

The ABC (Analysing Biomolecular Contacts)-database

Peter Walter, Sam Ansari and Volkhard Helms

Center for Bioinformatics, Saarland University, 66041 Saarbrücken, Germany

Abstract

As protein-protein interactions are one of the basic mechanisms in most cellular processes, it is desirable to understand the molecular details of protein-protein contacts and ultimately be able to predict which proteins interact. Interface areas on a protein surface that are involved in protein interactions exhibit certain characteristics. Therefore, several attempts were made to distinguish protein interactions from each other and to categorize them. One way of classification are the groups of transient and permanent interactions. Previously two of the authors analysed several properties for transient complexes such as the amino acid and secondary structure element composition and pairing preferences. Certainly, interfaces can be characterized by many more possible attributes and this is a subject of intense ongoing research. Although several freely available online databases exist that illuminate various aspects of protein-protein interactions, we decided to construct a new database collecting all desired interface features allowing for facile selection of subsets of complexes. As database-server we applied MySQL and the program logic was written in JAVA. Furthermore several class extensions and tools such as JMOL were included to visualize the interfaces and JfreeChart for the representation of diagrams and statistics. The contact data is automatically generated from standard PDB files by a tcl/tk-script running through the molecular visualization package VMD. Currently the database contains 536 interfaces extracted from 479 PDB files and it can be queried by various types of parameters. Here, we describe the database design and demonstrate its usefulness with a number of selected features.

Availability: The ABC-database can be accessed on <http://service.bioinformatik.uni-saarland.de/abc>.

1 Introduction

Protein-protein interactions play an important role for most biological processes such as signal transduction, the formation of enzyme inhibitor complexes, or metabolic reactions. Elucidating the principles of protein-protein interactions is likely to help us in various areas of interest. It might be beneficial for the prediction of putative binding sites when the structure of the complex is unknown. Characterizing the interface properties of a given protein complex may provide information about its function and its role in a cellular pathway [1,2]. It also facilitates the discovery of new ligands that exert an effect on interfaces which is of great interest for the pharmaceutical research. As a last example, characterizing the properties of typical protein interfaces may be useful for the prediction and evaluation of docked complexes [3].

Before addressing the main issues related to protein interactions, the definition of the interaction area should be specified here. An interface is denoted as the regions on the surfaces of two different protein chains that lie within a certain distance range from each other in the bound conformation. For our considerations we choose distance cutoffs from 4Å to 8Å in steps of 1Å that covers most distance-based definitions in literature. According to an alternative definition, an interface may be characterized by the residues exhibiting a certain change in surface accessible area (ASA) when comparing the values of the unbound state

with those of the complexed form [4]. The strictness of both definitions can be adjusted via the variable parameters distance-cutoff or Δ ASA. The groups of atoms selected according to the two criteria will generally be quite similar but not identical. In any case the residues of a typical interface do not have to form a contiguous chain. Instead, most interfaces are interrupted by parts that lie beyond the distance cutoff. Speaking from a physicochemical view, the residues of the interface are not covalently connected. In summary, an interface is a spatial entity that is responsible for a protein interaction.

The question may now come up whether one can identify interfaces on the basis of surface properties in the absence of experimental information. First of all, not every complex structure determined by X-ray crystallography contains biologically meaningful packings. Many contacts are only artifacts that are formed during the crystallization process. One goal of structural bioinformatics consists in finding appropriate attributes that characterize biological interface areas in order to distinguish them from non-biological interface regions. Several authors found that interface regions differ in various aspects from the remaining surface [5,6]. Bahadur et al., for instance, examined the distinction of specific and non-specific interactions by applying preference scores based on residue propensities and hydrophobicity [7]. Another approach considered the degree of residue conservation in a multiple sequence alignment for a family of related sequences [8]. Obviously, an interface region tends to be more conserved than the remaining protein surface because its biological function is likely maintained during evolution. Aloy et al. have shown that binding modes are conserved in families of homologous protein pairs with sequence identities of 30% and higher [9]. Besides, it is possible to define geometric and compositional attributes that allow distinctions among interfaces [10].

Because of the vast number of different biological functions there may be many different modes of interaction and constraints on interfaces. As general rules may not be applicable to all types of interfaces, it makes sense to introduce a classification for interfaces. Ofra et al. suggested six types of interfaces that differ from each other with respect to amino acid composition and residue pair propensities [11]. One of the most accepted classifications introduced by Noreen et al. is the distinction of protein interactions via the lifetime of the complex [12]. On the basis of this, one can split the group of interacting pairs in transient and permanent interactions. The former describes proteins that exhibit short-time interactions, the latter comprises proteins that are strongly connected leading to a long-term interaction. An example of permanent interactions are antigen-antibody complexes. The recognition of an antigen by an antibody should result in an irreversible binding between the two components leading to the inactivation of the foreign body. According to Ansari et al. enzyme inhibitor complexes, transport mechanisms, signalling transduction and hormone receptor system are typical examples of transient interactions [13]. Two or more proteins bind to each other to cause a certain effect or to inhibit the function of one partner and dissociate after some time. A further classification of interfaces after Noreen et al. refers to the stability of the complex. An interaction is denoted as obligate if the elements of the complex are only existent in the bound state. Non obligate interactions are those that can exist separately from each other. The last definition considered here addresses structural properties. Homodimeric interactions are made of two equal or homologous molecules, whereas the binding partners in heterodimeric interactions are of different nature [12].

One important application of statistical analysis on protein-protein interfaces is the separation of transient and permanent interactions. This classification is certainly strongly related to the functional aspects of protein-protein interactions. For example, transient interactions may play an important role in regulatory processes whereas permanent interactions are responsible for the structural integrity of protein complexes. Although exceptions exist, transient interactions often turn out to be heterodimeric and non-obligate whereas permanent interactions are

homodimeric and obligate. Schröder and co-workers have presented a machine learning algorithm that distinguishes both classes with 97% accuracy [14]. We recently achieved a comparable accuracy on our data set based on different features (unpublished results). This very successful classification problem provides a strong stimulus to tackle further classification tasks such as finding a correlation of binding constants and interface properties and correlating protein function with interface properties, for instance whether interfaces of signalling proteins are different from bioenergetic partners etc.

A number of databases dealing with various aspects of protein interactions already exist. Many of these are freely available on the Internet. AffinDB concentrates on protein-ligand interactions and provides information about affinity data and experimental conditions [15]. KDBI focuses on kinetic data only [16]. The database is searchable by molecule name and other classifiers. DIP gives an overview of proteins that form an interaction with each other [17]. SCOPPI deals with the classification and annotation of domain interactions. The database can be queried on SCOP and PDB identifiers and allows a comprehensive view of interface residues including conservation and physicochemical type [18]. PDBSum offers an alternative view of PDB files [19]. Interface residues can be graphically represented and statistics shed light on the distribution of amino acid-groups and types of contacts. Although all these databases provide a rich source of information we found it necessary to compile another database including additional features facilitating our analysis of protein contacts.

2 ABC database

Here we introduce the ABC database, short for ‘Analyzing Biomolecular Contacts’, an easy-to-use online database for the statistical analysis of biomolecular contacts. It is available at <http://service.bioinformatik.uni-saarland.de/abc>.

2.1 Database design

The relational database contains detailed information about protein-protein interfaces. An overview of the main relations is shown in figure 1. Every entry in ‘datasets’ stands for a certain interface and contains attributes such as the PDB id, PDB header and surface sizes of the complex. An interface consists of a number of residue pairs that are stored in the ‘contacts’ relation. A single residue pair can be considered at the atomic level. We distinguish between atoms belonging to the side-chain or to the backbone of the amino acid. Therefore, interacting residues may involve either side-chain/side-chain, side-chain/backbone or backbone/backbone atoms. The distribution is stored in the ‘sidechain/backbone’ relation. An interface is only a part of the chains that are involved in the interaction. For a certain interface all residues that belong to the two chains are stored in the ‘sequence’ relation. The ‘aa composition’ may be considered as an interface profile. It comprises the amino acid-composition of the interfaces split according to different distance criteria and chains. The ‘secondary structure composition’ relation and the ‘sidechain/backbone composition’ relation are constructed accordingly. Every interface is assigned one or more CATH, SCOP and Uniparc classifications using separate relations.

When taking a closer look at the schema it can be easily seen that some data are stored twice in different relations or that one piece of information can be derived from another source. For instance the number of amino acids belonging to a certain interface can be calculated by summing up all appropriate residues in the contacts relations instead of retrieving this information from the ‘aa composition’ relation. We decided to allow such redundancy because it allows a faster retrieval of certain data and an easier formulation of query terms.

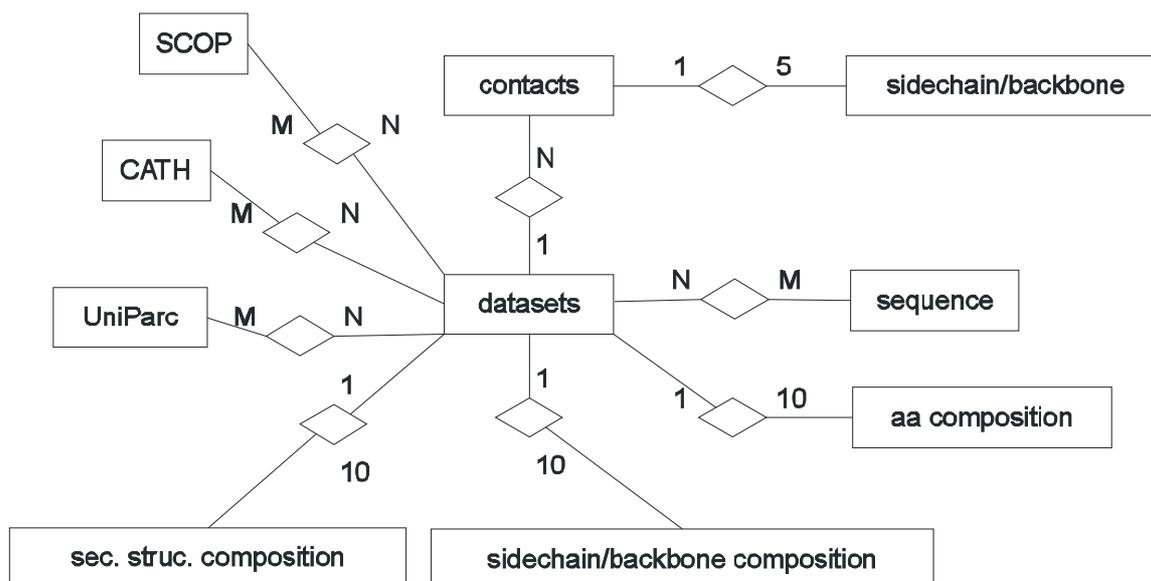


Figure 1: ER-diagram. 'datasets' is the central entity in the schema and represents the interfaces. Every interface may be assigned more than one CATH, SCOP and UniParc classification and vice versa. Therefore the classification-entities are linked with datasets by an N:M-relation. Every interface in datasets consists of a number of amino acid pairs that are represented by the 'contacts'-entity. A contact pair exhibits a number of side-chain/backbone compositions depending on the distance criterion. As we consider a range between 4Å and 8Å in steps of 1Å, we have a 1:5 relation. Similarly the relations between 'datasets' and 'aa composition', 'sec. struc. composition' and 'sidechain/backbone composition' come about. In addition to the range, the number of chains has to be taken into consideration resulting in a 1:10 relation. An interface including the chain consists of a number of residues. A residue may belong to more than one interface if the commensurating chain is involved in more than one interaction. This leads to the N:M relation between 'datasets' and 'sequence'.

2.2 Data Import

Up to the time of writing the database comprises structural and compositional data for 536 transient and permanent interfaces selected from the RCSB Protein Data Bank that constitutes the most important source for three-dimensional structure information about protein interactions [20]. Currently it contains approximately 19,000 putative protein complexes with two or more chains. We obtained suitable PDB identifiers for transient and permanent protein complexes from the literature [13, 21-25]. For every PDB file a tcl/tk script was applied under VMD that extracts the data which were of interest to us [26]. The parser searches for residue pairs that fulfil the distance criterion. By applying all interface criteria mentioned above, the script outputs residue name, position, secondary structure elements, interacting atoms and surface area. The results were written in a XML file for which we wrote an import filter under JDOM, a class extension that offers useful methods for the import and export of XML data [27]. The average size of the files was about 2-10 megabytes so that we decided to use the SAX parser. Besides, we included some additional information such as CATH or SCOP classification [28,29]. All these data are accessible for download from the commensurating websites. For some complexes kinetic data (binding constants or association rates) available in the literature were also imported [30].

2.3 Implementation

The database and webinterface were created with non-commercial or freely available tools. The program logic was written in JAVA. With Servlets and JSP it offers useful class

extensions especially meant for the development of dynamic web pages [31]. As webserver we use Tomcat for the management of Servlets and JSP pages as well as Apache for the static web content [32,33]. MySQL is one of the most popular relational databases that are free for use [34]. We chose a relational database system since they are wide-spread and well established [35]. Further open-source tools were included in our project such as the JAVA based class extension JfreeChart for the visual representation of statistics [36]. Figure 2 shows a diagram made by this program. We also implemented the Jmol-applet for a three-dimensional interactive view of molecules [37]. This program used in many bioinformatics related sites offers a large number of functions and a Javascript based programming interface for implementing additional features. Using an applet makes appliance easier for the user since no installation on the local machine is required.

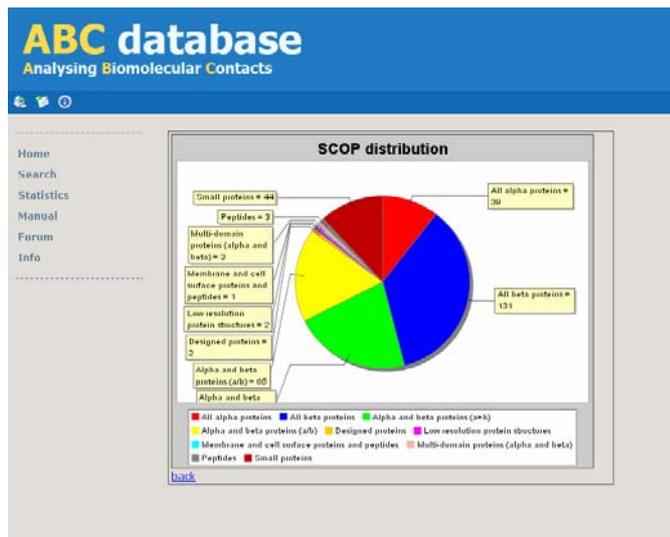
2.4 Database Access

The database can be queried by various types of parameters. First of all one can select data by choosing classification criteria such as the CATH or SCOP denotation. Every term can be linked together with logic operators 'AND' or 'OR' or 'AND NOT'. It is as well possible to apply filter criteria such as the composition of amino acids, composition of secondary structure elements or the pairing propensities between residues or secondary structure elements. Additionally, one can search for entries with respect to kinetic criteria such as certain k_{on} or k_{off} -ranges. After performing a query the user obtains a list of interfaces that fit to these constraints. If desired, further information can be displayed for every entry found such as an overview of the classification types mentioned above, statistics of the amino acid composition, and a list of the residue-residue contacts. The user can also view the three-dimensional structure of the interface residues or of the whole complex. The search result can be further refined by applying all filter options again to reduce the number of complexes that were found in the previous query.

As an additional feature we implemented the calculation of a scoring matrix. From the result list one can select one or several entries that are used for the computation. Various types of matrices can be displayed according to user-defined criteria. One matrix type belongs to the composition of amino acid pairs, another one to the pairing preferences of secondary structure elements. Molar fraction or contact fraction matrices are based on the amino acid matrix, but the values are normalized. An example is shown in figure 3. Processing the data in third party applications is facilitated by a function allowing the export in plain text or Microsoft Excel formats.

The following query example demonstrates an application of the database. For instance, one may be interested in finding PDB's that contain a serine protease and at least 50% lipophilic residues at the interface. In the query menu the first condition is represented in the classification area and the second one in the filter criteria section. First of all the user searches for serine protease complexes by selecting 'PDB header' as classification identifier and entering suitable keywords using wildcards as shown in figure 4a. After performing the query every entry in the resulting list can be checked so that the user has the opportunity to exclude improper hits. Then the query refinement is applied. The filter criterion is defined by choosing an appropriate interval for the amount of lipophilic residues (figure 4b). The final result list comprises the complexes fulfilling the aforementioned conditions (figure 4c). If desired one can download the list of PDB identifiers including the commensurating chain pair in plain text format or a scoring matrix can be computed as mentioned above.

a



b

c

Figure 2: Database screenshots: (a) shows the query screen of the ABC database. The diagram in (b) was made by JfreeChart and represents the distribution of SCOP classifiers in our dataset. A result list is represented as table in (c). The three-dimensional structure of an interface or complex can be displayed with the integrated molecule viewer JMol.

a

b

Figure 3: Scoring matrices calculated by ABC. The 20x20 matrix contains the pairing propensities of amino acids with each other. The values are normalized according to the molar

fraction. (a) shows a matrix derived from interfaces of permanent complexes, (b) is based on those of transient complexes.

Figure 4 consists of three panels. Panel (a) shows a search form for classification with three criteria: 'PDB header' with value '*serine*', 'PDB header' with value '*prot*', and 'PDB ID' with an empty field. Logical operators 'and', 'or', and 'and not' are available between criteria. Panel (b) shows a search form for amino acid properties with 'lipophilic' selected and a range of '75 % +/- 25 %'. Panel (c) shows a table of search results with columns for PDB ID, type, and description.

PDB ID	Type	Description
1ACB	EI	SERINE PROTEASE
1A7Z	AC	SERINE PROTEASE/INHIBITOR
1CA0	BD	SERINE PROTEASE/INHIBITOR
1CSE	EI	SERINE PROTEINASE-INHIBITOR
1S1B	EI	SERINE PROTEASE/INHIBITOR COMPLEX
1SPB	PS	SERINE PROTEINASE/PROSEGMENT
2SEC	EI	SERINE PROTEINASE-INHIBITOR
2TEC	EI	SERINE PROTEINASE-INHIBITOR
3TEC	EI	SERINE PROTEINASE-INHIBITOR

Figure 4: (a), (b) show excerpts of the search form that are related to the query example described in the text. The resulting list of hits is shown in (c)

3 Outlook

In the near future we plan to enhance the analytical power of the database by implementing further tools such as BioJAVA or NeoBio and further attributes like the conservation degree of interface residues. To further increase the current datasets we plan to screen putative protein-protein contacts from the PDB database with the help of a support vector machine. As reflected by the expression 'biomolecular' instead of only 'protein' in the name of our database other types of biological contact elements may be included in future work such as interactions between proteins and small molecules although our current work focuses on protein-protein interactions. We are certain that this database may facilitate the work of other researchers too.

4 Acknowledgement

We thank Carla Haid for compiling a list of experimental kinetic data and Sikander Hayat for comments on this manuscript.

5 References

- [1] S. Jones and J.M. Thornton. Prediction of Protein-Protein Interaction Sites using Patch Analysis. *J. Mol. Biol.*, 272: 133-143, 1997.
- [2] O. Keskin, C.J. Tsai, H. Wolfson and R. Nussinov. A new structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.*, 13: 1043-1055, 2004.
- [3] I. Halperin, B. Ma, H. Wolfson and R. Nussinov. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Function. *Proteins*, 47:409-443, 2002
- [4] H. Zhu, F.S. Domingues, I. Sommer and T. Lengauer. NOXClass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7: 27, 2006.
- [5] H.X. Zhou and Y. Shan. Prediction of Protein Interaction Sites from Sequence Profile and Residue Neighbor List. *Proteins*, 44: 336-343, 2001.

- [6] X. Gallet, B. Charlotheaux, A. Thomas and R. Brasseur. A Fast Method to Predict Protein Interaction Sites from Sequences. *J.Mol.Biol.*, 302: 917-926, 2000.
- [7] R.P. Bahadur, P. Chakrabarti, F. Rodier J. Janin. A Dissection of Specific and Non-specific Protein-Protein Interfaces. *J. Mol. Biol.*, 336: 943-935, 2004.
- [8] D.R. Caffrey, S. Shyamal, D.J. Hughes, J. Mintseris and E.S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein. *Prot. Science*, 13:190-202, 2004.
- [9] P. Aloy, H. Ceulemans, A. Stark and R.B. Russell. The relationship between sequence and interaction divergence in proteins. *J.Mol.Biol.*, 332, 989-998, 2003.
- [10] J. Jones and J. Thornton. Principles of protein protein interactions. *Proc. Natl. Acad. Sci. USA*, 93:13-20, 1996.
- [11] Y. Ofra and B. Rost. Analysing six types of protein protein interfaces. *J.Mol.Biol.*, 325: 377-387, 2003.
- [12] I. Noreen and J. Thornton. Diversity of protein protein interactions. *EMBO J*: 22(14): 3486-92, 2003.
- [13] S. Ansari and V. Helms. Statistical analysis of predominantly transient protein-protein interactions. *Proteins*, 61: 344-355, 2005.
- [14] S. Kottha and M. Schröder. Classifying permanent and transient protein interactions. *Proceedings of German Bioinformatics Conference GCB2006*, 2006.
- [15] P. Block, C.A. Sotriffer, I. Dramburg and G. Klebe: AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.*, 34, 522-526, 2006.
- [16] Z.L. Ji, X. Chen, C.J. Zhen, L.X. Yao, L.Y. Han, W.K. Yeo, P.C. Chung, H.S. Puy, Y.T. Tay, A. Muhammad and Y.Z. Chen, KDBI: Kinetic Data of Bio-molecular Interactions database. *Nucleic Acids Res.*, 31(1):255-257, 2003.
- [17] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucl. Acids Res.*, 32: D449-D451, 2004.
- [18] C. Winter, A. Henschel, W.K. Kim and M. Schroeder. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, 34(Database issue):D310-4, 2006.
- [19] R.A. Laskowski, V.V. Chistyakov and J.M. Thornton. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, 33:D266-D268, 2005.
- [20] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28: 235-242, 2000.
- [21] J. Mintseris and Z. Wen. Structure, function, and evolution of transient and obligate protein-protein interactions. *PNAS*, 102: 10930-10935, 2005.
- [22] D. Subhajyoti, O. Krishnadev, N. Srinivasan and N. Rekha. Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Bioinformatics*, 5 :1-16, 2005.
- [23] H. Neuvirth, R. Raz and G. Schreiber. ProMate: A Structure Based Prediction Program to Identify the Location of Protein-Protein Binding Sites. *JMB*, 338:181-199, 2004.

- [24] J.R. Bradford and D.R. Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21: 1487-1494, 2005.
- [25] J. Mintseris and Z. Weng. Atomic contact vectors in protein-protein recognition. *Proteins*, 53: 629-639, 2003.
- [26] <http://www.ks.uiuc.edu/Research/vmd/>
- [27] <http://www.jdom.org>
- [28] F.M. Pearl, C.F. Bennett, J.E. Bray, A.P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton and C.A. Orengo. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, 31(1): 452-455, 2003.
- [29] A.G. Murzin, S. Brenner, T. Hubbard and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247, 536-540., 1995.
- [30] G. Schreiber. Kinetic studies of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 12:41-47, 2002.
- [31] <http://www.java.sun.com>
- [32] <http://tomcat.apache.org>
- [33] <http://httpd.apache.org>
- [34] <http://www.mysql.com>
- [35] C.J. Date. *Database in Depth*. O'Reilly, 2005.
- [36] <http://www.jfree.org>
- [37] <http://sourceforge.jmol.net>