D150–D152 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj025

NPInter: the noncoding RNAs and protein related biomacromolecules interaction database

Tao Wu[1,3], Jie Wang[1,3], Changning Liu[2,3], Yong Zhang[1,3], Baochen Shi[1,3], Xiaopeng Zhu[1,3], Zhihua Zhang[1,3], Geir Skogerbø[1], Lan Chen[2,3], Hongchao Lu[2,3], Yi Zhao[2] and Runsheng Chen[1,2,*]

[1]Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China, [2]Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Beijing 100080, China and [3]Graduate School of the Chinese Academy of Sciences, Beijing, China

Received August 9, 2005; Revised and Accepted September 21, 2005

ABSTRACT

The noncoding RNAs and protein related biomacromolecules interaction database (NPInter; http://bioinfo.ibp.ac.cn/NPInter or http://www.bioinfo.org.cn/NPInter) is a database that documents experimentally determined functional interactions between noncoding RNAs (ncRNAs) and protein related biomacromolecules (PRMs) (proteins, mRNAs or genomic DNAs). NPInter intends to provide the scientific community with a comprehensive and integrated tool for efficient browsing and extraction of information on interactions between ncRNAs and PRMs. Beyond cataloguing details of these interactions, the NPInter will be useful for understanding ncRNA function, as it adds a very important functional element, ncRNAs, to the biomolecule interaction network and sets up a bridge between the coding and the noncoding kingdoms.

INTRODUCTION

NPInter (the noncoding RNAs and protein related biomacromolecules interaction database) aims at integrating the diverse body of experimental knowledge about functional interactions between noncoding RNAs (ncRNAs) (except tRNAs and rRNAs) and protein related biomacromolecules (PRMs) (proteins, mRNAs or genomic DNAs) into a single, easily accessible database. By functional interactions we mean both physical interactions between an ncRNA and a protein, and other forms of interaction where the combination of an ncRNA and an mRNA or a genomic DNA sequence elicits a cellular reaction. Although biological knowledge on this sort of interactions is contained in the scientific literature, retrieving it will be much easier with the NPInter database.

The data in NPInter are novel, in the sense that no earlier database has especially catalogued this type of data. ncRNAs are the functional molecules in the noncoding kingdom and PRMs, proteins and protein coding related molecules (mRNAs and genomic DNAs), are the functional molecules in the coding kingdom. Therefore, NPInter sets up a bridge between the coding and the noncoding realms.

The database now contains 700 published functional interactions from six model organisms, such as *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*. The amount of data is not large, but the NPInter covers almost all experimentally verified functional interactions between ncRNA and PRM published before the end of the year 2004.

THE NEED FOR NPInter

It is maintained that the diversity of genes cannot approximate the diversity of functions within an organism (1). Recent discoveries in the 'Modern RNA World' have made it clear that large-scale mRNA expression profiling data provide only a partial picture of gene expression. Most post-transcriptional events are mediated by the association of RNAs with specific proteins or macromolecular protein complexes (2).

Systems biology is becoming increasingly important for the discovery of new properties of biological systems. One major aim of systems biology is to understand how the various components of biological systems are combined to produce these new properties. However, to practice systems biology, one must capture global sets of combinatorial biological data,

*To whom correspondence should be addressed. Tel: +86 10 6488 8543; Fax: +86 10 6487 7837; Email: crs@sun5.ibp.ac.cn
Correspondence may also be addressed to Yi Zhao. Tel: +86 10 6256 5533, ext. 5717; Fax: +86 10 6256 7724; Email: biozy@ict.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

such as protein–protein, protein–DNA and protein–RNA interactions.

The primary goal of NPInter is to extract and integrate the wealth of information about functional interactions between ncRNAs and PRMs into a user-friendly environment. ncRNAs are found in all analyzed organisms, and participate in numerous cellular processes. Regulatory processes involving ncRNA molecules are very common (3). The conservation of multiple ncRNA families is found across a wide taxonomic range (4).

Recently, novel ncRNAs and their functional interactions with other biomolecules are continuously being reported. Functional interaction experiments are, nevertheless, concentrated to the six major model organisms, such as *E.coli*, *S.cerevisiae*, *C.elegans*, *D.melanogaster*, *M.musculus* and *H.sapiens*. Over the recent years, several databases have been established to collect, organize and classify ncRNA sequences and information, such as Rfam (4), RNAdb (5), noncoding regulatory RNAs database (6) and NONCODE (7). Simultaneously KEGG (8), DIP (9), IntAct (10), BIND (11), MIPS (12) and some other biomolecular interaction databases have been established, dealing primarily with the protein–protein interactions. Despite the growing number of databases, no database has been established to particularly collect ncRNA functional interaction data, and no biomolecular network including an ncRNA component has been described. Therefore, NPInter was created to catalog experimentally determined functional interactions between ncRNAs and proteins, mRNAs or genomic DNA sequences.

DESCRIPTION AND STRUCTURE OF THE DATABASE

When we collected our data, we set up some criteria which were followed strictly. First, ncRNA functional interaction data were entered into NPInter only after publication in peer-reviewed journals. Second, there had to be a clear description of the experimental evidence (e.g. co-immunoprecipitation, yeast two-hybrid, *in vitro* binding assays, etc.) for the interaction in the paper. Third, based on the experiment, we specified the concerned organism and divided the interaction to two types, *in vivo* or *in vitro*. This whole process was performed manually by a curator and, thereafter, double-checked by a second curator.

NPInter is a relational database written in the programming language SQL. SQL efficiently handles diverse types of data and enables rapid sorting and analysis. The database can be conveniently extended as required without altering the existing database content, by adding new fields and tables to the data structure.

Each interaction entered into the NPInter has three main components: General Information, Molecule Information and Reference. The General Information provides users with basic information such as Interaction ID (unique ID in NPInter), names of interacting molecules, classification (e.g. interaction type), organism, experimental information and a description of the interaction. In Molecule Information, we describe each molecular entity (the ncRNA and the PRM) participating in the interaction. All interactions in NPInter are binary interactions. In reference, we provide information on the literature from which we have obtained the information of the interaction.

The NPInter database is composed of four linked tables:

(i) The ncRNA–protein related bimolecular interaction table describes the two interacting molecules (ncRNA and PRM) and the organism in which the interaction was studied. The table gives a description of the details of the interaction between two molecules and the interaction class (according to the new classification we suggest). It also contains data on the experiments used to detect the interaction. The experimental techniques represented in NPInter include co-immunoprecipitation, yeast two-hybrid, *in vitro* binding assays and a few others. Each interaction in the table includes the MEDLINE standard article code PMID.

(ii) The ncRNA information table contains ncRNA identification codes from NONCODE, GenBank database (Gene ID) and specific databases [WormBase (13), Flybase (14), etc.], as well as the gene name of each ncRNA, aliases, ncRNA class, organism and description of the ncRNA function.

(iii) The PRM information table contains PRM identification codes from Swiss-Prot (15), GenBank (16) database (Gene ID) and specific database (WormBase, Flybase, etc.), as well as each PRM's gene name, aliases, description and organism.

(iv) The Reference table details the literature citations in the interaction table. Each record in the table includes the MEDLINE standard article code (PMID), as well as general publication information.

THE NPInter CLASSIFICATION SYSTEM

Because ncRNAs participate in many different cellular interaction processes, we have introduced a classification system in which the ncRNA functional interactions are divided into eight different interaction processes. These are 'ncRNA binds protein', 'ncRNA regulates mRNA expression', 'ncRNA indirectly regulates a gene activity', 'ncRNA expression is regulated by protein', 'ncRNA affects protein activity', 'ncRNA activity is affected by protein', 'genetic interaction between ncRNA gene and protein gene' and 'other linkages' (see Supplementary Data: Supplementary I).

We use a few characters of the 'Oracle Bone Script' (the most ancient known form of Chinese written language, dating back to 4800 years before present) to symbolize the different classes of functional interactions. We believe that these pictographs should efficiently visualize the eight different interaction processes. To accommodate the non-Chinese user with a similar simple view of the basic content of each of the different types of interactions, we have also devised a set of 'expressions' using Latin letters (see Supplementary Data: Supplementary I).

DATABASE ACCESS

NPInter currently provides a search interface that can be used to search the database by names, IDs or text. For example, ncRNA name, protein name, their name alias, ncRNA class,

class alias, organism name all work as keywords, also do PubMed ID, ncRNA identifier in NONCODE and NCBI (17) Entrez, protein identifier in Swiss-Prot and NCBI Entrez, ncRNA and protein identifier in the species-specific databases [i.e. EcoCyc (18), SGD (19), etc.]. All sections in the database including interaction description, ncRNA description, protein description and experiments can be searched.

CURRENT STATE AND FUTURE DEVELOPMENTS

As of August 2005, the NPInter contains 700 pair-wise interactions and more than 40 classes of ncRNAs in 7 model organisms.

Although the NPInter has grown to its current state by manual entry of the interaction data, we plan to implement automatic literature search and text mining methods, such as the Textpresso (20), to update NPInter. The interaction data in *Arabidopsis thaliana* and *Xenopus laevis* will appear in NPInter's next update.

In the near future, we also hope to link the NPInter with other bimolecular interaction database to construct the first bimolecular interaction network containing the noncoding component.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the National High Technology Development Program of China under Grant No. 2002AA231031, National Key Basic Research, National Sciences Foundation of China Grant Nos 30500104 and 30570393 and Development Program 973 under Grant Nos 2002CB713805 and 2003CB715907. Funding to pay the Open Access publication charges for this article was provided by National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. (2004) The hidden genetic program of complex organisms. *Sci. Am*., **291**, 60–67.
2. Bompfünewerer,A.F., Flamm,Ch., Fried,C., Fritzsch,G., Hofacker,I.L., Lehmann,J., Missal,K., Mosig,A., Müller,B., Prohaska,S.J. *et al*. (2005) Evolutionary patterns of non-coding RNAs. *Theor. Biosci*., **123**, 301–369.
3. Barciszewski,J. and Erdmann,V.A. (2003) *Noncoding RNAs: Molecular Biology and Molecular Medicine*. Lands Bioscience, Georgetown.
4. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*., **33**, D121–D124.
5. Pang,K.C., Stephen,S., Engstrom,P.G., Tajul-Arifin,K., Chen,W., Wahlestedt,C., Lenhard,B., Hayashizaki,Y. and Mattick,J.S. (2005) RNAdb—a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res*., **33**, D125–D130.
6. Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2003) Noncoding regulatory RNAs database. *Nucleic Acids Res*., **31**, 429–431.
7. Liu,C., Bai,B., Skogerbo,G., Cai,L., Deng,W., Zhang,Y., Bu,D., Zhao,Y. and Chen,R. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res*., **33**, D112–D115.
8. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
9. 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.
10. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
11. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250.
12. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al*. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*., **32**, D41–D44.
13. Chen,N., Harris,T.W., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Canaran,P., Chan,J., Chen,C.K. *et al*. (2005) WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res*., **33**, D383–D389.
14. Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res*., **33**, D390–D395.
15. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31.**, 365–370.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res*., **33**, D34–D38.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
18. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res*., **33**, D334–D337.
19. Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Fisk,D.G., Hirschman,J.E., Hong,E.L., Nash,R. *et al*. (2005) Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the *Saccharomyces* Genome Database (SGD). *Nucleic Acids Res*., **33**, D374–D377.
20. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*., **2**, e309.