BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 5 2009, pages 690–691
doi:10.1093/bioinformatics/btp025

Databases and ontologies

MatrixDB, a database focused on extracellular protein–protein and protein–carbohydrate interactions

Emilie Chautard[1], Lionel Ballut[1], Nicolas Thierry-Mieg[2,*] and Sylvie Ricard-Blum[1,*]

[1]UMR 5086 CNRS - Université Lyon 1, 7 passage du Vercors, 69367 Lyon Cedex 07 and [2]TIMC-IMAG, UMR 5525 CNRS - Université Grenoble 1, Faculté de Médecine, 38706 La Tronche Cedex, France

Received on October 23, 2008; accepted on January 8, 2009

Advance Access publication January 15, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Summary: MatrixDB (http://matrixdb.ibcp.fr) is a database reporting mammalian protein–protein and protein–carbohydrate interactions involving extracellular molecules. It takes into account the full interaction repertoire of the extracellular matrix involving full-length molecules, fragments and multimers. The current version of MatrixDB contains 1972 interactions corresponding to 4412 experiments and involving 259 extracellular biomolecules.

Availability: MatrixDB is freely available at http://matrixdb.ibcp.fr

Contact: nicolas.thierry-mieg@imag.fr; s.ricard-blum@ibcp.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Most of the cells in multicellular organisms are surrounded by an extracellular matrix (ECM), which is comprised of proteins (~500 genes) and complex polysaccharides termed glycosaminoglycans (GAGs). GAGs play crucial roles either within the ECM or at the cell surface (Rodgers *et al*., 2008). A global view of the extracellular network is required to understand how the ECM conveys information, and how its assembly is regulated. Interactions involving extracellular biomolecules are under-represented in interaction databases, which mainly focus on interactions occurring inside the cell and report few protein–carbohydrate interactions. Furthermore a number of extracellular proteins are multimeric, and are able to form supramolecular assemblies either alone (Ricard-Blum *et al*., 2005) or in association with other molecules (Kielty, 2006). This is important because the oligomerization state may modulate the ability of proteins to interact with other molecules (Jokinen *et al*., 2004). Enzymatic cleavage of full-length extracellular biomolecules also affects the interaction network by releasing fragments with specific molecular recognition properties referred to as matricryptins (Davis *et al*., 2000, Ricard-Blum *et al*., 2005). Matricryptins are enzymatic fragments of ECM containing exposed biologically active cryptic sites (matricryptic sites) that are revealed after structural or conformational alteration of these molecules. Matricryptic sites and matricryptins have been reported within protein components of the ECM as well as in GAGs (Davis *et al*., 2000). We have built a database to report protein–protein and protein–carbohydrate interactions involving extracellular molecules. MatrixDB takes into account all the above features and mainly focuses on mammalian molecules. It integrates data from four major interaction databases, in-house literature curation and binding experiments performed with protein and polysaccharide arrays. The current version of MatrixDB contains 1972 interactions involving 259 extracellular biomolecules. MatrixDB is a member of the International Molecular Exchange (IMEx) consortium.

*To whom correspondence should be addressed.

2 BUILDING MATRIXDB

2.1 Biomolecule sources

Protein data were imported from the UniProtKB/Swiss-Prot database (Bairoch *et al*., 2005) and identified by UniProtKB/Swiss-Prot accession numbers. In order to list all the partners of a protein, interactions are associated by default to the accession number of the human protein. The actual source species used in experiments is indicated in the page reporting interaction data. Intracellular and membrane proteins were included to obtain a comprehensive network of the partners of extracellular molecules. Indeed, ECM proteins and GAGs bind to a number of membrane proteins or cell-associated proteoglycans and some of them interact with intracellular partners upon internalization (Dixelius *et al*., 2000). ECM proteins were identified by the UniProtKB/Swiss-Prot keyword 'extracellular matrix' and by the GO terms 'extracellular matrix', 'proteinaceous extracellular matrix' and their child terms. The proteins annotated with the GO terms 'extracellular region' and 'extracellular space', which are used for proteins found in biological fluids, were not included because circulating molecules do not directly contribute to the extracellular scaffold. Additionally, 96 proteins were manually (re-)annotated through literature curation. These new or corrected annotations will be submitted to UniProtKB/Swiss-Prot. Glycan-binding proteins were annotated with cross-references to the KEGG GLYCAN database when available (Hashimoto *et al*., 2006). We created specific identifiers for fragments (PFRAG_number), multimers (MULT_number), cations (CAT_number), lipids (LIP_number) and GAGs (GAG_number), with cross-references to the ChEBI database when available (Degtyarenko *et al*., 2008).

2.2 Interaction sources

MatrixDB integrates 1378 interactions from the Human Protein Reference Database (HPRD, Prasad *et al*., 2009), 211 interactions

from the Molecular INTeraction database (MINT, Chatr-Aryamontri *et al*., 2007), 46 interactions from the Database of Interacting Proteins (DIP, Salwinski *et al*., 2004), 232 interactions from IntAct (Kerrien *et al*., 2007a) and 839 from BioGRID (Breitkreutz *et al*., 2008) involving at least one extracellular biomolecule of mammalian origin. We added 283 interactions from manual literature curation and 65 interactions from protein and GAG array experiments. All interaction data, except those imported from HPRD and BioGRID, are reported according to the MIMIx standard (Minimum Information required for reporting a Molecular Interaction experiment, Orchard *et al*., 2007). The coverage of the extracellular interaction network is not complete and MatrixDB is regularly updated.

2.3 Architecture of the database

We used AceDB, a database system developed by Durbin and Thierry-Mieg (1994). SwissKnife (Hermjakob *et al*., 1999) was used to extract data from UniProtKB/Swiss-Prot. HPRD, DIP, MINT, IntAct and BioGRID data, available in PSI-MI XML 2.5 (Kerrien *et al*., 2007b), were transformed into MatrixDB format. AcePerl (Stein and Thierry-Mieg, 1998) and AceBrowser, a set of Perl-CGI scripts, provided a customizable and straightforward browsable interface to MatrixDB. An Apache HTTP server hosts MatrixDB, which is freely available at http://matrixdb.ibcp.fr

3 QUERYING MATRIXDB

All members of a category (i.e. all the extracellular molecules, proteins, fragments, multimers, GAGs, lipids or cations) can be displayed. Biomolecules can be searched by their common name, UniProtKB/Swiss-Prot (or CheBI for GAGs) accession number or gene name. GO terms, UniProtKB/Swiss-Prot keywords, PubMed IDentifiers or author names can be used to query the database. Interactions from each database can be retrieved. All the binding partners of a biomolecule are listed on the 'Biomolecule Report' page. The list of experiments reporting an interaction is accessible on the 'Association Report' page. Experimental data (binding site, kinetic and affinity constants) are displayed on the 'Experiment Report' page. Selected molecules can be saved in a cart for building the corresponding interaction network. A tutorial is available on-line and interaction data can be downloaded from the MatrixDB website in PSI-MI XML 2.5 and MITAB 2.5.

4 VISUALIZATION OF INTERACTION NETWORKS

We have developed a set of scripts based on Cytoscape (Shannon *et al*., 2003) to visualize the extracellular interaction network or sub-networks. Biomolecule types are shape- and color-coded. Besides non-covalent physical interactions, links are displayed between a multimer and its constitutive monomers, a fragment and its parent molecule, and the protein and GAG moieties of proteoglycans (Fig. 1, Supplementary Material). Sub-cellular localization of the biomolecules can be visualized. Biomolecule and interaction data are imported into Cytoscape and displayed when clicking on a node (biomolecule) or on an edge (link) of the network. Alternatively, interaction networks can be visualized on MatrixDB website using Medusa (Hooper and Bork, 2005).

5 CONCLUSIONS

The architecture of MatrixDB takes into account the structural and functional complexity of the ECM organization, including protein–carbohydrate interactions. The analysis of the extracellular interaction network will be used to investigate the mechanisms of ECM assembly and homeostasis, and to determine how genetic and acquired diseases interfere with these processes.

ACKNOWLEDGEMENTS

We thank R. Salza and M. Fatoux for their help in literature curation and C. Blanchet for his strong support in building the web site.

Funding: Contrat de Plan Etat-Région Rhône-Alpes; Institut Rhône-Alpin des Systèmes Complexes (to S.R.B., E.C. and N.T.M.).

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Breitkreutz,B.J. *et al*. (2008) The BioGRID interaction database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.

Chatr-Aryamontri,A. *et al*. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.

Davis,G.E. *et al*. (2000) Regulation of tissue injury responses by the exposure of matricryptic sites within extracellular matrix molecules. *Am. J. Pathol.*, **156**, 1489–1498.

Degtyarenko,K. *et al*. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*., **36**, D344–D350.

Dixelius,J. *et al*. (2000) Endostatin-induced tyrosine kinase signaling through the Shb adaptor protein regulates endothelial cell apoptosis. *Blood,* **95**, 3403–3411.

Durbin,R. and Thierry-Mieg,J. (1994) The ACeDB genome database. In Suhai,S. (ed.) *Computational Methods in Genome Research.* Plenum Press, New York, pp. 45–55.

Hashimoto,K. *et al*. (2006) KEGG as a glycome informatics resource. *Glycobiology*, **16**, 63R–70R.

Hermjakob,H. *et al*. (1999) Swissknife-'lazy parsing' of SWISS-PROT entries. *Bioinformatics*, **15**, 771–772.

Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.

Jokinen,J. *et al*. (2004) Integrin-mediated cell adhesion to type I collagen fibrils. *J. Biol. Chem*., **279**, 31956–31963.

Kerrien,S. *et al*. (2007a) IntAct - open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.

Kerrien,S. *et al*. (2007b) Broadening the horizon-level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.

Kielty,CM. (2006) Elastic fibres in health and disease. *Expert Rev. Mol. Med.*, **8**, 1–23.

Orchard,S. *et al*. (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, **25**, 894–898.

Prasad,T.S. *et al*. (2009) Human Protein Reference Database - 2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Ricard-Blum,S. *et al*. (2005) The collagen superfamily. *Curr. Topics Chem*., **247**, 7–33.

Rodgers,K.D. *et al*. (2008) Heparan sulfate proteoglycans: a GAGgle of skeletal-hematopoietic regulators. *Dev. Dyn.*, **237**, 2622–2642.

Salwinski,L. *et al*. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Stein,L.D. and Thierry-Mieg,J. (1998) Scriptable access to the Caenorhabditis elegans genome sequence and other ACEDB databases. *Genome Res.*, **8**, 1308–1315.