Characterisation of model species interactome available from primary molecular interaction databases

Vitalii Kleshchevnikov

21 December 2016

Contents

The report was published on 2017-05-13. Data used in the report was available on 2017-04-24.

Chapter 1

Outline

1. Abstract
2. Introduction
3. Methods
4. Results and discussion
    - how available interactome covers the proteome
    - IntAct(IMEx) vs Biogrid
    - which proteins are missing
    - interaction detection biases
    - study bias: the more articles, the more interactions
    - which protein function are interaction databases and high-throughput datasets enriched in?
5. Conclusion

Chapter 2

Abstract

Chapter 3

Introduction

The structure and the function of the cell arise from interactions between molecules inside and outside it [Hein et al., 2015]. Though proteins, nucleic acids, lipids and small molecules can all form important interactions, studies and literature focus mainly on interactions between proteins and other macromolecules. We can discover and study these molecular interactions using a number of experimental and computational techniques. This study aims to describe the coverage and the biases of currently available molecular interaction data. We focus on molecular interactions identified in the experimental setting most of which are represented (in the literature and databases) by protein-protein interactions (although, there is a considerable amount of data on protein-DNA interactions, for example, ChIP-Seq data, which is traditionally incorporated into the genomic or specialist databases). We focus on specifically on protein-protein interaction data.

Molecular interactions can be classified using multiple criteria: the information interaction detection methods can provide towards ground truth, interaction detection method, biological role of the interaction (covalent binding, enzymatic reaction, e.g.) which can be explored in molecular interaction ontology [Hermjakob et al., 2004, http://www.ebi.ac.uk/ols/ontologies/mi/]. A standard way of describing interaction allows record published interactions into databases, assign interactions a score based on the evidence and reliability and, least but not last, reuse interaction data for computational analyses, gaining insight into the novel fucntion of proteins and generate hypothesis. In the recent years, the fact that biases in molecular interaction data can mislead network-driven studies has become evident [Schaefer et al., 2015], which motivates the need to study the coverage and bias of currectly available molecular interaction data, identify missing proteins and molecular functions those protiens perform. These results may help in selecting appropriate network data for data integration studies.

We investigate the data in the IntAct database [Orchard et al., 2014] which is provider of high-depth manually curated molecular interaction data, a part of IMEx consortium. The IMEx consortium [Orchard et al., 2012] is an international collaboration between a group of major public interaction data providers who share curation effort. This study can aid literature curation effort by IntAct group by pointing to the publications contating interaction

data on proteins with no interactions currently deposited in the IntAct database.

interac we can classify interactions into different types: binary interactions and associations. Binary interactions are the interactions between two components, for example, two specific proteins, some detection methods (e.g. two-hybrid) identify those. To understand associations, we need to imagine we know proteins A, B and C constitute a complex and interact as shown in a figure 1 A. When we conduct an experiment, we choose the bait (the molecule experimentally treated to capture its interacting partners - called preys) to be protein A, and by detection method (e.g. affinity-purification mass spectrometry) we get both protein B and protein C detected as preys. Next step is to translate bait-prey relationship into a model of reality like the one shown in the figure 1 A. We call interactions between A-B and A-C associations because we cannot infer the true relationship between A, B, and C from this experiment design. In the other words, establishing that proteins are in direct physical contact is really challenging. However, to represent associations in a tabular format with each row corresponding to one interaction (e.g. A-B) we need to expand those. Two ways are commonly used to expand interactions, matrix and spoke expansion, both shown in the figure 1 B.

Defining interactome

The aggregation of all components and their interactions into a single network results in what we call interactome, the whole of all molecular interactions. You can also look into the subset of this network, for example, you can select only those proteins that are expressed in the brain, and only the interactions between these proteins identified experimentally in the brain cells. This example reflects the complexity and the diversity of the interactome - which is what you would expect from a system underlying the complexity and the diversity of the cell types, cellular behaviours, and functions. For the same reason, only by studying these interactions and how they change in specific cell types and under specific circumstances in combination with the functional analysis we can decipher cellular regulatory networks. The ultimate goal of the research in the field would be to capture all physical interactions and thoroughly describe them while avoiding false positive results.

Experimental approaches for discovering interactions

Numerous experimental protein interaction detection methods are currently widely used. The classification we provide is far from comprehensive but gives a short description of the methods analysed in this study. Based on the evidence provided and possibility to scale up to high-throughput studies methods can be classified into 3 main categories.
The first category is formed by methods using affinity purification to capture all proteins associated with the bait. Only proteins that have a direct or indirect physical connection with the bait will be purified. Following purification procedure, those proteins can be identified using western-blotting and specific antibody staining or using mass-spectrometry, latter can be done in a high-throughput manner. The main advantage of these methods is the potential

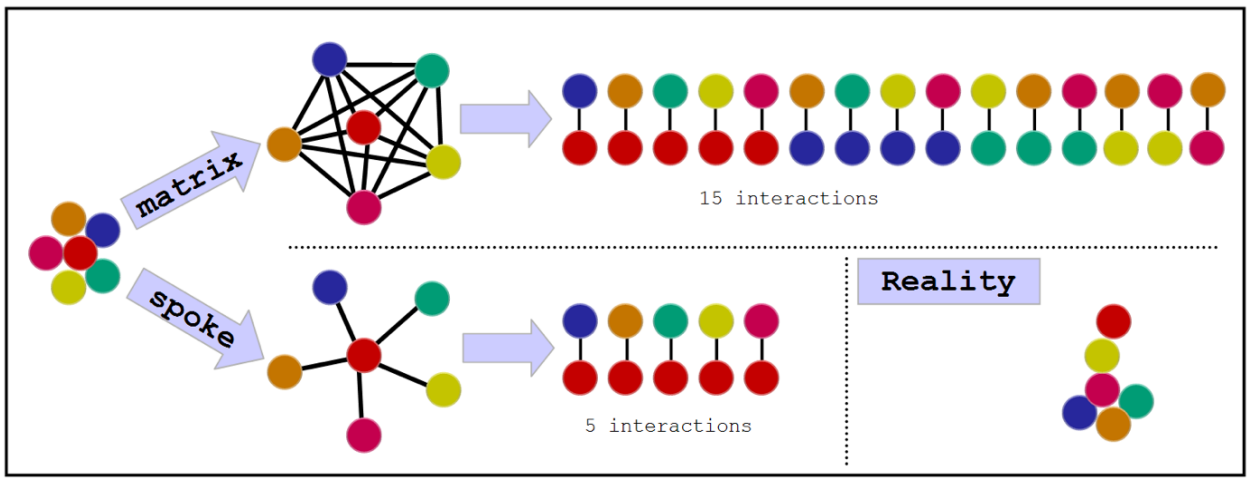

Figure 3.1: Figure 1-B. Association and physical associations identified using AP-MS approaches fundamentally don't provide binary interactions. Binary interactions have to be inferred from the list of proteins represented as cluster on the left. Matrix expansion links every protein to every other protein in the AP-MS-derived cluster. Spoke expansion only links the bait with all other proteins in a cluster. As you can see, none of these methods generate the exact set of binary interactions that occur in reality. Spoke expansion tends to generate less false-positives and is used more often.

ability capture any protein, in different cellular contexts and to quantitatively characterise interaction properties [Hein et al., 2015]. The main disadvantage of these techniques comes from the fact that experiments indentify both direct and indirect interactions between the bait and captured proteins and no way of distinguishing those (although, one may delete identified proteins one by one from the cell and decipher direct interactions). This type of interactions in called associations and do be represented in the network requires the use of expansion methods (Figure 1-B, adopted from IntAct website). The other drawback of this method is dependency on the availability and quaility of antibodies for affinity-purification step. We will call these methods affinity-purification followed mass-spectrometry (AP-MS) across this report.
As defined by molecular interaction ontology [Hermjakob et al., 2004, http://www.ebi.ac.uk/ols/ontologies/mi/], an association is an interaction between molecules that may participate in formation of one, but possibly more, physical complexes, association will be called physical if experiments show enough evidence that proteins are in the same physical complex but don't show direct interaction. The second category of methods is formed by protein complementation techniques which include two-hybrid (transcription factor complementation), the most widely used interaction detection method (including high-throughput experiments). In this method, pairs of proteins are tested for interaction and therefore discovered interactions are more likely to be direct (the main advantage of this method). Classic implementation of two-hybrid is performed in yeast cells and requires studied proteins to be non-membrane, however, two-hybrid for membrane proteins and for mammalian cells was also developed [Lemmens et al., 2015, Saraon et al. [2017]]. The main disadvantage of two-hybrid methods is that every protein has to be cloned into a plasmid or other vector and exogenously expressed.

Ability to clone protein-coding sequence and, in case of yeast two-hybrid, correct protein folding, are the limiting factors for two-hybrid but not AP-MS techniques. As a side note, the lack of antibodies for a specific bait protein may force researches to tag, clone and express protein which is a subject to similar problems, however, AP-MS would allow identification of the binding partners, no cloning needed. Final category is formed by structure-based methods (co-crystallisation and X-ray crystallography, e.g.). These methods can provide valuable information on how exactly physical interaction occurs, however, these methods are extremely labor-intensive and, therefore, non-scalable and will always need complementary experiments showing if the proteins actually interact in the cellular context. To conclude, all protein interaction detection methods have their strengts and weaknesses, so it is important to accept that every protein interaction detection method has it's limitations in the ability to identify true physical interaction and serves as evidence we use to infer protein interaction. By combining different methods to identify protein interaction we can gain more confidence in our findings and by disrupting interaction under specific cellular context we can identify it's function.

Challenges of interactomics

Four big challenges substantially complicate the study of molecular interactions, especially on the whole organism scale. The first being that we don't know the true nature of underlying our experimental results (all assays provide evidence that interaction is possible and some can provide quantitative description, but all are prone to error and the problem described in the figure 1 A) which lead to the necessity of combining interaction data from multiple experiments and complex statistical evaluation of how probable the interaction is based on that data (such as Bayesian approach, Braun et al. [2009], Zhang et al. [2011]) rather than receiving confident yes-or-no result from a single experiment. Interaction databases make an effort to score the interactions based on supporting evidence, however, this is usually done with non-probabilistic heuristic approaches, such as MI score implemented in IntAct [Villaveces et al., 2015]. Every database that aggregates interaction data from other resources will develop an algorithm to score interactions. The challenge is to identify when to put the threshold betweem high and low confidence interactions or when to say "I am confident the interaction exists".

The second big challenge is the problem of "noise" - or the problem of false positives. Different interaction detection experiments are prone to these errors for different reasons, for example, in-vitro experiments (e.g. TAP-MS) may allow the interaction between proteins which are normally separated between different cellular compartments. Specific groups of proteins (based on their physical or chemical properties, abundance) may have a higher susceptibility to false positives, for example, highly abundant proteins are easier to detect and may also be less efficiently diluted during the affinity purification procedure, which may lead to artifactual results. Contamination is another common problem for AP-MS experiments which have motivated the creation of contaminant database, CRAPOME [Mellacheruvu et al., 2013]. A more general problem of noise can be addressed by proteome-scale interactomics experiments (which can include enough samples to guarantee low false positive rate while still identifying

interactions).

The third big challenge is that our knowledge of interactome is incomplete which arises from the fact that experimental approaches have low statistical power and often miss out some real interactions. Many efforts to reproduce protein interactions find little overlap between the new and the original study [Braun et al., 2009]. Also, many proteins, especially in non-model species have no know interactions.

The final challenge contributes to the "incomplete interactome" problem but is grounded in the fact that not all protein interaction discovered and published are included in protein interaction databases. In the other words, this is database curation problem. More than 100 public databases containing protein interactions are available now [http://pathguide.org/]. These databases differ: - by the types of data they include (e.g. computational prediction, manual curation from experimental articles - primary, aggregated data from many primary databases - secondary),
- the level of detail captured from articles to describe interactions,
- how often and if they are updated with new data.
The level of detail ranges from only mentioning the pairs of interactors and heuristic score assigned to them (STRING, updated once in 2 years) to the ones containing experiment details (detection method, bait/prey status, if available - quantitative data, experiment setup, protein variants), such as IntAct [PMCID: PMC3703241]. The amount of interaction data generated per year is growing exponentially making manual curation of all this data into primary databases a daunting task. To prioritise curation efforts and reduce redundancy between databases (to curate different data using the same standards) IMEx consortium was formed in 2012 [PMCID: PMC3703241]. IMEx-compliant databases include all big primary databases excluding only BioGRID (which curates at the lower level of detail) and not active legacy databases.

Motivation for this study

Solving some of these challenges may be easier than the others. In particular, to solve the last challenge we can prioritise curation efforts for already published interactions to cover unrepresented proteins and we can encourage authors to submit their results to the databases prior to publishing. We can also encourage research of underrepresented parts of the interactome. However, for both of those aims, we need to characterise the interactome already present in interaction databases. Specifically, to learn how available interactome covers the proteome of main model species, if there are any biases to proteins with no available interactions and if any major protein interaction detection methods exhibit any biases towards specific groups of proteins. The other helpful to look at the problem is to search for underrepresented in interaction databases but in general well-researched proteins.

Chapter 4

Aims of the study

1. Find out how available interactome covers the proteome of main model species. Considering either all UniProtKB or SwissProt entries only as the proteome (canonical identifiers as well as protein isoforms). Consider all interactions from IMEx-compliant databases as interactome.
2. Compare the coverage of proteome by interactome from IMEx to the interactome from BioGRID (the other major primary database).
3. Find out if proteins with no available interactions stand out by specific functions (Gene Ontology, GO: biological process and molecular function), cellular localisation (GO), molecular mass, or protein evidence status from SwissProt
4. Find out if major protein interaction detection methods (two-hybrid and AP-MS, AP-WB?) exhibit any bias towards biochemical properties of the proteins involved (mass, disordered regions, hydropathy, the fraction of charged residues)
5. What is the relationship between the number of interactions or MI score and the number of publications or GO terms per protein?
6. Are proteins with a higher fraction of intrinsically disordered domains more likely to have interactions available and do they have more interactions (if normalised for how well-studied proteins are)?
7. Find out if there are any proteins which are in general well researched (many associated publications or manual GO annotations) but underrepresented in IntAct (low MI score)
8. If that is possible to measure: do intermediate filaments (or other highly insoluble proteins) really have higher rates of false-discovered interactions?

Chapter 5

Methods - data processing and analysis

Getting proteome from UniProtKB

Whole proteome (all UniProtKB) for each species was downloaded programmatically in R using UniProt rest API. SwissProt-proteome was subset from the whole proteome by reviewed status column. UniProt identifies proteins by UniProtKB/AC (e.g. P04637, accession) which does not distinguish between protein isoforms. UniProt aggregates isoform information and identifiers (e.g. P04637-4) in a separate column with zero to many isoforms per each UniProtKB accession. To generate proteome list which includes protein isoforms, isoform accessions were extracted and combined with the list of generic accessions. In this analysis, protein evidence status and protein mass are only attributed to generic accessions.

Getting and transforming interactome data from IMEx databases and BioGRID

Interactome from all IMEx databases was downloaded programmatically in R using PSIQUIC package from Bioconductor [Shannon, 2017]. IMEx databases include IntAct, MINT, bhf-ucl, MPIDB, MatrixDB, HPIDb, I2D-IMEx, InnateDB-IMEx, MolCon, UniProt, MBInfo. The list of interactions (pairs of interactors) was transformed into the list of interactors preserving interactor identifiers, the type of interactor identifier, species information and the database interaction originates from. Only unique proteins wereIMEx databases contain interactions between proteins, RNA, DNA and small molecules, moreover, these interaction may involve molecules originating from different species. Therefore, to perform by species interactome/proteome comparison there is a need to remove non-UniProtKB/AC molecule identifiers (which removes non-protein molecules, although, may also remove a small fraction of proteins which have no UniProtKB/AC) and there is a need to remove proteins originating from other species. Also, entries in IMEx databases has to be cleaned of tags and textual descriptions ("taxid:9606(human-h1299)|taxid:9606(Homo sapiens lung lymph node

carcinoma)" to "9606") to make further analysis easier and cleaner. Next, when provided in the research articles protein isoform information is always included in IMEx databases, so to perform analysis excluding isoform information UniProtKB/AC were cleaned of -N suffix (P04637-4 to P04637). IMEx consortium databases such as IntAct, MINT, BHF-UCL, MPIDB, MatrixDB, HPIDb, I2D-IMEx, InnateDB-IMEx, MolCon, UniProt, MBInfo are currently integrated into IntAct [1].

Physicochemical properties and disordered region prediction

Information on disordered region content and biochemical properties of individual proteins were obtained from the dataset generated by Vincent and Schnell in 2015 []. Briefly, Vincent and Schnell used a number of disorder prediction algorithms (IUPred and DisEMBL) and their consensus to generate disordered region predictions for each protein which was be used to calculate the fraction of disordered regions in a protein. In addition, Vincent and Schnell used localCIDER version 0.1.7 (Classification of Intrinsically Disordered Ensemble Regions) to calculate physical properties such as a fraction of charged residues, mean hydrophobicity or charge separation for each protein. This was done for 10 eukaryotic proteomes and written to SQLite-database which was made available online.

Gene ontology enrichment analysis

coming soon

Bibliography

Pascal Braun, Murat Tasan, Matija Dreze, Miriam Barrios-Rodiles, Irma Lemmens, Haiyuan Yu, Julie M Sahalie, Ryan R Murray, Luba Roncari, Anne-Sophie de Smet, Kavitha Venkatesan, Jean-Francois Rual, Jean Vandenhaute, Michael E Cusick, Tony Pawson, David E Hill, Jan Tavernier, Jeffrey L Wrana, Frederick P Roth, and Marc Vidal. An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, 6(1):91–97, Jan 2009. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.1281.

Marco Y Hein, Nina C Hubner, Ina Poser, Jurgen Cox, Nagarjuna Nagaraj, Yusuke Toyoda, Igor A Gak, Ina Weisswange, Jorg Mansfeld, Frank Buchholz, Anthony A Hyman, and Matthias Mann. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3):712–723, Oct 2015. ISSN 1097-4172 (Electronic); 0092-8674 (Linking). doi: 10.1016/j.cell.2015.09.053.

Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jerome Wojcik, Lukasz Salwinski, Arnaud Ceol, Susan Moore, Sandra Orchard, Ugis Sarkans, Christian von Mering, Bernd Roechert, Sylvain Poux, Eva Jung, Henning Mersch, Paul Kersey, Michael Lappe, Yixue Li, Rong Zeng, Debashis Rana, Macha Nikolski, Holger Husi, Christine Brun, K Shanker, Seth G N Grant, Chris Sander, Peer Bork, Weimin Zhu, Akhilesh Pandey, Alvis Brazma, Bernard Jacq, Marc Vidal, David Sherman, Pierre Legrain, Gianni Cesareni, Ioannis Xenarios, David Eisenberg, Boris Steipe, Chris Hogue, and Rolf Apweiler. The hupo psi's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–183, Feb 2004. ISSN 1087-0156 (Print); 1087-0156 (Linking). doi: 10.1038/nbt926.

Irma Lemmens, Sam Lievens, and Jan Tavernier. Mappit, a mammalian two-hybrid method for in-cell detection of protein-protein interactions. *Methods Mol Biol*, 1278:447–455, 2015. ISSN 1940-6029 (Electronic); 1064-3745 (Linking). doi: 10.1007/978-1-4939-2425-7{_}29.

Dattatreya Mellacheruvu, Zachary Wright, Amber L Couzens, Jean-Philippe Lambert, Nicole A St-Denis, Tuo Li, Yana V Miteva, Simon Hauri, Mihaela E Sardiu, Teck Yew Low, Vincentius A Halim, Richard D Bagshaw, Nina C Hubner, Abdallah Al-Hakim, Annie Bouchard, Denis Faubert, Damian Fermin, Wade H Dunham, Marilyn Goudreault, Zhen-Yuan Lin, Beatriz Gonzalez Badillo, Tony Pawson, Daniel Durocher, Benoit Coulombe, Ruedi Aebersold, Giulio Superti-Furga, Jacques Colinge, Albert J R Heck, Hyungwon Choi, Matthias Gstaiger, Shabaz Mohammed, Ileana M Cristea, Keiryn L Bennett, Mike P

Washburn, Brian Raught, Rob M Ewing, Anne-Claude Gingras, and Alexey I Nesvizhskii. The crapome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods*, 10(8):730–736, Aug 2013. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.2557.

Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti, Fiona S L Brinkman, Gianni Cesareni, Andrew Chatr-aryamontri, Emilie Chautard, Carol Chen, Marine Dumousseau, Johannes Goll, Robert E W Hancock, Linda I Hannick, Igor Jurisica, Jyoti Khadake, David J Lynn, Usha Mahadevan, Livia Perfetto, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Lukasz Salwinski, Volker Stumpflen, Mike Tyers, Peter Uetz, Ioannis Xenarios, and Henning Hermjakob. Protein interaction data curation: the international molecular exchange (imex) consortium. *Nat Methods*, 9(4):345–350, Apr 2012. ISSN 1548-7105 (Electronic); 1548-7091 (Linking). doi: 10.1038/nmeth.1931.

Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C Lovering, Birgit Meldal, Anna N Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The mintact project–intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, 42(Database issue):D358–63, Jan 2014. ISSN 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/gkt1115.

Punit Saraon, Ingrid Grozavu, Sang Hyun Lim, Jamie Snider, Zhong Yao, and Igor Stagljar. Detecting membrane protein-protein interactions using the mammalian membrane two-hybrid (mamth) assay. *Curr Protoc Chem Biol*, 9(1):38–54, Mar 2017. ISSN 2160-4762 (Electronic); 2160-4762 (Linking). doi: 10.1002/cpch.15.

Martin H Schaefer, Luis Serrano, and Miguel A Andrade-Navarro. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet*, 6:260, 2015. ISSN 1664-8021 (Linking). doi: 10.3389/fgene.2015.00260.

Paul Shannon. *PSICQUIC: Proteomics Standard Initiative Common QUery InterfaCe*, 2017. R package version 1.14.0.

J M Villaveces, R C Jimenez, P Porras, N Del-Toro, M Duesbury, M Dumousseau, S Orchard, H Choi, P Ping, N C Zong, M Askenazi, B H Habermann, and Henning Hermjakob. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database (Oxford)*, 2015, 2015. ISSN 1758-0463 (Electronic); 1758-0463 (Linking). doi: 10.1093/database/bau131.

Wangshu Zhang, Fengzhu Sun, and Rui Jiang. Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach. *BMC Bioinformatics*,

12 Suppl 1:S11, Feb 2011. ISSN 1471-2105 (Electronic); 1471-2105 (Linking). doi: 10.1186/1471-2105-12-S1-S11.