FEBS 25674 FEBS Letters 513 (2002) 135–140

Minireview

MINT: a Molecular INTeraction database

Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, Gianni Cesareni*

Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica, 00133 Rome, Italy

Received 20 November 2001; revised 29 November 2001; accepted 3 December 2001

First published online 20 December 2001

Edited by Gianni Cesareni and Mario Gimona

Abstract Protein interaction databases represent unique tools to store, in a computer readable form, the protein interaction information disseminated in the scientific literature. Well organized and easily accessible databases permit the easy retrieval and analysis of large interaction data sets. Here we present MINT, a database (http://cbm.bio.uniroma2.it/mint/index.html) designed to store data on functional interactions between proteins. Beyond cataloguing binary complexes, MINT was conceived to store other types of functional interactions, including enzymatic modifications of one of the partners. Release 1.0 of MINT focuses on experimentally verified protein–protein interactions. Both direct and indirect relationships are considered. Furthermore, MINT aims at being exhaustive in the description of the interaction and, whenever available, information about kinetic and binding constants and about the domains participating in the interaction is included in the entry. MINT consists of entries extracted from the scientific literature by expert curators assisted by 'MINT Assistant', a software that targets abstracts containing interaction information and presents them to the curator in a user-friendly format. The interaction data can be easily extracted and viewed graphically through 'MINT Viewer'. Presently MINT contains 4568 interactions, 782 of which are indirect or genetic interactions. © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Protein modules; Target recognition; Interaction networks; Peptide repertoires; Protein interaction

1. Introduction

Molecular interactions are at the hearth of cell physiology. In principle, if we knew the concentrations of each molecule in a cell and the details of the interaction network, including kinetic and thermodynamic parameters, we could be in the position to assemble a virtual cell in silico [1,2]. In practice information about molecular interactions is dispersed in the scientific literature and difficult to retrieve in a structured format. Furthermore, in the forthcoming few years we will be flooded by an exponentially increasing amount of interaction data from latest generation high-throughput technology [3,4]. Well organized and user-friendly databases containing information about molecular interactions are, and will become even more in the future, an essential resource for biologists. At least a couple of public databases, containing information about protein interactions, have been described and are currently accessible on the web [5,6]. The database of interacting proteins (DIP) documents experimentally determined protein–protein interactions. Presently it stores 10 432 interactions, 90% of which are from high-throughput experiments in microorganisms. Entries describing interactions among proteins from mammalian proteomes are approximately 750. BIND (Biomolecular Interaction Network Database) has a somewhat larger scope and contains information about bimolecular interactions, complexes and pathways. BIND contains 5939 interactions approximately 300 of which describe interactions among mammalian proteins. Even if these databases are still largely incomplete, they have provided bioinformaticians, and biologists in general, with a unique and easily searchable depository of interaction information. The progress in extending the coverage of these interaction databases is slowed by the time required, even for an expert annotator in extracting information from a published article and completing the entry. Several groups are developing software in order to be able, in the future, to extract reliable interaction information by automatic text mining [7–9]. Presently, however, these efforts are hampered by the heterogeneity of the language used in the scientific literature and by the absence of a uniform and universally accepted vocabulary of protein names. Thus databases that contain entries annotated by expert biologists, although requiring in principle a much larger effort, represent an unmatched collection of reliable interaction information and can be used as a benchmark for automatic methods and to test the performance of newly developed text mining approaches.

MINT was conceived to store molecular interaction data in a relational database and present it to the user in a friendly format so that it could be easily retrieved to extract meaningful and reliable biological information. The structure of MINT is such that it is not confined to physical interactions between proteins and can in principle be used to store any type of interaction between any 'molecular entity' of biological interest. The molecular entities can be either natural (proteins, nucleic acids, lipids) or artificial, in order to take into account the growing information load derived from screening of artificial repertoires with particular reference to peptide repertoires.

Most of the interactions presently stored in MINT are of the type entity A binds to entity B where the entities are proteins. As already stated, however, the entities in the database can in principle be any molecule in the cell. Some entities

*Corresponding author. Fax: (39)-6-2023500.
E-mail address: giovanni.cesareni@uniroma2.it (G. Cesareni).

0014-5793/02/$22.00 © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
PII: S0014-5793(01)03293-8

are 'primitive': genes, promoters, other regulatory sites, transcripts or the primary translation products of any transcript. The primitive entities are identified in a MINT table with the accession numbers and relative annotation extracted from primary databases like, for instance, SWISS-PROT and TREMBL for proteins [10]. Other entities like promoters and transcripts might be more difficult to define unequivocally at present because of the absence of well annotated reference databases. In this first release of MINT only interactions between proteins, or between a protein and peptide ligands selected from peptide repertoires, are considered and the reference database is SWISS-PROT/TREMBL.

Any entity can interact with any other entity in the database not only by binding it but also by performing a series of enzymatic activities including Y phosphorylates, Y dephosphorylates, S/T phosphorylates, S/T dephosphorylates, ubiquitinates, de-ubiquitinates, sumolates, acetylates, deacetylates, hydrolyzes and finally activates and inactivates.

The interaction between two primary entities leads to secondary entities that are identified by a MINT accession number. Secondary entities, like protein complexes or modified proteins (for instance phosphorylated proteins), are the result of the action of an entity onto another entity. Some of these actions, like for instance 'bind' are symmetric while actions implying an enzymatic activity are asymmetric (entity A modifies entity B) and provide a directionality to some parts of the interaction network.

Most of the interactions in the database are presently binary interactions (between two entities). It is becoming increasingly apparent, however, that the interaction network that one may derive from the assembly of many binary interactions is somewhat misleading or, at least, it does not represent all the biological wisdom contained in the scientific literature. For instance it is clear that molecules in the cell do not form a single large web but are organized in discrete complexes that can be isolated by standard affinity purification procedures. Furthermore, one would like to distinguish from the complex and intricate interaction web the set of interactions that represent a biological pathway.

Similarly to the BIND database, MINT permits to enter information about molecular complexes and biological pathways. We define a complex as the ensemble of entities that can be co-purified from a cell culture or a tissue. Although this definition is rather vague, the emerging technology that relies on affinity purification and the identification by mass spectrometry of the proteins that co-purify experimentally defines a 'complex'[11].

Differently from BIND we do not plan to predefine pathways (for instance the 'EGF pathway') and then edit the relevant interactions as outlined by experts in the field. However, MINT also contains indirect interactions between two entities. Genetic and 'long range' interactions belong to this class. For instance if the overexpression of protein A leads to suppression of a mutation in protein B, this information will be entered in an appropriate format as 'protein A' 'suppresses by overexpression' 'protein B'. We are currently defining a list of possible indirect interactions between two molecular entities. We envisage that biologically meaningful pathways should emerge from the thoughtful combination of this type of information and the information describing direct interactions.

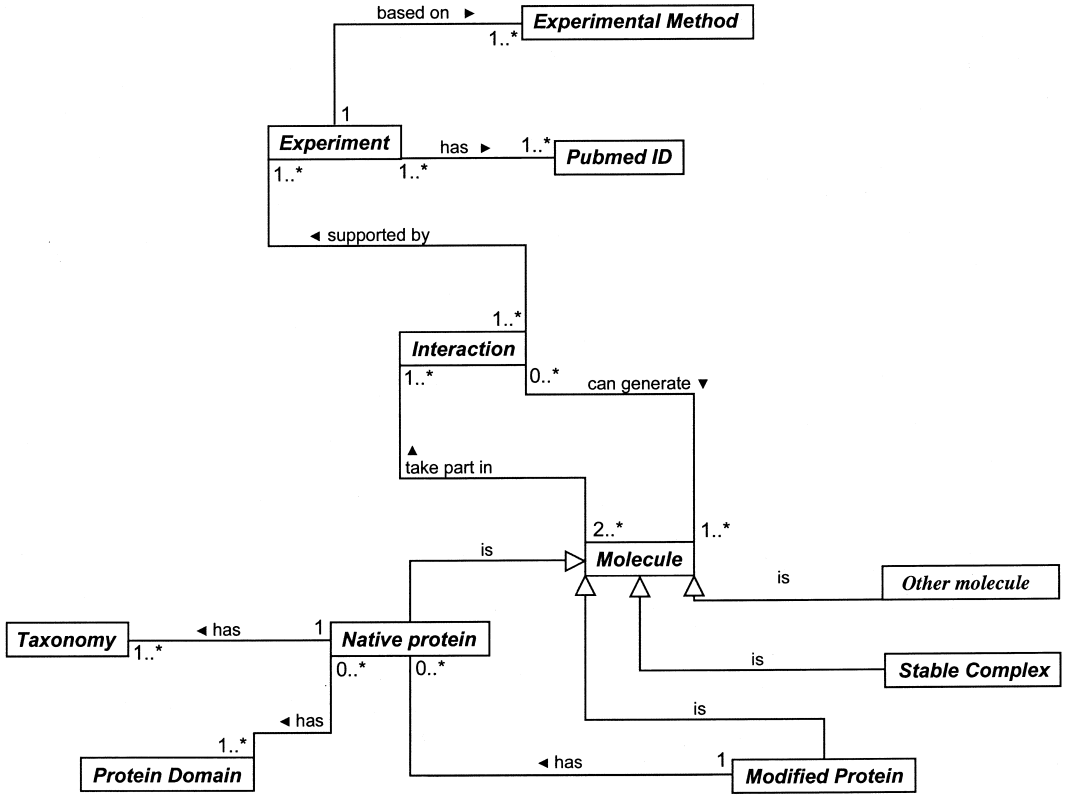

Fig. 1. MINT relational structure. The entities are represented as boxes and the relationships are indicated by a line connecting two entity symbols. Associated with each end of the line is a cardinality notation. The cardinality is indicated by a notation of the form 'lower-bound..upper-bound'. The asterisk character (*) denotes an unlimited upper bound. This diagram was drawn with DIA, an open source vector-based drawing tool (http://www.lysator.liu.se/~alla/dia/).

MINT a Molecular INTeractions database

#754

CELLULAR TUMOR ANTIGEN P53

Binary Interactions - Pathways - Complexes

ID	754
AC	P04637 (S)
Name(s)	cellular tumor antigen p53 tumor suppressor p53;phosphoprotein p53
Length	393 aa
Gene	TP53;P53
Organism	Homo sapiens (Human)
Function	acts as a tumor suppressor in many tumor types. induces growth arrest or apoptosis depending on the physiological circumstances or cell type, but both activities are involved in tumor suppression. acts in cell cycle regulation, is a trans- activator that acts to negatively regulate cell division by controlling a set of genes required for this process. one of the activated genes is an inhibitor of cyclin-dependent kinases. apoptosis induction seems to be mediated either by stimulation of bax and fas antigen expression, or by repression of bcl-2 expression
Subcelluar location	nuclear
Post-translational modifications	o-linked glycosylation in the c-terminal basic region was studied in eb-1 cell line
Genbank	AAA61212 AAA59988 AAA59989 CAA25652 AAA61211 CAA26306 AAC12971 AAD28628 CAA42627 CAA38095
Prosite	PS00348
Pfam	PF00870
PDB	1OLG 1OLH 1SAE 1SAF 1SAG 1SAH 1SAI 1SAJ 1SAK 1SAL 1TSR 1AIE 1PES 1PET
Keywords	Anti-oncogene DNA-binding Transcription regulation Activator Nuclear protein Phosphorylation Glycoprotein Apoptosis Alternative splicing Disease mutation Polymorphism 3D-structure Li-Fraumeni syndrome

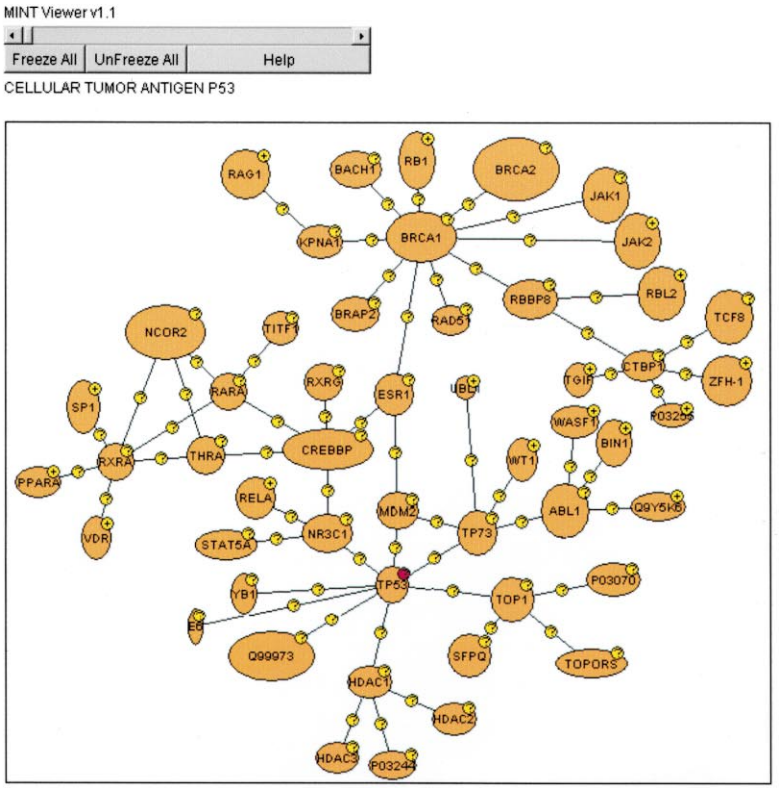

Fig. 2. A typical output page of a MINT query. The interaction web on the right was obtained through MINT Viewer by expanding the interaction network that nucleates around the tumor suppressor protein p53. By clicking on nodes or edges information about the protein or the interaction is displayed in the left frame respectively.

MINT is a relational database designed to collect and integrate experimental protein interaction data, in a unique database accessible via a user-friendly web interface written in an HTML embedded scripting language named PHP (http://www.php.net) The MINT core is stored in an SQL server (PostgreSQL). The entity relationship model underlying the database structure is shown in a simplified form in Fig. 1.

2. Data submission and MINT Assistant

MINT entries are curated by expert biologists who carefully screen the interaction information published in peer-reviewed journals. Presently the curator team consists of 20 Ph.D. students, in the program of Molecular and Cellular Biology of the University of Rome Tor Vergata, who fill appropriate web submission forms. However, any scientist is encouraged to register in order to be able to enter the interactions he is expert of.

Each entry contains a 'core' information consisting of the SWISS-PROT/TREMBL accession numbers of the two proteins and the specification of the functional interaction (binds, activates, phosphorylates...). Most of the entries in the database currently refer to a Pubmed identification (PMID) number. Unpublished observations, however, can also be added to the database. Furthermore, the curator can enter information about the domains that are demonstrated to be involved in the interaction, the binding and/or kinetic constants and the experimental method(s) utilized to characterize the interaction.

To support the curator task we have implemented a web tool named 'MINT Assistant'. This tool helps MINT curators in the identification of protein–protein interaction papers and assists them in the insertion of the relevant information into the database. The software scans titles and abstracts, extracted from the scientific literature, by counting words that are frequent in papers describing protein–protein interactions, essentially as already described [7].

The top ranking abstracts are further analyzed to identify protein names and whenever described in the abstract, experimental methods. To recognize protein names, each abstract is first compared with a simple scientific English dictionary to quickly eliminate common words. Any 'non-English' word is then matched to a vocabulary of 214 038 protein names extracted from SWISS-PROT, TREMBL and EMBL databases. When a protein name is identified, the program also registers the gene name, protein accession number and any other information that is required to complete a protein–protein interaction entry in MINT. The software output consists of several html pages that can be viewed by an internet browser. The front page displays the abstract titles, ranked according to the likelihood to contain information about protein interactions, as assessed by a statistical algorithm. By clicking a title the MINT curator has access to a new html page that provides the information needed to complete the entry. When an entry is completed, the information is stored in temporary tables where the data are automatically double-checked and then entered in the MINT database tables.

The number of entries was also increased by adding the interactions detected in genome-wide two-hybrid analysis [4,12–15]. The MIPS (Munich Information Center for Protein Sequences) yeast physical and genetic interactions tables [16] have also been incorporated into MINT.

3. Searching, browsing and visualizing a protein network

Searches can be performed via protein name, accession number or keywords. The search returns a list of entries containing the query names or keywords (only if present in the KEYWORDS line of the SWISS-PROT entries). By clicking the corresponding protein ID all the interactions described in MINT and having the selected protein as one of the partners are displayed. Each interaction ID in the output page is hyperlinked to a MINT entry. In order to produce the output, the information about a specific interaction is retrieved from one of the MINT tables and composed in two frames. The first frame contains information about the interacting proteins, the second shows the features of the interaction itself and the corresponding experimental procedure. Finally a third frame permits to display graphically the network of the interacting proteins as produced by 'MINT Viewer' (Fig. 2).

This tool is based on a java applet derived from the Sun's applet 'Graph' (http://java.sun.com) and adapted for use in our database. Proteins are represented by ovals whose size is proportional to molecular weight. Protein interactions are represented by lines (edges) connecting the proteins (nodes). Both nodes and edges are interactive and the action of clicking results in the display of additional information about the partner proteins and their interaction or in the expansion of the displayed network.

4. Current status of MINT

At 1 November 2001, the MINT database contains interaction information about 3556 proteins from 64 different organisms. These proteins participate in 3786 pairwise interactions, three multimeric complexes, and 782 'indirect' interactions. Furthermore, we have started to collect the published and unpublished experiments describing screening of peptide repertoires with protein recognition domains. 76% of the interactions rely on a single experimental procedure, mostly yeast two-hybrid (Fig. 3c,d). Nevertheless, as many as 206 interactions are supported by three independent approaches (Fig. 3d).

A large majority of the interactions are derived from large genomic projects. However, more than 700 articles have been processed manually by curators and 569 entries describe interactions between proteins of mammalian organisms. Cluster analysis of the MINT entries (Fig. 3a,b) reveals that, at the present stage, although most of the yeast proteins form a large cluster of 1537 proteins, most of the clusters range in size from two to eight proteins. Although the number of interactions for mammalian proteins is considerably lower, also in this case relatively large clusters begin to emerge. It is likely that, as the number of interactions in MINT is increased, the smaller clusters will merge into a single network.

5. Perspectives

One of the major efforts in the coming years will consist in the attempt to speed up the survey of the scientific literature in order to be able to enter a larger number of well curated interactions. We envisage that text mining software will play a major role in this. It is clear, however, that the most adequate annotators of any given entry are the scientists that have experimentally discovered the interactions. Scientific

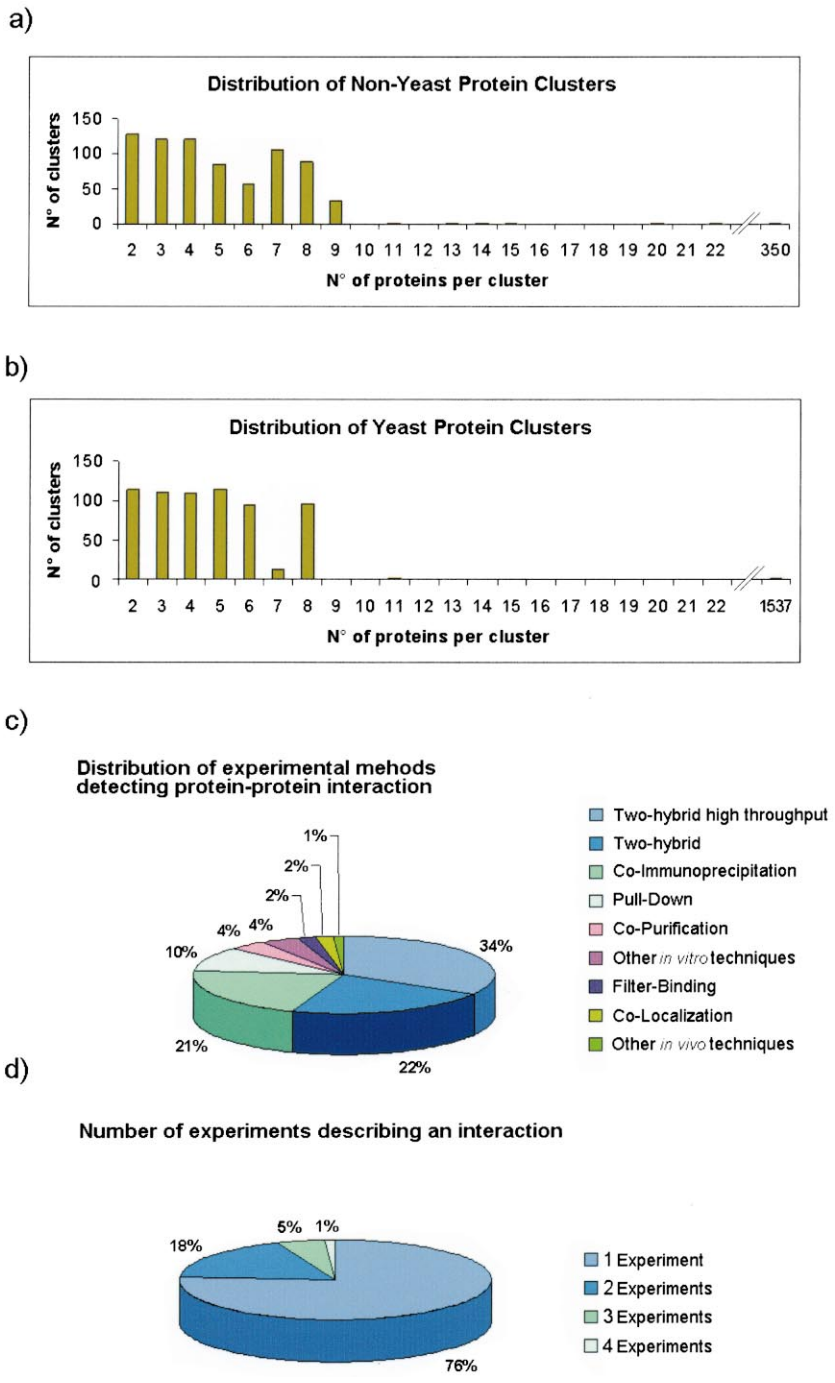

Fig. 3. MINT statistics. a: Distribution of the protein cluster sizes in all non-yeast proteins. The largest cluster contains 350 proteins. b: Distribution of the cluster sizes in the subset of yeast interactions. The largest cluster includes 1537 proteins. c: Pie chart representing the percentage of interactions described by several experimental methods. d: Pie chart describing the distribution of the interactions documented in MINT according to the number of experimental methods: about 76% of the interactions described were supported by a single experiment, 18% by two, 5% by three and 1% by four.

journals should ask the authors to submit interaction data to protein databases as they have done for DNA sequencing data and as they are doing for protein structure data. FEBS Letters has started, on an experimental basis, to ask authors to curate the interaction data described in their accepted manuscript and submit it to MINT.

Finally, although the co-existence of diverse interaction databases may play some role at this stage by stimulating discussion and by increasing the 'annotation effort', it is clear that these public endeavors should soon agree on a common database structure and pull together their efforts to avoid duplications.

Acknowledgements: This work is supported by an AIRC grant and by EU Grant QLRI-CT-2000-00127. We like to thank the students of the Ph.D. program in Molecular and Cellular Biology of the University of Rome Tor Vergata for help in database curation. We are grateful to the SMART team, for their help and cooperation. We also thank the staff of the SWISS-PROT and MIPS databases for providing data.

References

[1] Schaff, J.C., Slepchenko, B.M. and Loew, L.M. (2000) Methods Enzymol. 321, 1–23.
[2] Cagney, G., Uetz, P. and Fields, S. (2000) Methods Enzymol. 328, 3–14.
[3] Mendelsohn, A.R. and Brent, R. (1998) Nat. Biotechnol. 16, 520–521.
[4] Uetz, P. et al. (2000) Nature 403, 623–627.
[5] Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M. and Eisenberg, D. (2001) Nucleic Acids Res. 29, 239–241.
[6] Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T. and Hogue, C.W. (2001) Nucleic Acids Res. 29, 242–245.
[7] Marcotte, E.M., Xenarios, I. and Eisenberg, D. (2001) Bioinformatics 17, 359–363.
[8] Blaschke, C., Andrade, M.A., Ouzounis, C. and Valencia, A. (1999) Proc. Int. Conf. Intell. Syst. Mol. Biol., 60–67.
[9] Andrade, M.A. and Valencia, A. (1998) Bioinformatics 14, 600–607.
[10] Bairoch, A. and Apweiler, R. (2000) Nucleic Acids Res. 28, 45–48.
[11] Andersen, J.S. and Mann, M. (2000) FEBS Lett. 480, 25–31.
[12] Ito, T. et al. (2000) Proc. Natl. Acad. Sci. USA 97, 1143–1147.
[13] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) Proc. Natl. Acad. Sci. USA 98, 4569–4574.
[14] Bartel, P.L., Roecklein, J.A., SenGupta, D. and Fields, S. (1996) Nat. Genet. 12, 72–77.
[15] McCraith, S., Holtzman, T., Moss, B. and Fields, S. (2000) Proc. Natl. Acad. Sci. USA 97, 4879–4884.
[16] Mewes, H.W. et al. (2000) Nucleic Acids Res. 28, 37–40.