Characterisation of model species interactome available from primary molecular interaction databases

Vitalii Kleshchevnikov

21 December 2016

Contents

The report was published on 2017-05-12. 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. 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 focuses on molecular interactions identified in the experimental setting, most of which are represented in the literature and databases by protein-protein interactions (also protein-DNA interactions obtained, for example, by ChIP-Seq, but those are traditionally incorporated into genomic databases).

Based on the information detection methods can provide towards ground truth we can classify interactions in three 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, hub and spoke expansion, both shown in the figure 1 B.

Defining interactome

The aggregation of all components and their interactions into a single network result 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 proteins, only those proteins that are expressed in the brain, and only the interactions between this protein 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 discoveries.

Experimental approaches for discovering interactions

Experimental protein interaction detection methods can be classified into 3 main categories based on the evidence they provide and whether they can be used in a high-throughput manner: The first category is formed by methods using affinity purification of the bait and all the prey associated with it. Following that, preys can be identified using western-blotting and specific antibodies or using mass-spectrometry, which can be done in a high-throughput manner [Mann,]. The main advantage of these methods is the ability to quantitatively characterise interactions [Mann,] and capture many prey proteins per bait - the latter, however, presents the disadvantage of dealing with associations. The main disadvantage of these techniques is that for the reliable result it requires all interacting proteins to be soluble []. The second category 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 all discovered interactions are binary (the main advantage of this method). Classic implementation of two-hybrid requires proteins to be soluble as well [], however, two-hybrid for membrane proteins was also developed []. The main disadvantage of two-hybrid methods are that they allow only qualitative characterisation of interactions [], are usually performed in yeast (thus, have a lower sensitivity) and are highly prone to false-positive results []. Final category consists of methods based on the structure of the protein complex. They can provide valuable information on how exactly physical interaction occurs but as for now are extremely labor-intensive and will always need complementary experiments showing if the proteins actually interact in the cellular context.

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 (Bayesian approach [1]) rather than receiving confident yes-or-no result from 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, like MI score [PMCID: PMC4316181].

The second big challenge is the problem of "noise" - or 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 included in separate cellular compartments. Specific groups of proteins (based on their physical or chemical properties) may have a higher susceptibility to false positives, for example, intermediate filaments (e.g. nuclear lamins) have low solubility under non-denaturing conditions necessary for affinity-purification based techniques, which may lead to artifactual results. However plausible, this particular problem lacks empirical evidence and requires more investigation. A more general problem of noise will be addressed by more 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 and arises from the fact that experimental approaches have low statistical power and often miss out some real interactions. Also, many proteins, especially for non-popular model species, were not researched for protein 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. 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).

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

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