi) (Xenarios et al., 2002) was used to identify the proteins associating with the unknown-function proteins. The functions of proteins were analyzed from their associating components, indicating that some specific associating components can verify the predicted functions of the proteins.

GPPD

The GPPD system server is hosted by Red Hat Enterprise Linux. The database for storing information on the proteins was created using MySQL, whereas the website was developed using Perl and PHP.

ACKNOWLEDGMENTS

We thank Dr. Huanming Yang, Dr. Jian Wang, and Dr. Jun Wang of BGI Shenzhen for sharing the draft sequence of the giant panda genome. We also thank Professor Zihe Rao for his insightful discussion of the manuscript, Dr. Feng Xu for his help with building the GPPD server, as well as Dr. Taijiao Jiang of Institute of Biophysics, Chinese Academy of Sciences, and Dr. Mark Bartlam for their helpful suggestions on the manuscript.

REFERENCES

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res 14, 988–995.

Bocca, S. N., Magioli, C., Mangeon, A., Junqueira, R. M., Cardeal, V., Margis, R., Sachetto-Martins, G. (2005). Survey of glycine-rich proteins (GRPs) in the Eucalyptus expressed sequence tag database (ForEST). Genet Mole Biol 28, 608–624.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol 268, 78–94.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., and Chen, Y.Z. (2003). SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31, 3692–3697.

Hao, Y.Z., Hou, W.R., Hou, Y.L., Du, Y.J., Zhang, T., and Peng, Z.S. (2009). cDNA, genomic sequence cloning and overexpression of ribosomal protein S25 gene (RPS25) from the Giant Panda. Mol Biol Rep 36, 2139–2145.

Hama, N., Kanemitsu, H., Tanikawa, M., Shibaya, M., Sakamoto, K., Oyama, Y., Acosta, T.J., Ishikawa, O., Pengyan, W., and Okuda, K. (2009). Development of an enzyme immunoassay for urinary pregnanediol-3-glucuronide in a female giant panda (Ailuropoda melanoleuca). J Vet Med Sci 71, 879–884.

Krause, J., Unger, T., Nocon, A., Malaspinas, A.S., Kolokotronis, S.O., Stiller, M., Soibelzon, L., Spriggs, H., Dear, P.H., Briggs, A.W., et al. (2008). Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol Biol 8, 220.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305, 567–580.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. (2010). The sequence and de novo assembly of the giant panda genome. Nature 463, 311–317.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011, bar009.

Nash, W.G., Wienberg, J., Ferguson-Smith, M.A., Menninger, J.C., and O'Brien, S.J. (1998). Comparative genomics: tracking chromosome evolution in the family ursidae using reciprocal chromosome painting. Cytogenet Cell Genet 83, 182–192.

Pages, M., Calvignac, S., Klein, C., Paris, M., Hughes, S., and Hanni, C. (2008). Combined analysis of fourteen nuclear genes refines the Ursidae phylogeny. Mol Phylogenet Evol 47, 73–83.

Pagni, M., Ioannidis, V., Cerutti, L., Zahn-Zabal, M., Jongeneel, C.V., Hau, J., Martin, O., Kuznetsov, D., and Falquet, L. (2007). MyHits: improvements to an interactive resource for analyzing protein sequences. Nucleic Acids Res 35, W433–437.

She, R., Chu, J.S., Wang, K., Pei, J., and Chen, N. (2009). GenBlastA: enabling BLAST to identify homologous gene sequences. Genome Res 19, 143–149.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2, ii215–225.

Williamson, M.P. (1994). The structure and function of proline-rich regions in proteins. Biochem J 297 (Pt 2), 249–260.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30, 303–305.

Zhan, X., Li, M., Zhang, Z., Goossens, B., Chen, Y., Wang, H., Bruford, M.W., and Wei, F. (2006). Molecular censusing doubles giant panda population estimate in a key nature reserve. Curr Biol 16, R451–452.

