ProDGe: investigating protein-protein interactions at the domain level

Finja Büchel[1,*], Clemens Wrzodek[1], Florian Mittag[1], Andreas Dräger[1], Adrian Schröder[1], and Andreas Zell[1]

[1]Departement of Cognitive Systems, Sand 1, 72076 Tübingen, Germany

August 4, 2011

Abstract

An important goal of systems biology is the identification and investigation of known and predicted protein-protein interactions to obtain more information about new cellular pathways and processes. Proteins interact via domains, thus it is important to know which domains a protein contains and which domains interact with each other. Here we present the Java™ program ProDGe (**Pro**tein **D**omain **Ge**ne), which visualizes existing and suggests novel domain-domain interactions and protein-protein interactions at the domain level. The comprehensive dataset behind ProDGe consists of protein, domain and interaction information for both layers, collected and combined appropriately from UniProt, Pfam, DOMINE and IntAct. Based on known domain interactions, ProDGe suggests novel protein interactions and assigns them to four confidence classes, depending on the reliability of the underlying domain interaction. Furthermore, ProDGe is able to identify potential homologous interaction partners in other species, which is particularly helpful when investigating poorly annotated species. We further evaluated and compared experimentally identified protein interactions from IntAct with domain interactions from DOMINE for six species and noticed that 31.13% of all IntAct protein interactions in all six species can be mapped to the actual interacting domains. ProDGe and a comprehensive documentation are freely available at http://www.cogsys.cs.uni-tuebingen.de/software/ProDGe.

Keywords: Protein, domain, interaction, domain-domain interaction, protein-protein interaction, interaction visualization

1 Introduction

Proteins are macromolecules that fulfill several important cell functions, such as enzymatic catalysis or transmission of information. Many of those tasks require the interplay of two or more proteins, which is called a protein-protein interaction. Protein-protein interactions occur in a wide range of contexts and play a central role in biological systems [10, 5, 9]. A protein-protein interaction is established by protein domains, which are encoded in protein sequences and form individual and independent structures. These domains and the resulting protein interactions are highly regulated and evolutionary conserved [12, 3]. One major topic of systems biology is the investigation and identification of known and predicted protein-protein interactions to reveal new cellular pathways, disturbed cell processes or even create complete interactomes of organisms [1, 5]. Especially the construction and comparison of protein interaction networks of different diseases can improve the understanding of the disease cause and can help to determine new drug targets [13].

A multitude of online databases, containing predicted or experimentally validated data, are available for the detailed investigation of proteins and their interaction partners. For instance, the UniProt database contains well curated protein information and cross-references to other databases [4]. The Pfam database provides domain and protein-domain information [7], and the DOMINE database combines and presents known and predicted domain-domain interactions by integrating both, experimental and predicted protein interaction data [14]. Here, we present the application ProDGe, which stands for **Pro**tein **D**omain **Ge**ne. ProDGe integrates protein, domain, gene, domain-domain interaction, and protein-protein interaction databases in one application. These interacting layers are presented in a neatly arranged view, so that users obtain an overview of a particular protein or domain and its interaction partners at the first glimpse. ProDGe bridges the gap between proteins,

*To whom correspondence should be addressed.
Email: finja.buechel@uni-tuebingen.de

their domains, domain interactions and protein interactions. Especially the combination of domain interaction databases with protein-domain annotation databases reveals novel protein-protein interactions and helps researchers of not well annotated organisms to find homologous interaction partners in other organisms.

2 Results and Discussion

2.1 Data integration

ProDGe integrates the data of four well curated online databases: UniProt, IntAct, Pfam and DOMINE. Protein sequences, identifiers, gene names and many other protein related information for millions of proteins are obtained from the UniProt knowledge database [4]. Experimentally validated protein interactions are collected by parsing the integrated IntAct database [2]. IntAct currently contains more than 250,000 binary interactions coming from curated literature mining or user submissions. Mapping all these data to the domain layer is done by taking protein and domain annotations from the Pfam database [7]. The Pfam database (release 25, March 2011) contains 12,273 protein families with mappings from proteins to domains and further domain information. We collect these domain information and map the data on the protein information from IntAct and UniProt. Thus, ProDGe can exactly mark and describe the domains in an amino acid sequence and can explain the IntAct protein interactions in more detail by displaying the actual interacting domains. To complete the domain layer, the Pfam data is extended with 26,219 domain interactions from the DOMINE database (v2.0, September 2010) [14]. DOMINE integrates domain interactions from 15 different sources that range from experimentally inferred domain interactions to predicted interactions. Each domain interaction is assigned one of four confidence classes: (i) experimentally validated, (ii) high prediction confidence, (iii) medium prediction confidence or (iv) low prediction confidence. This classification scheme is also used by ProDGe to characterize both, the domain and the protein interactions. The combination of the selected databases results in a comprehensive, high quality dataset for protein, domain, protein interaction, and domain interaction information. All datasets have been combined and stored in a MySQL database at our institution to maximize the performance and minimize memory requirements of the application. The combined database is updated whenever one of the integrated databases is updated.

2.2 Novel protein interactions

ProDGe suggests for each protein, containing one or more domains, a variety of new protein interactions. These suggestions are based on the idea that most domains and domain interactions are evolutionary conserved, and consequently, proteins will interact if they contain domains that are known to associate [12, 3]. Thus, the suggestion of novel protein interactions is performed by looking at all domains of a protein, using the Pfam database. The integrated DOMINE database is subsequently queried for interacting domains and their domain interaction partners. Mapping these interacting target domains back to a list of proteins that contain them results in a list of potential protein interaction partners. These potential protein interactions are finally assigned to the same interaction confidence class as the domain interaction. That means, if the domain interaction is experimentally validated, predicted with high, medium or low interaction confidence, then the predicted protein interaction belongs to the same class. If there are more domain interactions describing a protein interaction with different interaction classes, the best interaction class is assigned to the protein interaction. ProDGe supports a multitude of different species. By default, only interactions occurring in the same species are considered. However, this restriction can be lifted to reveal homologous protein interactions in related species. This helps to deduce unknown protein interactions in poorly annotated species, based on interacting proteins of other species that contain the same domains, and helps to complete missing links when findings from one species should be transferred to another. Since ProDGe combines both, protein and interaction data from IntAct and DOMINE [2, 14], we compared the overlap between these sources for six species: (i)*Homo sapiens*, (ii)*Mus musculus*, (iii) *Rattus norvegicus*, (iv) *Saccharomyces cerevisiae*, (v) *Caenorhabditis elegans*, and (vi) *Drosophila melanogaster*. The evaluation for each species has been performed in the following way:

Step 1: Mapping IntAct interactions on domains

First, all experimentally validated IntAct protein interactions are separated into interactions with and without domain information, and those with domain information that

Table 1: Comparison of IntAct and DOMINE interactions

Species (no of all experimental PPIs from IntAct)		PPIs without annotated domains	PPIs without interactions in DOMINE	PPIs mapped to DDIs				
				Total	PDB	HC	MC	LC
D. melanogaster	2486	545 (21.92%)	987 (39.70%)	954 (38.37%)	4.14%	18.06%	4.14%	12.03%
C. elegans	818	281 (34.35%)	184 (22.49%)	353 (43.15%)	5.38%	18.83%	4.52%	14.43%
S. cervisiae	59257	16534 (27.90%)	29029 (48.99%)	13694 (23.11%)	1.10%	7.40%	5.23%	9.38%
R. norvegicus	1022	106 (10.37%)	706 (69.08%)	210 (20.55%)	3.23%	7.63%	2.25%	7.44%
M. musculus	3402	347 (10.20%)	1916 (56.32%)	1139 (33.48%)	5.38%	17.70%	3.67%	6.73%
H. sapiens	29337	5510 (18.78%)	15581 (53.11%)	8246 (28.11%)	3.48%	14.07%	3.31%	7.24%
Average	16053.67	3887.17 (20.59%)	8067.17 (48.28%)	4099.33 (31.13%)	3.79%	13.95%	3.86%	9.54%

ProDGe contains both, the protein-protein interactions (PPIs) from IntAct and domain interactions from DOMINE. To compare both data sources and evaluate the overlap between them, the experimental IntAct protein interactions have been mapped the corresponding domain interactions and compared to the DOMINE content. PDB denotes protein interactions containing domains whose interaction is experimentally validated. HC, MC and LC denotes protein interactions whose domain interactions have a high, medium or low prediction confidence.

have been mapped to the domains, using the integrated Pfam database.

Step 2: Comparing with DOMINE interactions In a second step, we searched matching domain interactions in the DOMINE dataset. Hereby, a protein interaction could possibly be explained by several domain interactions. In that case, the domain interactions with the highest confidence class is chosen to explain the protein interaction.

The results are shown in Table 1. In the first column, the species and total number of experimentally validated protein interactions for the species are listed. These interactions are separated into: (i) interactions that have no domain description, (ii) interactions for which no corresponding domain interaction could be found, (iii) number of protein interactions from IntAct that have corresponding domain interactions in DOMINE. The last column is further separated into experimentally validated protein interactions that can be mapped to domain interactions of confidence class experimentally validated (PDB), high predictive confidence (HC), medium predictive confidence (MC), and low predictive confidence (LC). The table reflects the advantage of integrating protein and domain interaction data. This integration allows to predict new protein interactions with high confidence and to explain and further characterize known protein interactions. On the one hand, 31.13% IntAct protein interactions of all species in our evaluation can be mapped to domain interactions and thus explained in more detail. On the other hand, the overlap is not too big and hence, the combination of both datasets results in a more complete picture than taking just one interaction dataset. For example, 76.89% experimentally validated protein interactions in *S. cervisiae* are not overlapping with domain interactions from DOMINE. And 60.36% (702 in absolute numbers) of all experimentally validated DOMINE interactions are not overlapping with protein interactions from IntAct (data not shown). Thus, the combination of both databases leads to a high quality network of protein interactions on different layers. On the application side, ProDGe tracks and displays sources and confidences for each domain or protein interaction. This includes a clear separation and visualization of experimentally validated interactions and predicted interactions as well as displaying the source for each interaction. The number of protein interactions, coming from the combination of DOMINE and Pfam, can even be increased by including predicted domain interactions. In *S. cervisiae*, the number of domain interactions including predictions increases from 702 to 7,444, whereby 52.29% of the domain interactions can now be mapped to protein interactions (data not shown). Including predicted data and especially having a large amount of predicted versus experimentally verified data is a very common situation in proteomics. For example, the UniProt database contains millions of proteins. But experimental evidence for the existence of proteins is only available for 3.46% of all proteins in the database. 20% are inferred from homology and the existence of 76.53% of all proteins relies just on predictions [6]. Hence, the number of protein interactions, inferred from domain interactions, is much bigger than the number of experimentally validated protein interactions in IntAct, because no protein

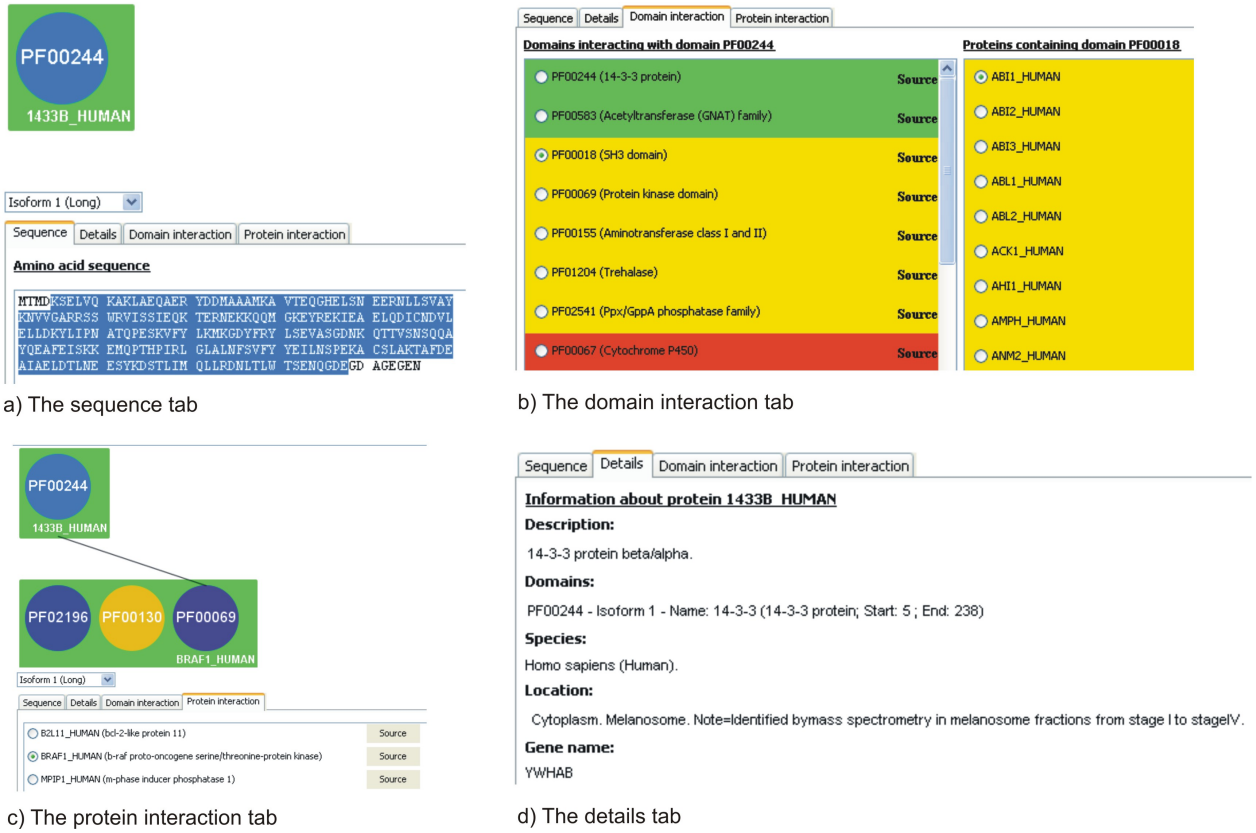

Figure 1: Some functionalities of ProDGe. Subfigure (a) visualizes the long isoform of protein `1433B_HUMAN` and its amino acid sequence. The protein is depicted as a green rectangle and the domain as colored circle. The domain part of the amino acid sequence is highlighted in the same color as the corresponding domain circle. (b) The domain interaction tab: On the left is a domain list, providing all domains interacting with the selected domain `PF00244`. The protein list on the right shows all proteins containing the selected domain `PF00018` from the left list. The different colors represent the interaction confidence classes (green=experimental validated interaction, yellow=high prediction confidence and red=low prediction confidence). A legend is also shown on this tab, providing the possibility to show or hide some interactions and further descriptions (not shown in the picture). (c) Visualization of two interacting proteins (interacting domains are connected with a line) and the protein interaction tab, providing known protein interactions. (d) Example of the details tab, providing various protein and domain related information.

interaction for a protein whose existence is not even proven can be measured experimentally.

2.3 The application

Usage

To provide direct access to information about proteins, domains and their interactions, ProDGe can handle UniProt identifiers, UniProt accession numbers, Pfam identifiers, Ensembl identifiers, and gene symbols. After entering one of theses identifiers, the corresponding protein or domain is depicted in an internal window. Now the following information is provided: (i) visualization of the protein as a rectangle with domains as circles, (ii) a list of the protein isoforms, (iii) the amino acid sequence of the protein, whereas the domain parts are colored differently, (iv) protein and/or domain information like description, cellular location and several identifiers, (v) predicted and experimentally validated domain interaction partners for a selected domain and the proteins containing the interacting domain, and (vi) experimentally validated protein interaction partners (see Figure 1). Furthermore, it is possible to export the visualized information in a PDF document and to obtain additional information of the genetic context of a protein from Ensembl [8]. ProDGe distin-

guishes between known protein interactions, which are collected from the cross-references provided by the UniProt database and IntAct, and predicted protein interactions based on the combination of DOMINE and Pfam data (see section 2.2). For instance, if a domain has one interaction partner, which is experimentally validated, then ProDGe searches for other proteins containing the same domain and suggests a novel protein interaction. The combination of different interaction layers can identify currently unknown protein interactions, which play an important role in cellular processes. Furthermore, it is possible to look for interactions between different species, which is particularly helpful for the investigation of organisms whose protein interactions are rare or currently unknown.

Availability and requirements

ProDGe and a comprehensive documentation are freely available from http://www.cogsys.cs.uni-tuebingen.de/software/ ProDGe. There are two possibilities to start ProDGe: running it directly as a Java™ webstart application or by downloading the application as ZIP archive. The ZIP file contains an executable JAR and scripts for various operating systems to start the JAR so that no further installation of any library other than the Java™ virtual machine is required. In any way, an active internet connection and a Java™ virtual machine (version 6 or later) are required.

Implementation

ProDGe is entirely written in Java™ and runs on all operating systems for which a Java™ virtual machine is available. For obtaining the latest protein information from UniProt on-the-fly, the WSDBfetch library from EMBL-EBI has been integrated into the application [11]. The integrated dataset has been stored in a MySQL database that is automatically queried by ProDGe. This database is updated regularly and located on a server of the chair for Cognitive Systems at the University of Tübingen.

3 Conclusion

We have integrated databases containing information about proteins (UniProt), domains (Pfam), protein interactions (IntAct), and domain interactions (DOMINE). These information help to obtain a complete picture of proteins and their interactions. Furthermore, the integrated dataset allows for suggesting novel protein interactions by taking domain interactions and mapping the corresponding domains back to proteins that contain these domains. This procedure can also be used to identify homologous interaction partners in other species and thus, is particularly useful for investigations of not well annotated species. The platform-independent tool ProDGe visualizes both, proteins and domains, with their experimentally validated and predicted interaction partners. ProDGe uses a simple interaction classification to carefully distinct between experimental and predicted interactions with different confidences and displays the source for each interaction. This classification scheme, paired with a careful layout of the application, guarantees an easy usage. The application is freely available and can be downloaded as stand-alone version or executed directly as Java™ webstart.

Acknowledgements

We thank Rainer Nagler for creating the ProDGe logo.
Funding: German Federal Ministry of Education and Research (BMBF) [National Genome Research Network (NGFN+) under grant number 01GS08134].

References

[1] P. Aloy and R. B. Russell. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, 7(3):188–197, Mar 2006.

[2] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, 38(Database issue):D525–D531, Jan 2010.

[3] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 13(1):190–202, Jan 2004.

[4] U. Consortium. Ongoing and future developments at the universal protein resource.

Nucleic Acids Res, 39(Database issue):D214–D219, Jan 2011.

[5] M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill. Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No. 2:R171–R181, Oct 2005.

[6] European Bioinformatics Institute. UniProtKB/TrEMBL protein database release 2011_05 statistics. http://www.ebi.ac.uk/uniprot/TrEMBLstats/, 2011.

[7] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–D222, Jan 2010.

[8] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S. Riat, D. Rios, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. e M Fernández-Suarez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Vogel, and S. M. J. Searle. Ensembl 2011. *Nucleic Acids Res*, 39(Database issue):D800–D806, Jan 2011.

[9] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, Jan 1996.

[10] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–14121, Oct 2002.

[11] H. Mcwilliam, F. Valentin, M. Goujon, W. Li, M. Narayanasamy, J. Martin, T. Miyar, and R. Lopez. Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Research*, 37:W6–W10, 2009.

[12] J. Park and D. Bolser. Conservation of protein interaction network in evolution. *Genome Inform*, 12:135–140, 2001.

[13] L. Sam, Y. Liu, J. Li, C. Friedman, and Y. A. Lussier. Discovery of protein interaction networks shared by diseases. *Pacific Symposium on Biocomputing*, 12:76–87, 2007.

[14] S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari, and R. Jothi. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res*, 39(Database issue):D730–D735, Jan 2011.