Combining Interactomes from Multiple Organisms: a Case Study on Human-Mouse

Juan J. Cáceres and Alberto Paccanaro
Department of Computer Science - Centre for Systems and Synthetic Biology
Royal Holloway, University of London
Egham, Surrey, UK

Abstract—**The amount and quality of available data on different organisms varies greatly. While model organisms benefit from extensive experimental studies, there is often a lack of detailed experimental data for more specific organisms. Additionally, even among model organisms there are noticeable differences in the amount and type of data available, due to the different suitability of experiments in different organisms. The combination of interactomes for closely related species, represents a viable tool to increase the amount of protein- protein interaction data for a given organism. The Human-Mouse case of study is particularly relevant, as many experiments cannot be carried out on humans. This paper describes a general method to construct a combined interactome from different organisms. The construction is achieved through the integration of data from different sources and formats, including gene-protein relations, protein homology classes and protein-protein interactions. We show that the Human-Mouse combined interactome increase the mouse gene coverage by over 150% and the interaction coverage by over 430%. We also provide a novel mathematical formalisation for the interactome combination.**

I. Introduction

In modern biology, organism functions are modelled biological networks. Different biological networks are suited to describe the different layers in witch biological components act. The interactome of an organism is the set of all molecular interactions that can occur in this organism.

A common issue in large scale analysis of biological networks is the lack of available data. Although the genome for many organisms [2] is available, the function and interactions of their proteins remain, in many cases, unknown. From a Computer Science perspective, an interactome is an undirected graph, where the vertices represent proteins, and the edges represent an interaction between the connected proteins. We propose the construction of a combined interactome by integrating publicly available data into a wider-covering network of protein interactions.

A common approach when studying human diseases is the use of model organisms, such as *M. musculus*, *C. elegans* and *D. melanogaster*. These model organisms are comprehensively studied and biologists conduct experiments that would be impossible in human beings. While the model organisms allow experiments in controlled environments, the work model results in a fragmented environment: certain model organisms are used for specific diseases. For example, the *M. musculus* or house mouse, is extensively used for Cancer research and Diabetes [3], while the nematode worm *C. elegans* is used for the early stages of brain development [4].

To analyse this diseases, researchers would mutate individuals of their population until groups of susceptible and resistant individuals appear. The distinct groups present defining DNA traits that are related to the resistance to the disease. An experiment would, then, yield candidate genes that are related to the disease. Since this genes are found in a certain model organism, the resulting genes need now to be incorporated to a human context to find their relation with the known genes and proteins of the disease.

The method presented in this paper produces a combined interactome, in which different source networks are combined into a single, wider covering protein-protein interaction network. The main challenge in the construction of a combined interactome are the different data sources and formats, which should be taken into account to avoid inaccurate representation of the biological data.

In the following section we present a description of some important biological concepts that are key in understanding the method. Then, the our General Method to combine interactomes is presented, with a formal description of how the combined interactome is built and the expected result. Following that, we present the combined Human-Mouse interactome study case.

II. Biology

A. Genes and Proteins

The central dogma of molecular biology states that DNA is used as a permanent data storage, and that a gene is a region of DNA with code that defines proteins. In order to construct proteins, genes are first transcribed to messenger RNA (mRNA) molecules, which then are transported out of the nucleus where the ribosomes translates them into proteins [7]. A gene that has produced proteins is said to be expressed.

Proteins are the essential biological molecules that perform most functions in a cell, however they do not perform their function independently. Proteins usually interact with one another forming complexes that drive most of the cell's chemistry [7]. Although indirect gene interactions may be considered as part of the interactome, we only consider protein-protein interactions (PPI) to make the interactome. For sake of simplicity, we consider that experimental data is related to genes, so the resulting interactome can must be a gene interactome, while

978-1-5090-1633-4/16/$31.00 ©2016 IEEE

the available data may be either a gene or a protein interactome [1]. From a Graph Theory perspective, the interactome is an undirected graph where nodes represent the proteins and links represent the physical interactions between the proteins.

B. Homology

Although each different organism has its own defining DNA sequence, and each organism has its own characteristics, genes in the different organisms might have similar code sequences and therefore express proteins with similar function [12].

A pair of genes is said to be homologous if they descend from a common ancestral DNA, and are expected to retain similar functions [12]. Pairs of homolog genes might belong to either a single organism or a two different organisms (to be more precise, when the organisms are different, proteins are called orthologs, but for the sake of simplicity, we will just use the term homolog for most of this study).

A frequent approach to study homolog genes is to establish homology classes to group them together. Homology identification methods are generally based on sequence similarity [8], [9]. The existence of homology is key in allowing researchers to perform genomic experiments in one organism and infer the results in a related organism.

C. Model Organisms

A model organism is a species that has been widely studied, usually because it is easy to maintain and breed in a laboratory setting and has particular experimental advantages [11].

Notable examples of model organisms are the house mouse, and the E. coli bacteria (Escherichia coli). The house mouse has a very high breeding rate for a mammal, and a extremely high diet variety. This characteristics and the fact that mice are among the closest relatives of humans excluding primates, makes them the most frequently mutated mammal in scientific research. On the other hand, E. coli is the most studied prokaryotic organism, as it can grow and develop easily in a research environment.

D. Interaction Prediction

Modern techniques that allow protein interaction transfer between different organisms are based on the interolog concept [14]. If interacting proteins p_a and p_b in an organism, have interacting orthologs in a different organism $p_{a'}$ and $p_{b'}$, the pair of interactions are said to be interologs. The *de facto* standards to define the possible interologs are based on reciprocal best match [15], and generalised interologs [16].

Large databases were built upon these concepts. STRING [17] provides interactomes for many organisms (some of which may only have proteins known, and no known interactions), transferring interologs for any organism with experimental data. Other more specialised databases, such as MGI [6] provide curated homology classes for human and mouse (among other genetic, genomic and biological data concerning the laboratory mouse).

To the best of our knowledge, there is no resource that provides a joint interactome with genes or proteins from different organisms.Such an interactome can be relevant for biologists (i.e. in the analysis of diseases with high co-morbidity, gene synthetic lethality, gene fusion, and others).

III. Producing a combined interactome

We formalise a General Method to integrate PPI networks from different organisms into a single interactome. The resulting combined interactome R will be represented by the graph $R = \langle V_R, E_R \rangle$, the first goal is to create the vertices of the graph V_R, then, the second goal is to establish the edges E_R. The proposed process to obtain R starts from the original PPI network for each organism, then creates an intermediate extended network for each organism, to finally integrate the extended networks into the combined interactome.

A. Vertices

Since the result of the method is a gene interactome, and the interactions are presented as protein-protein interactions, the vertices of the resulting network will require information about genes and the proteins they express.

Let a gene g be defined as $g = \langle id_g, name_g, sym_g \rangle$, where id_g is the gene identifier, $name_g$ is the gene's official name, and sym_g is the gene's official symbol. Likewise, let a protein p be defined as $p = \langle id_p, name_p, sym_p \rangle$, where id_p is the protein identifier, $name_p$ is the protein's official name, and sym_p is the protein's official symbol. Then, to build the resulting network, the genes must be matched with the proteins they express.

In many cases, the official symbol is the only information available for genes and proteins. An identifier convention for the genes and proteins has to be defined for these cases in which information is lacking. This convention will match gene or protein names and identifiers to the symbols using a database for the selected convention. It is likely that the PPI networks used as source will already have a defined identifier convention for the proteins, then, the selection of the protein identifier is a simple decision in most cases. The gene identifier convention can be selected in a convenient way to build the combined interactome.

Formally, we define a gene or protein database as a function [1] $DB : ID \to Name \times Sym^+$, where ID is the set of molecule identifiers, $Name$ is the set of official molecule names, and Sym is the set of molecule symbols ($Sym^+ = \{sym_i\}$ are the symbols that represent the molecule, where sym_1 is the official symbol and the remaining symbols are synonyms). To obtain identifiers from the symbols, we define a function $DB^c : Sym \to ID^+$, where:

$$id \in DB^c(sym) \leftrightarrow sym \in DB(id). \qquad (1)$$

Symbols do not always have a single identifier in frequently used databases. Since each database identifier stands for a unique molecule, keeping multiple identifiers per symbol can lead to inaccurate results. Unless there is a clear criteria to select one identifier over another, the only safe decision is to

[1] A^+ represents a subset of one or more elements from the set A

avoid using the symbols that have multiple identifiers in the combined interactome. After solving the multiple identifier problem, the reverse database identifier function $DB^R : Sym^v \rightarrow ID$ maps every valid symbol $Sym^v \subseteq Sym$ from the database to one identifier.

The set of available genes for an organism O_g is given by:

$$\begin{aligned}\langle id_g, name_g, sym_g\rangle \in O_g \leftrightarrow\ & sym_g \in Sym^v_{O,g}\\ & \wedge DB^R_{O,g}(sym_g) = id_g\\ & \wedge DB_{O,g}(id_g) = \langle name_g, Syms_g\rangle\end{aligned} \tag{2}$$

where $DB_{O,g}$ is the gene database function, $DB^R_{O,g}$ is the reverse gene database function and $Sym^v_{O,g}$ are the valid gene symbols of the organism O, and $Syms_g$ are the symbols that are mapped by the identifier id_g in the database (note that $sym_g \in Syms_g$ is the official symbol). Likewise, the set of available proteins for an organism O_p is given by:

$$\begin{aligned}\langle id_p, name_p, sym_p\rangle \in O_p \leftrightarrow\ & id_p \in ID_{O,g}\\ & \wedge DB_{O,g}(id_p) = \langle name_p, Syms_p\rangle\\ & \wedge sym_p \in Syms_p\end{aligned} \tag{3}$$

where $DB_{O,p}$ is the protein database function, and $ID_{O,g}$ is the set of protein identifiers of the organism O. This first step is to use Equations 2 and 3 on each organism, and obtain the sets of genes and proteins available for each organism.

The following task is to translate genes to proteins. This will require the usage of a convention mapping resource of identifiers in addition to the simple symbol comparison. Since a gene g can express multiple proteins $\{p_i\}$, the gene expression function[2] is $X : G \rightarrow P^*$, where G is a set of genes, and P is a set of proteins. The resulting protein set P^* can be empty if there is no proteins associated for the gene, this could happen when there is no known protein expressed for the gene or due to mapping problems between the different databases. We propose the following mapping function from genes to proteins:

$$p \in X(g) \rightarrow \begin{cases} \hfill sym_p = sym_g \vee \\ (sym_p \neq sym_g \wedge id_p \in T(id_g)) \end{cases} \tag{4}$$

where $T : ID_g \rightarrow ID_p$ is an identifier convention translation function, provided in databases by official nomenclature committees. Note that we choose the molecule symbol as the primary translation element, since the gene identifier is already obtained from the gene symbol.

The gene to protein translation will extend the information of the vertices from the original PPI network for an organism O. Since the original protein interactome only has proteins as vertices, we propose to combine the protein vertices O_p given by Equation 3 with the valid genes O_g given by Equation 2, yielding a set $O_{g,p}$ of gene-protein tuples defined by:

$$gp =< g, p >\in O_{g,p} \leftrightarrow p \in X(g). \tag{5}$$

The vertices of each organism must be extended in this step in order to continue the process.

The final task is to combine the vertices of the different organisms into one set of vertices for the combined interactome. This procedure is independent of the chosen homology identification method.Homolog genes among different organisms can be treated as equivalent genes, this defines homology classes that group all homolog genes. Without loss of generality, let O and Q be a pair of organisms. The homology class set $H = \bigcup H^i$ establishes a partition in the gene sets O_g and Q_g, defined as:

$$\begin{aligned} H^i &= O^i_g \cup Q^i_g \qquad & O^i_g \cup Q^i_g \neq \varnothing, \quad & \forall i\\ O_g &= \bigcup O^i_g \qquad & O^i_g \cap O^j_g = \varnothing, \quad & \forall i \neq j\\ Q_g &= \bigcup Q^i_g \qquad & Q^i_g \cap Q^j_g = \varnothing, \quad & \forall i \neq j, \end{aligned} \tag{6}$$

where $\{H^i\}$ is a partition of H, $\{O^i_g\}$ is a partition of O_g, and $\{Q^i_g\}$ is a partition of Q. Note that either $O^i_g \subset O^i$ or $Q^i_g \subset Q^i$ can be empty when a gene has no homolog in the other organism. We propose to create the resulting vertex set V_R as:

$$\begin{aligned} v^i = \langle gp_O, gp_Q\rangle \in V_R \leftrightarrow p_O \in H^i \wedge q_O \in H^i\\ v^i = \langle gp_O\rangle \in V_R \leftrightarrow p_O \in H^i \wedge Q^i_g = \varnothing\\ v^i = \langle gp_Q\rangle \in V_R \leftrightarrow p_Q \in H^i \wedge O^i_g = \varnothing \end{aligned} \tag{7}$$

where $gp_O =< g_O, p_O >\in O_{g,p}$ is an extended vertex for organism O, and $gp_Q =< g_Q, p_Q >\in Q_{g,p}$ is an extended vertex for organism Q. In this approach, v^i holds at most one homolog gene of a homology class of each organism. This combined vertex is related to the concept of protein interologs [13].

B. Edges

The interaction between proteins of the original PPI network for an organism O is symmetric (if protein p interacts with protein p', then protein p' interacts with protein p), this means that the graph edges are undirected. Therefore, the edges of the graph established by the original PPI are defined as $\{p, p'\} \in E_I \subset O_p \times O_p$ when protein p and protein p' interact. We define the edges of the extended gene interactome $E_{I,g} \subset O_{g,p} \times O_{g,p}$ as:

$$\{gp, gp'\} \in E_{I,g} \leftrightarrow \{p, p'\} \in E_I, \tag{8}$$

where $p, p' \in O_p$ are extended to $gp =< g, p >, gp' =< g', p' >\in O_{g,p}$ respectively, according to Equation 5.

The final step is to create the set of edges E_R for the resulting combined graph R. We define $E_R \subset H \times H$ as:

$$\{v, v'\} \in E_R \leftrightarrow \{gp_O, gp'_O\} \in E_{I,O} \vee \{gp_Q, gp'_Q\} \in E_{I,Q}, \tag{9}$$

where $gp_O \in O_{g,p}, gp_Q \in Q_{g,p}$ and $gp'_O \in O_{g,p}, gp'_Q \in Q_{g,p}$ are combined into $v \in V_R$ and $v' \in V_R$ respectively, according to Equation 7.

[2] A^* represents a subset of zero or more elements from the set A

IV. THE HUMAN-MOUSE INTERACTOME

The construction of a Human-Mouse Interactome is of particular interest since the the house mouse is the most similar model organism to the human. The details of an implementation of the general method are presented in this subsection.

The source of the original molecule interaction network was BioGRID, this is a public database of hand curated molecule interaction datasets from model organisms and humans [5]. The process of building a combined Human-Mouse interactome from independent Human and Mouse BioGRID databases is described in this section.

The process starts with the Human and Mouse gene interaction databases provided by BioGRID v3.2. The format of this database is an entry per curated publication on each interaction. The database fields that are related to the network structured mentioned in the General Method section are:

- *From interactor A*: Entrez Gene ID, Systematic name, Official symbol, and Synonyms/Aliases.
- *From interactor B*: Entrez Gene ID, Systematic name, Official symbol, and Synonyms/Aliases.

A. Vertices

To obtain every vertex from the database, each database entry must be parsed. The database format holds the definition $DB : ID \to Name \times Sym^+$ as:

- Entrez Gene ID $\in ID$
- Systematic name $\in Name$
- Official symbol and Synonyms/Aliases $\in Sym$,

from this, we map the format for a gene vertex $g = \langle id_g, name_g, sym_g \rangle$ directly from the database fields as:

- id_g =Entrez Gene ID
- $name_g$ =Systematic name
- sym_g =Official symbol.

An interesting detail here is that we do not recommend to use the gene synonyms for the network nodes. The synonyms for each gene can be defined by a function function $S : Sym- > Sym^+$, where Sym is the set of gene symbols. Looking further into the synonym concept, the whole gene symbol set Sym should be partitioned into synonym subsets, as $Sym = \bigcup Sym_i$ given that $Sym_i \cap Sym_j = \emptyset, \forall i \neq j$, where $Sym_i \subseteq Sym$ is a synonym subset. Then, a synonym subset Sym_i verifies the following property:

$$sym \in Sym_i \to S(sym) \subseteq Sym_i, \forall sym \in Sym_i. \quad (10)$$

The synonym subsets for the BioGRID genes presented an unexpected result that lead to giving up this approach. The synonym subsets were too big, one of them had over 2500 protein names, this was clearly a set that contained different genes, making the whole process of combining them into a single element not only pointless, but incorrect. The problem was traced to names such as P14, synonym of proteins such as Peroxidase 14 and Ragulator complex protein LAMTOR2 (Endosomal adaptor protein p14), that are not related.

As the original interactome is a gene interactome, there is no need to translate genes to proteins. This can be regarded already as the intermediate network where the only step needed is to create the homology classes. We chose to use the Human-Mouse homology classes resource provided by MGI [6] for this step. This resource is used to build the homology classes H. The format of the database includes one gene to the homology class per entry, in this way several entries refer to the same class. The fields that are relevant to build the class are:

- HomoloGene ID
- Common Organism Name
- EntrezGene ID.

With this data, all the genes that belong to the homology class can be identified by the HomoloGene ID. Therefore, the number of different homology classes is given by the number of different HomoloGene ID values in the database. In order to use the same notation as in the General Method section, let human be organism O and mouse be organism Q. The gene from the database entry will belong to O if the Common Organism Name field value is ”human”, and it will belong to Q if the Common Organism Name field value is ”mouse, laboratory”. The gene node $g = \langle id_g, name_g, sym_g \rangle$ can be obtained matching id_g with the EntrezGene ID field value.

To avoid the same problem presented by the gene synonyms, we choose to obtain a representative gene on each homology class per organism in the combined vertex set V_R. Since the O and Q hold homologs of the same organism, there are classes $\{O_i\}$ and $\{Q_j\}$ with more than one gene inside. We decided to choose between them according to characteristics of the gene symbol in those sets. As the gene symbol is hand curated, homolog genes usually share a common base string (i.e. Pnp and Pnp2). The first criteria to select the class representative is to find if a gene symbol is contained among the other gene symbols of the class. If this criteria is not met, the gene with the smallest Entrez identifier is kept. The set of vertices V_R is built according to Equation 7. An important point to notice is that the BioGRID Human database also includes genes from other species such as the house mouse or the Norwegian rat, likewise the BioGRID Mouse database includes genes from organisms such as human and the Norwegian rat. We create nodes for each of those genes, but if the BioGRID Human database includes a mouse gene, it is identified as such in the homology classes and its human homolog is found, correspondingly, if the BioGRID Mouse database includes a human gene, it is identified as such in the homology classes and its mouse homolog is found.

B. Edges

Building the edge set E_R for the combined network is an easier task compared to building the vertex set V_R. As the BioGRID databases includes one entry per curated interaction evidence between pair of genes, a new entry does not translate into a new edge in every case. The handling of the multiple evidences per interaction depends on the needs of the implementation, we will not develop into that topic as it does not modify the structure of the resulting combined graph.

Recall that each BioGRID entry provides the fields Entrez Gene ID, Systematic name, Official symbol, and Synonyms/Aliases for each interactor. As all the gene interactors g of the organism O are included in the set O_g, g can be identified by its Official Symbol (sym_g equal to Official Symbol). This means that the edge $\{g, g'\} \in E_{I,g}$ if there is any entry in the BioGRID database for O, where g is the interactor A and g' is the interactor B, or g' is the interactor A and g is the interactor B.

The construction of the set E_R from $E_{I,g}$ and V_R is finally straightforward following equation 9.

C. *Results*

The number of vertices and edges of the original BioGRID graphs, compared to the number of entries of each database is described in Table I.

TABLE I
ORIGINAL DATABASES

	Vertices	Edges	DB Entries	Repeated Edges
Human	19402	162484	261175	98691
Mouse	8503	19032	23635	4603

Table I shows that 98691 entries do not provide additional interactions in the Human BioGRID database. Repeated edges are included to the database to add evidence for the interaction of a pair of proteins (i.e. different experimental sources that confirm the interactions). Analogously 4603 entries do not provide additional interactions in the Mouse BioGRID database.

TABLE II
INDEPENDENT DATABASES WITH HOMOLOGY

	Vertices	Repeated Vertices	Edges	Repeated Edges
Human	16959	2443	160316	2168
Mouse	6824	1679	18227	805

Table II shows the results of finding homologs independently on each interactome. We identified 2443 homolog genes in the BioGRID Human database. This high number of homolog genes is the consequence of having both human and mouse genes in the database. Likewise, the BioGRID Mouse database contains 1679 homolog genes. Although the percentage of repeated vertices is over 14% in the human interactome and over 24% in the mouse interactome, the percentage of repeated edges is significantly smaller in both cases. The percentage of repeated vertices is less than 2% in the human interactome, and less than 5% in the mouse interactome. Our hypothesis for this phenomenon is that homolog genes are generally added to increase the number of connections in the BioGRID interactomes.

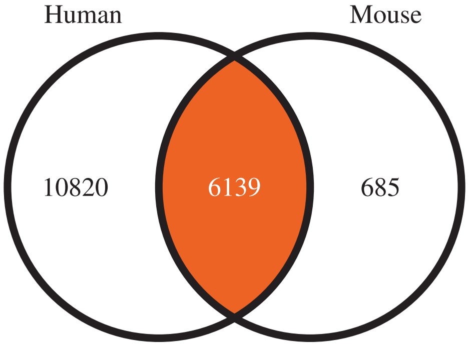
Finally, the number of vertices and edges of the combined Human-Mouse Interactome are:

TABLE III
HUMAN-MOUSE INTERACTOME

	Vertices	Edges
Total	17644	169458
Human	16959	168084
Mouse	6824	97348
Exclusively Human	10820	12452
Exclusively Mouse	685	36

To add clarity to the data form Table III, total vertices refers to the complete Human-Mouse interactome, Human vertices is when the source is the human interactome, and Mouse vertices is when the source is the mouse interactome. Note that there are overlaps between both sets of vertices since pairs of homolog genes from both networks become only one vertex. Vertices exclusively from one organism are those which do not have an homolog vertex in the other interactome. Total edges is number of edges in the complete Human- Mouse interactome, and the rest of the edges is the subset of those edges that lie between the vertices of each category, for instance, Human edges are the edges that lie between Human vertices.

Table III shows that, as expected, most of the benefits are for the mouse interactome, that uses the human interactome to significantly enrich its number of vertices and edges.



Gene contribution to the Human-Mouse Interactome

As the main intended usage of this interactome is to increase the coverage of mouse genes, therefore we compare the Mouse row from Table II with the Total row from table III. The most important result is an increase of over 150% in the number of vertices, and an increase of over 430% in the number of edges. This gain is not only due to the new vertices added by the human interactome, but because there is more interaction data. This can be clearly seen comparing the initial 18227 edges from the mouse interactome and the 97348 edges between the same set of vertices but using also the human edges in between them.

V. CONCLUSION

The General Method proposed in this paper presents a formal outline to construct a combined interactome between related organisms. As the General Method does not fix the input databases, nor the homology mapping, the methodology can be further customized for the required tests and organisms involved in the process. This also allows the methodology to stay relevant as gene relations are discovered.

The practical Human-Mouse study case clearly demonstrates the benefits of using combined databases, to enhance the information in organisms that have less information. More importantly, the Human-Mouse interactome can be used for studies that use biological experiments in mice, such as the large majority of cancer experiments.

Acknowledgment

This work was supported, in part, by the CONACYT Paraguay Grant INVG01-112 (14-INV-088).

References

[1] Lage, Kasper, et al. *A human phenome-interactome network of protein complexes implicated in genetic disorders*. Nature biotechnology 25.3: 309-316, 2007.

[2] Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins*. Nucleic acids research 35.suppl 1: D61-D65, 2007.

[3] Chua, Streamson C., et al. *Phenotypes of mouse diabetes and rat fatty due to mutations in the OB (leptin) receptor*. Science 271.5251: 994-996, 1996.

[4] Yuan, Junying, et al. *The C. elegans cell death gene ced-3 encodes a protein similar to mammalian interleukin-1β-converting enzyme*. Cell 75.4: 641-652, 1993.

[5] Chatr-aryamontri, Andrew and Breitkreutz, Bobby-Joe and Oughtred, Rose and Boucher, Lorrie and Heinicke, Sven and Chen, Daici and Stark, Chris and Breitkreutz, Ashton and Kolas, Nadine and O'Donnell, Lara and others, *The BioGRID interaction database: 2015 update*, Nucleic acids research, pp. gku1204, Oxford Univ Press, 2014.

[6] Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE; The Mouse Genome Database Group. 2015. *The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease*. Nucleic Acids Res. 2015 Jan 28;43(Database issue):D726-36.

[7] Cohen, William. *A Computer Scientist's Guide to Cell Biology*, 1st ed. Springer, 2007.

[8] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. *Basic local alignment search tool*. Journal of Molecular Biology, pp. 215:403-410, 1990.

[9] S. R. Eddy, *Multiple Alignment Using Hidden Markov Models*. Proc. Third Int. Conf. Intelligent Systems for Molecular Biology, pp. 114-120, 1995.

[10] World Health Organization. *Standards and operational guidance for ethics review of health-related research with human participants*. 2011.

[11] Miklos, George L. Gabor, and Gerald M. Rubin. *The role of the genome project in determining gene function: insights from model organisms*. Cell 86.4: 521-529, 1996.

[12] Sattler, Rolf. *Homology-a continuing challenge*. Systematic Botany: 382-394, 1984.

[13] Matthews, Lisa R., et al. *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"*. Genome research 11.12 (2001): 2120-2126.

[14] Walhout, Albertha JM, et al. *Protein interaction mapping in C. elegans using proteins involved in vulval development*. Science 287.5450 (2000): 116-122.

[15] Li, Siming, et al. *A map of the interactome network of the metazoan C. elegans*. Science 303.5657 (2004): 540-543.

[16] Yu, Haiyuan, et al. *Annotation transfer between genomes: proteinprotein interologs and proteinDNA regulogs*. Genome research 14.6 (2004): 1107-1118.

[17] Szklarczyk, Damian, et al. *STRING v10: proteinprotein interaction networks, integrated over the tree of life*. Nucleic Acids Res. 2015 43(Database issue):D447-52.