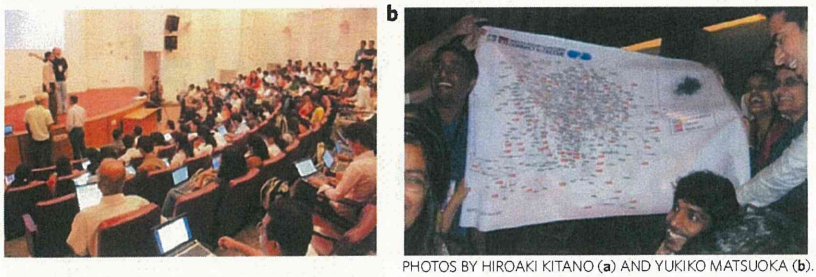


PHOTOS BY HIROAKI KITANO (**a**) AND YUKIKO MATSUOKA (**b**).

Figure 1 | Large projects require many enthusiastic and fulfilled participants. (**a**) A scene from a Connect2Decode session, a portion of the OSDD project funded by CSIR, India. More than 200 participants gather for a week for the final assembly of the pathway. Each student is given a laptop and input pathway-interaction data from publications. A detailed workflow of the project is in **Box 1**. (**b**) The map of *Mycobacterium tuberculosis* metabolic pathways on a poster presented by joyous students. The upper right box of the poster shows the entire dense 'hair-ball' network comprising more than 1,000 reactions involved in an entire metabolic network. The main component of the poster represents a part of the network in the central metabolic cycle.

fully funded, it is impractical to assume that sufficiently large numbers of researchers will be willing to focus on a single species and biological problem and put their own established systems aside.

Existing resources such as pathway databases, developed increasingly by manual curation of publications in which curators read all relevant papers one by one for precise computational model developments are also not a viable solution. Pathway databases do not necessarily cover all relevant molecules and interactions, nor are they necessarily accurate. The current gold standard for such databases is manually curated models carefully built from the literature by small groups of people hired by the project, who curate every associated study for a small subsystem[2]. This has been termed 'deep curation' and is exemplified by a series of comprehensive molecular-interaction maps[3,4]. However, the deep curation of large-scale network maps from the literature is extremely labor-intensive and stressful work. Also, it is very difficult at a sociological level to motivate manual curators to continuously update maps and models to keep up with new discoveries over many years. Automated literature mining has been extensively pursued, but replacing manual curators is decades away. At the same time, quality control is dependent upon the individual groups, and updating and correcting errors can be slowed by this centralization. The solution to this problem affects the productivity and practicality of computational approaches for drug discovery.

Is WikiBiology a solution?

One approach toward building a comprehensive and rich data resource is to follow the success of Wikipedia, which is continuously updated and covers every possible subject. There is increasing interest in a Wikipedia-like approach, also called a 'Web 2.0' or 'community-based' approach, in biology. There are several attempts to create Wikipedia-like resources such as WikiGenes and WikiPathways[5], as well as open-access approaches such as Science Commons, which aims to promote extensive sharing of knowledge through a community-driven, bottom-up strategy. Whereas there is often a core group that receives funding to create an initial seed of these types of Web sites, the wiki-like approach fundamentally relies on voluntary contributions from community members.

One of the remarkable features of Web 2.0, such as Wikipedia and Google Earth Community, is the collective contribution of knowledge and experiences to a globally shared Web space. Although considerable resources must be diverted for quality control and prevention of spam[6], services such as Wikipedia have been extremely effective and have grown to be indispensable resources. The question remains, however, whether such a model can be applied to biology successfully.

Motivational and sociological factors are critical to the success of the community-driven system. Skeptical views exist on whether biologists are willing to spend time to provide feedback on community efforts. Indeed, there is no mechanism to reward contribution in any formal way. Contributing to a wiki-based biology site is certainly not a factor in hiring and promotion, nor is it considered an honor within the scientific community[7]. In the extreme case, disclosing one's knowledge may help to speed up a competitor's research. It should be noted that this does not mean that large-scale collaboration for pathway curation is not feasible. In fact, there are efforts to construct consensus pathway maps as a community activity. An example of such a project was the reconstruction of the yeast metabolic network[8]. A series of genome annotation projects fall into the same category. In these projects, members are well defined, and they gain the spoils of authorship on published works. The social dynamics therefore differ and are in fact quite the opposite of those in the Web 2.0 approach.

We should carefully look at reasons why projects such as Wikipedia and Linux have soared and keep flying. In the case of Linux, a hacker culture supported open sourcing and sharing of knowledge, as signified by the Free Software Foundation, founded by Richard Stallman, where contribution to the community at large was the pride of the hackers. At the same time, there was a practical need to develop open-source operating systems as opposed to closed commercial systems. Among the efforts for open-source operating systems exemplified by FreeBSD, Linux survived mainly because it happened at the right moment and had more applications and publications than other initiatives. In the absence of Linux, FreeBSD or other initiatives would have filled this space. Wikipedia essentially inherited a similar culture. Having goals that are widely shared, are exciting and provide a sense of participation has been the key factor driving the community-based initiative. Whether these motivations are sustainable over time is yet to be seen, although certainly they were effective in getting these projects to maturity.

Unfortunately, such a hacker culture does not exist in biology today. At least, it is not a mainstream idea. Short of a formalized recognition system, any chance of a successful Web 2.0 approach will require a cultural shift in the community. Assignment of microimpact factors, or microattribution, for contribution to such an initiative should be considered. However, for such a system to be truly effective, such indices must be a core part of the merit system[9]. The citation index, for example, is simple and widely used as a measure of scientific influence. For microattribution to be accepted in the merit system, it has to acquire universality and attain the same "currency" status as the citation index in merit evaluation. Receiving microattributions in exchange for contributions should still be considered a weak motivating factor. In the most successful projects, people are driven when the vision, passion and dedication of the project are aligned with individual aspirations.

Emergent collaborations in engineering

It is useful to learn from successful emergent collaboration projects in other fields. One of the most successful of these projects in robotics and artificial intelligence is RoboCup[9], an initiative started in the mid-1990s with the aim of developing a series of high-impact technologies. A landmark goal was set that states, "by mid-21st century, a team of fully autonomous humanoid robot soccer players shall win the soccer game, comply with the official rule of the FIFA, against the winner of the most recent World Cup" (http://www.robocup.org/). Over the last 15 years, the project successfully attracted more than 4,000 researchers for scientific and technological research and hundreds of thousands of students from elementary to undergraduate levels for educational practices from at least 35 countries annually. There is no central research funding, so each researcher acquires his or her own funding. Extensive collaborations are taking place without top-down coordination. The governing body, the RoboCup Federation, only organizes the annual convention, publishes proceedings and approves regional activities, akin to an academic society. However, what is different from academic societies is that all the federation's activities are ultimately focused on achieving a defined goal rather than general promotion of the field. At the same time, it is not a traditional project, because there is no principal investigator (PI) to govern it as a whole, and the management team of the RoboCup Federation is democratically elected every year from the community. This seems to promote a sense of ownership of the project, and it thereby avoids becoming an orphan.

This emergent collaboration is sustainable for the following reasons: the goal is wildly exciting, widely shared and understood even by nonscientists; it is clear that the goal cannot be achieved by a single group or country; collaboration to promote the technology benefits everyone involved; and participants have the satisfaction of knowing that some spin-off activities have emerged that contribute to major humanitarian efforts, such as RoboCup Rescue, which was involved in the rescue operation at the World Trade Center on 9/11 (ref. 10) and more recently in the earthquake in Japan on 11 March. In addition, contributions to RoboCup offer professional opportunities beyond participation in the project itself, and many contributions will be published as well.

With regard to knowledge sharing, a principle is imposed in emergent collaborations that technical details must be disclosed at the end of the annual world convention. In some cases, facilitating a multidirectional open flow of knowledge involves mandatory disclosure of technical details and source code as a precondition for obtaining access to the rich source-code repository. This feature distinguishes this type of large-scale emergent collaboration from a wiki-type project. Proper means of collaboration would be different based on the interest of potential stakeholders, and it is critical to choose the right level of incentive mechanisms, commitment and benefits for each interested party[11]. RoboCup is one of the rare projects in which vision, social appeal, passion and ego fulfillment have been well aligned with the professional and personal aspirations of the participants.

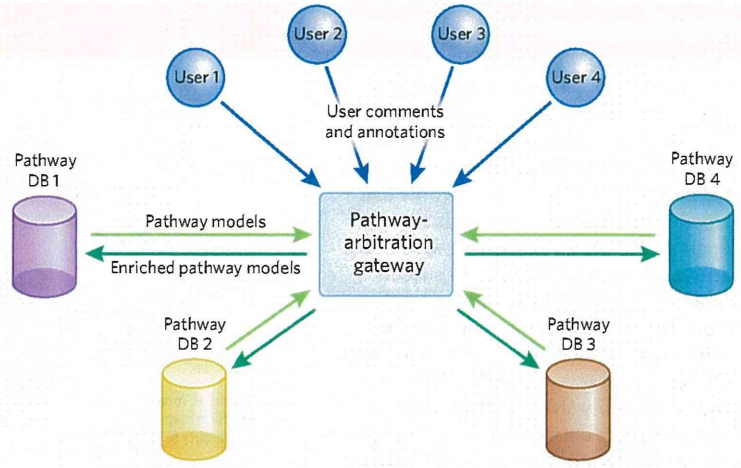


Figure 2 | Conceptual diagram of an open-flow model of knowledge sharing and integration. In the open-flow model, contributors benefit as well. Each pathway-database provider may contribute his or her pathway information to enrich and correct pathways placed on the open-flow gateway or an arbitration site, then may also obtain information enriched through other pathway databases (DB) as well as individual contributors. This model is essentially a combination of community-based pathway annotation, Creative Commons Attribution-ShareAlike and RoboCup-like social engineering.

Toward the open-flow paradigm of knowledge aggregation

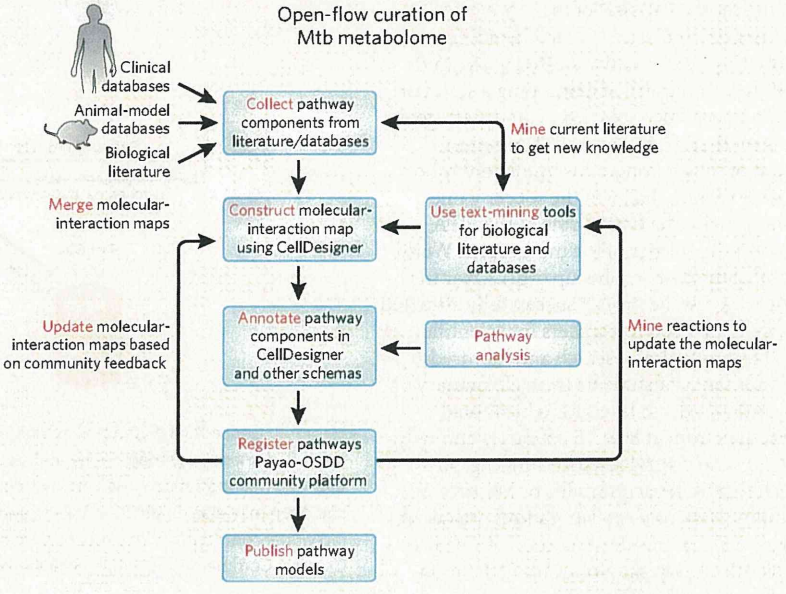
The field of drug discovery for neglected diseases could meet the criteria for a successful emergent collaboration project. It is a good cause, is socially appealing, needs collective efforts and can impart a sense of pride to participants for their contributions. A recent attempt is the Open Source Drug Discovery (OSDD) project initiated by the Council of Scientific and Industrial Research (CSIR), a branch of the Indian government[12]. OSDD is aimed specifically at drug discovery for tuberculosis through open collaboration. As an initial phase of the program, major funding within India was allocated for genome annotation, network reconstruction and a set of initial screens. More than 830 researchers and students are participating in the project, with 200 students tasked with curating the literature to construct a comprehensive and detailed metabolic map of *Mycobacterium tuberculosis* (**Fig. 1**). Unlike past efforts for collective pathway reconstruction and curation, such as the yeast metabolic map initiative, any researcher or student can join the effort on a volunteer basis, with a core principal investigator team having received funding from CSIR to drive the project in a sustainable manner. Thus, this project is distinctively open ended in terms of both quantity and quality of participation.

This project satisfies some of the criteria for a successful emerging project, such as having a clear and appealing goal and an alignment with participant motivations. A distributed and collective pathway reconstruction that was performed with the OSDD project was a true social experiment. With this experiment, a large-scale distributive reconstruction of biological networks was shown to be possible with the proper software platform, a well-defined workflow (**Box 1**), and project management when the objectives of the project were designed to motivate potential participants.

Unresolved issues include quality assurance over time and the need for continuous maintenance and updating of the map. Although not quite a strategy, the hope is that some of the participants will continue to work on these issues, but owing to the open-ended and emergent nature of the project, this cannot be expected or guaranteed. An alternative to the continued work by the participants is a call to the wider community of scientists for verification in a wiki-based challenge. Because of the

Box 1 | Distributed pathway reconstruction workflow

A project-wide distributed reconstruction has been tested using CellDesigner network editing software[18], a Payao-based community curation system[19], and a workflow (at right) for a large-scale distributed curation of biological networks that was provided by the Systems Biology Institute, Tokyo, Japan, as part of an agreement with CSIR. Participants use CellDesigner to draw subsets of molecular interactions along with annotation information. The interactions are merged and relaid out in CellDesigner. The interaction data are then sent to a Payao-based server installed in Tokyo for storage and community sharing. The integrated molecular-interaction database is accessible to all team members. Multiple iterations of pathway construction and integration take place. During this process, members can add notes and comments on the interaction map. After a few months of distributed curation sessions, everyone from the Indian side and Tokyo side got together in Delhi for a week for the final assembly of the entire network. Mtb, *Mycobacterium tuberculosis*.



social significance of finding an effective cure for drug-resistant tuberculosis, such a proposal may attract people who are willing to contribute even without personal or professional benefit. Medical practitioners, retired scientists and those with scientific backgrounds who are not directly engaged in science as their profession may be willing to participate in such a challenge. It is critical, therefore, to design the project in the context of proper social dynamics so that potential contributors become and remain motivated. Of course, this may still be wishful thinking and therefore may fail to generate sufficient contributions.

A more deliberate strategy is to apply the open-flow approach among pathway-database providers. As with the code-access principle in RoboCup, the basic tuberculosis map could be made accessible to pathway-database providers under an appropriate licensing scheme such as the Attribution-ShareAlike licensing in Creative Commons. This allows pathway-database providers to integrate the tuberculosis map into their database, but any corrections and enrichment they make are to be shared with the community. Creating a central gateway to various pathway databases would allow for the integration of currently fragmented knowledge. An integrated collection of pathway databases would also allow for a centralized forum for feedback to the original map providers. Such bidirectional and open exchange of information among the community and data-resource providers is certainly critical in maintaining quality data (**Fig. 2**). In the long run, it may solve the intrinsic trade-off between quality and coverage inherent to current pathway databases. The flagship project for tuberculosis can be used as a proof of concept, and similar efforts such as the GlaxoSmithKline deposition of malaria-related compounds to ChEMBL may provide the other case of open collaboration[13].

It should be noted that such widespread and large collaborations are possible now because of the development of various standards and software that complies with those standards. Systems Biology Markup Language (SBML)[14], Systems Biology Graphical Notation (SBGN)[15] and BioPAX[16] all ensure a certain level of interoperability. However, technology alone cannot make things work, particularly when projects necessarily involve large numbers of parties with varying motivations, career aspirations and opinions. Broader social consideration can be a major key for success when launching increasingly complex projects[17]. Social engineering will be recognized as an indispensable part of research activity in the coming years for large-scale and complex big science, because it is the people who do science, not technology or machines. ■

Hiroaki Kitano is in the Systems Biology Institute, Tokyo, Japan, Sony Computer Science Laboratories, Tokyo, Japan, and the Okinawa Institute of Science and Technology, Okinawa, Japan. Samik Ghosh is in the Systems Biology Institute, Tokyo, Japan. Yukiko Matsuoka is in the Systems Biology Institute, Tokyo, Japan, and the Japan Science and Technology Agency, Tokyo, Japan.
e-mail: kitano@sbi.jp

References

1. Donald, P.R. & van Helden, P.D. *N. Engl. J. Med.* **360**, 2393–2395 (2009).
2. Bauer-Mehren, A., Furlong, L.I. & Sanz, F. *Mol. Syst. Biol.* **5**, 290 (2009).
3. Caron, E. *et al. Mol. Syst. Biol.* **6**, 453 (2010).
4 Oda, K. & Kitano, H. *Mol. Syst. Biol.* **2**, 2006.0015 (2006).
5. Pico, A.R. *et al. PLoS Biol.* **6**, e184 (2008).
6 Kittuer, A., Suh, B., Pendleton, B. & Chi, E. in *CHI 2007 Proceedings*. 453–462 (ACM, 2007).
7. Callaway, E. *Nature* **468**, 359–360 (2010).
8. Herrgård, M.J. *et al. Nat. Biotechnol.* **26**, 1155–1160 (2008).
9. Kitano, H. *et al. AI Mag.* **18**, 73–85 (1997).
10. Murphy, R. *et al.* in *Springer Handbook of Robotics* (eds. Siciliano, B. & Khatib, O.) Ch. 50, 1152–1174 (Springer-Verlag, New York, 2008).
11 Altshuler, J.S. *et al. Sci. Transl. Med.* **2**, 52cm26 (2010).
12. Singh, S. *Cell* **133**, 201–203 (2008).
13. Nwaka, S. & Ridley, R.G. *Nat. Rev. Drug Discov.* **2**, 919–928 (2003).
14. Hucka, M. *et al. Bioinformatics* **19**, 524–531 (2003).
15. Le Novère, N. *et al. Nat. Biotechnol.* **27**, 735–741 (2009).
16. Demir, E. *et al. Nat. Biotechnol.* **28**, 935–942 (2010).
17. Hill, C.T. *Issues Sci. Technol.* **24**, 78–84 (2007).
18. Funahashi, A. *et al. Proc. IEEE* **96**, 1254–1265 (2008).
19. Matsuoka, Y., Ghosh, S., Kikuchi, N. & Kitano, H. *Bioinformatics* **26**, 1381–1383 (2010).

Competing financial interests

The authors declare no competing financial interests.

Bioinformatics Advance Access published July 28, 2011

Tissue Specific subnetworks and characteristics of publicly available human protein interaction databases

Tiago J.S. Lopes[1], Martin Schaefer[2], Jason Shoemaker[1], Yukiko Matsuoka[1,3], Jean-Fred Fontaine[2], Gabriele Neumann[4], Miguel A. Andrade-Navarro[2], Yoshihiro Kawaoka[1,4,5], Hiroaki Kitano[3,6,7,8,*]

[1]*JST ERATO KAWAOKA Infection-induced Host Responses Project, Tokyo, Japan*
[2] *Computational Biology and Data Mining, Max Delbrück Center for Molecular Medicine, Berlin, Germany*
[3]*The Systems Biology Institute, Tokyo, Japan*
[4]*Department of Pathobiological Sciences, Influenza Research Institute,University of Wisconsin-Madison, School of Veterinary Medicine, Madison, Wisconsin, USA*
[5] *Institute of Medical Science, Division of Virology, Department of Microbiology and Immunology,University of Tokyo, Tokyo, Japan*
[6]*Sony Computer Science Laboratories, Inc., Tokyo, Japan*
[7]*Open Biology Unit, Okinawa Institute of Science and Technology, Okinawa, Japan*
[8]*Division of Cancer Systems Biology, Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan*

Associate Editor: Prof. Burkhard Rost

ABSTRACT

Motivation: Protein-protein interaction (PPI) databases are widely used tools to study cellular pathways and networks, however there are several databases available that still do not account for cell type-specific differences. Here, we evaluated the characteristics of six interaction databases, incorporated tissue-specific gene expression information and finally, investigated if the most popular proteins of scientific literature are involved in good quality interactions.

Results: We found that the evaluated databases are comparable in terms of node connectivity (i.e., proteins with few interaction partners also have few interaction partners in other databases), but may differ in the identity of interaction partners. We also observed that the incorporation of tissue specific expression information significantly altered the interaction landscape and finally, we demonstrated that many of the most intensively studied proteins are engaged in interactions associated with low confidence scores. In summary, interaction databases are valuable research tools but may lead to different predictions on interactions or pathways. The accuracy of predictions can be improved by incorporating datasets on organ- and cell type-specific gene expression, and by obtaining additional interaction evidence for the most 'popular' proteins.

Availability: Supplementary information is available at the Bioinformatics Journal website.

Contact: kitano@sbi.jp

*To whom correspondence should be addressed.

1 INTRODUCTION

Traditionally, studies that assess the cellular metabolism, disease and cancer development, pathogens infections, or drug-protein interaction have focused on single genes or proteins. While such studies have created large amounts of data, they typically do not account for the multiple interactions that regulate cellular networks. Recently, high-throughput approaches including yeast two-hybrid screens (Rual, et al., 2005), immunoprecipitation studies followed by mass-spectrometry analysis (Ewing, et al., 2007), transcriptomics (Wilhelm, et al., 2008) and metabolomics studies (Braaksma, et al., 2011) have become important research tools to identify protein-protein interaction partners (Krogan, et al., 2006) or cellular factors that are up- or down-regulated in response to specific stimuli (Bhattacharya, et al., 2004). With the availability of the resulting large datasets, the challenge now lies in the generation of comprehensive and robust interactome maps, ideally capturing all protein-protein interactions within a cell and between cells at any given moment in time.

The human proteome is estimated to encompass 130,000–650,000 protein-protein interactions (Stumpf, et al., 2008; Venkatesan, et al., 2009). Of those, only a subset has been described at this point, establishing protein-protein interaction (PPI) databases that provide valuable information about the reactions occurring at the proteome level. Previous studies analyzed and compared some of these databases (Mathivanan, et al., 2006; Ramirez, et al., 2007; von Mering, et al., 2002); however, these analyses were based on the significantly smaller data sets available at the time of the analysis, and included only

subsets of currently popular PPI databases. Therefore, we analyzed the following four popular PPI databases (Table 1): HPRD (Human Protein Reference Database, (Prasad, et al., 2009)); MINT (Molecular INTeraction, (Ceol, et al., 2010)); INTACT (Aranda, et al., 2010), and BioGRID (Biological General Repository for Interaction Datasets, (Breitkreutz, et al., 2008)). In addition, we also included in the comparison a recently published database named HIPPIE (Human Integrated Protein Protein Interaction reference - http://cbdm.mdc-berlin.de/tools/hippie/) (M.S. et al. submitted). It is assembled through the compilation of several PPI sources, including the previously mentioned databases. Lastly, for the sections of this study not involving network topological characteristics, we also included the STRING database (Search Tool for the Retrieval of INteracting Genes/Proteins) (Jensen, et al., 2009), a popular resource that in addition to protein interactions, also contains protein associations from several pathway databases. MINT, HPRD, BIOGRID and INTACT are manually curated and have thousands of interactions submitted by the community; thus, since they offer original interactions used by other databases, we refer to these four databases as 'primary resources'. HIPPIE and STRING are composed of interactions taken from primary databases and other sources; hence, we refer to HIPPIE and STRING as 'derived databases'. In addition, for the purpose of this study we removed all predicted functional associations present in STRING.

Here, we focused on the human subset of interaction databases, and as an improvement over most current analyses, we demonstrated the usefulness of organ or cell type-specific subnetworks. We analyzed these databases for their basic features including protein coverage, number of interactions and neighborhood characteristics (i.e. we compared the number and identity of interactions partners, and asked whether proteins that are a hub in one database occupy a similar position in other databases). Finally, using three databases that assign confidence scores to its interactions, we demonstrated that there is a lack of interaction data with high confidence scores for many intensively studied proteins. Additional experimental evidence for those interactions - either confirming or refuting - would significantly increase the robustness of current PPI databases.

2 METHODS

2.1 Databases

The databases were obtained from their respective websites in the following versions or latest updates: HPRD Release 9; HIPPIE 1.1; STRING 8.3; MINT 15.December.2010; INTACT 21.April.2011; BIOGRID 3.1.76. Before initiating the analysis, the following pre-processing steps were carried out: (A) We removed all redundant interactions, keeping just the interaction with the highest score. (B) For all protein entries, their database-specific identification tags were converted to a common nomenclature (Entrez Gene IDs). Proteins that did not have a matching ID in Entrez Gene were discarded. Approximately 10% of interactions had to be removed from each database.

In the STRING database we performed additional pre-processing step: we removed all interactions involving non-human proteins, left only interactions with experimental evidence or obtained from pathway and other interaction databases (i.e. removed interactions derived from co-expression, genomic neighborhood, text-mining and other predictive techniques).

2.2 Network and Statistical Analysis

All interaction databases were converted to an undirected graph and further analyzed using R (version 2.10.1) and the iGraph library (version 0.5.4). From this library we used routines to find the degree, betweenness, diameter, shortest path, immediate neighbors and clustering coefficient. The other statistical tests (Welch, Kolmogorov-Smirnov, Wilcoxon, z-score) were performed using R with 0.95 confidence interval. Pathway and Gene Ontology enrichment analysis were performed with DAVID (Huang, et al., 2008) and ConsensusPath DB (Kamburov, et al., 2011) using the default parameters values. For the enrichment analysis tests, we used the list of proteins present in the tissue-specific subnetworks as background.

2.3 Popular Genes

The file gene2pubmed from the NCBI public FTP site contains a table with Pubmed IDs and the genes present in this each abstract (sorted by species). This file was used to rank the human genes according to the number of abstracts in which they appear and to select the 10% most popular genes (2,911 entries). The file was obtained on April 22nd 2010.

2.4 Gene expression data

We obtained an Affymetrix dataset containing the transcription levels of 84 human tissues and cell lines. This dataset is publicly available for query and download from the BioGPS project (Su, et al., 2004; Wu, et al., 2009).

We obtained the normalized expression data (pre-processed using GCRMA - GeneChip Robust Multiarray Averaging (Gentleman, et al., 2004)) and divided our analysis in the following steps: first, we defined that each probe must have an absolute intensity greater than 50 for at least one condition, thus removing any probe not being moderately or strongly expressed in at least one tissue (the original datasets have specific no background level). After this cut-off, 16,704 probes remained from the original dataset of 44,775 probes. Second, we performed the Kolmogorov-Smirnov test to evaluate the normality of each probe's intensity distribution, keeping probes with suitably normal distributions. Only 211 probes had a p-value greater than 0.1 and were excluded from further analysis. With 16,493 probes remaining, we converted their Affy_ID to Entrez Gene IDs and in this conversion 3,537 probes had no matching ID. In the end, our dataset consisted of 12,956 probes that mapped to 9,176 different genes. Finally, we calculated the z-score for each probe across all tissues. Using a Z-score cut-off of 0.1 (p-value 0.46), we determined which genes were moderately to highly expressed in each tissue.

2.5 Protein Degree Categorization

We classified the proteins into three categories (high-, middle- and low-degree) according to their number of interactions. To define the appropriate ranges, we ranked the proteins in decreasing order according to their number of interactions. With this list, we used a procedure which selected two random numbers: the first in the interval [80, 98] (we refer to it as *value1*) and the second in the interval [60, *value1*] (we call it *value2*). Subsequently, we considered high-degree proteins as those that occupied a position among the top *value1*% of the ranked list. Middle-degree proteins were those that occupied a position in the interval [*value2*, *value1*]% of the ranked list and finally, the low-degree proteins were on the [1, value2]% of the list. For a visual explanation of the procedure, please refer to Supplementary Figure 1. To verify the robustness of the results, this

procedure was repeated 100 times for each pair of databases being compared and the mean and standard deviations determined. We used this procedure instead of defining a fixed number of neighbors that a protein should have to belong to each category. The differences in the network sizes would cause the results to be unfairly dependent on the ranges selected.

3 RESULTS

3.1 Database Features

Table 1 compares the features of the six databases included in the analysis. The number of proteins (nodes) in these databases ranges from ~5,200 to ~12,000. STRING and HIPPIE contain the largest numbers of proteins since they include data from several other databases in addition to their own unique data.

For all databases except STRING, the total number of interactions ranges from ~12,500 to ~73,000 (Table 1). MINT has relatively few proteins and interactions, all of which are covered by one or several of the other databases. By contrast, over 140,000 interactions are reported in STRING, which comes close to the number of estimated interactions in the human proteome (Stumpf, et al., 2008; Venkatesan, et al., 2009). We found that 4,361 proteins and 5,589 protein-protein interactions were reported in at least two different databases, with the largest overlap between STRING and HIPPIE (Supplementary Table 1). Only 1,453 proteins and 1,619 protein-protein interactions are reported in all six databases. These interactions are reported in primary resource databases and are likely to stem from the same portion of literature that was manually curated by the authors(Turinsky, et al., 2010).

Table 1. Database characteristics

	HPRD	HIPPIE	STRING'	MINT	INTACT	BIOGRID
Proteins	9,117	11,835	10,546	5,206	8,310	9,057
Interactions	36,239	72,916	144,099	12,579	33,299	37,469
Average degree[1]	8	12	-	4.83	8.01	8.27
Average betweenness	13,528	15,840	-	8,009	11,909	13,639
Diameter[3]	14	13	-	12	13	12
Average Path Length[4]	4.25	3.79	-	4.43	3.96	4.21
Clustering Coefficient[5]	0.05	0.05	-	0.03	0.03	0.06

[1]Average degree describes the average number of interactions; [2]Average betweenness describes the 'centrality' of a factor in a network; [3]Diameter describes the maximal distance between the two most distant nodes in a network; [4]Average path length describes the average number of steps that connect any two components; [5]Clustering coefficient describes the tendency of nodes to interact among each other forming groups. * STRING is not a PPI database, thus we did not compute the features that are commonly used for network structure analysis.

Next, we compared the average degree and betweenness of the proteins in each in database. The average degree (average number of interactions per protein) ranges between 5-12, with HIPPIE showing the highest average number of neighbors for each protein (Supplementary Figure 2A shows the distributions of degree and betweenness in each database). Betweenness, in a broader sense describes the significance of a node (i.e. a protein in a PPI network) for the flow of information between different points in the network. It is calculated as follows:

$$B(v) = \sum \frac{Sij(v)}{Sij}, \qquad with\ i \neq j, v \neq i\ and\ v \neq j \qquad (1)$$

where s_{ij} is the number of shortest paths between the nodes i and j and $s_{ij}(v)$ is the fraction of those shortest paths passing through node v. High betweenness thus indicates that the respective protein has a 'central' position in the network, and that the perturbation of this protein may significantly affect the flow of information through the network. The average betweenness of the analyzed databases are similar (Table 1), with the exception of MINT, which has a slightly lower value. This was expected for all networks since they have similar structure, observed in their clustering coefficients, average degree and path lengths. The majority of proteins in all databases have medium to high betweenness values (defined here as 4.5 to 10.5 on a natural logarithm scale; see Supplementary Figure 2B), even though the number of interaction partners may be limited for these proteins. This suggests that even proteins with few interaction partners occupy important intermediate positions in a network (Joy, et al., 2005).

Finally, several measures of the overall network structure were compared for each database. The 'diameter' of a network defines the *maximal* distance between the two most distant nodes in the network while the average path length (APL) is the mean distance between all protein pairs in the network. As summarized in Table 1, the diameters and APLs of each network are comparable.

These findings collectively show that the databases have a similar network structure, although primary (MINT, INTACT, HPRD) and the derived database (HIPPIE) have a considerable difference in the number of interactions.

3.2 Conserved Topological characteristics between databases

After characterizing the basic features of the databases selected for this study, we next assessed their topological characteristics. 'Topology' describes the arrangements in which nodes are connected to each other in a database. Important topological parameters are the number and the identity of interaction partners. Such information is critical for the identification of hubs, which are often targeted for the identification of possible lethal genes (Albert, et al., 2000; Coulomb, et al., 2005; Jeong, et al., 2001), the development of novel drugs (Hase, et al., 2009; Yildirim, et al., 2007), or network disruption (Quayle, et al., 2007).

To this end, we adopted a strategy used for drug target identification and protein essentiality studies in which proteins are grouped into one of three categories based on the number of interactions (Han, et al., 2004; Hase, et al., 2009; Patil and Nakamura, 2006). We ranked the proteins according to their number of interactions and classified them as high-, middle- or low-degree proteins (see Methods). STRING was excluded from this analysis because it comprises not only protein interactions but also other types of non-physical, protein associations derived from pathway databases, in addition to co-expression of genes and genomic neighborhood.

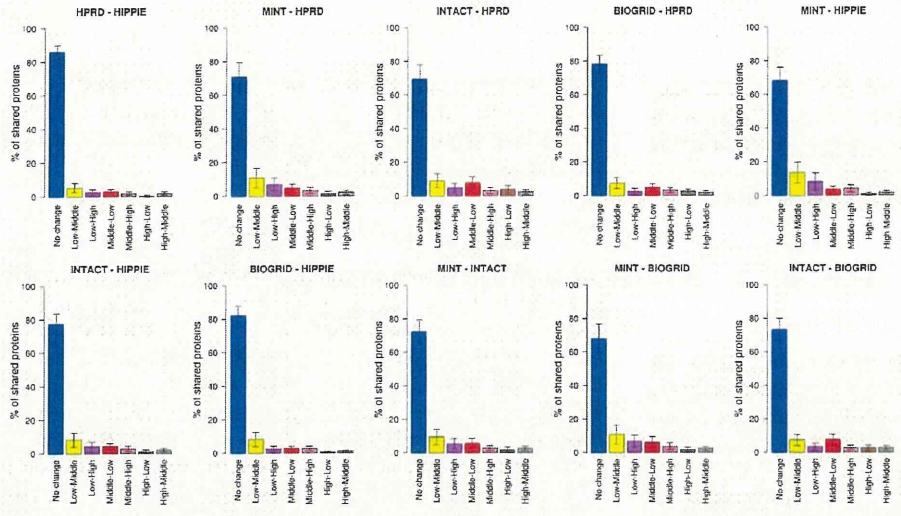


Fig. 1. Proteins were grouped into three categories: low-, middle- and high degree (see Methods). Then, we assessed the percentages of proteins that fall into the same (or different) categories in pair-wise comparisons of two databases. For most comparisons, 60-80% of proteins fall into the same category in both databases compared.

After categorizing all proteins, we assessed the percentages of proteins that fall into the same or different categories in pair-wise database comparisons. Figure 1 shows that 60-80% of the proteins shared between two databases fall into the same category in both databases. This shows that although the databases differ in the number of proteins and interactions, their shared proteins still have similar connectivity levels.

In our pair-wise comparisons, we matched the smaller database (with fewer interactions; e.g. HPRD) against the larger database (with more interactions; e.g. HIPPIE) (Figure 1). As a result, most proteins that fall into different categories between the databases shift into a higher degree category (e.g., the protein shifts from 'low degree' to 'middle degree'). However, we observed that when INTACT is matched against HPRD and BIOGRID, around ~10% of the proteins that are in the 'middle degree' category in the smaller database (i.e. INTACT) shift to the 'low degree' category in the larger database (i.e. HPRD or BIOGRID) (Figure 1). Most likely, this is a consequence of the different experimental datasets used in the different databases and we observed that those proteins show enrichment for translational elongation and RNA processing Gene Ontology categories (p-value < 0.01).

Notably, very few proteins changed between the 'high degree' and 'low degree' categories (or vice versa) when comparing databases (Figure 1). This further supports our notion that the five databases included in this analysis are in fairly good agreement regarding the connectivity of the proteins.

The only exception is the comparison of MINT with HIPPIE and other larger databases, with almost 10% of the proteins falling into the 'low degree' category in MINT, but into the 'high degree' category in HIPPIE. We attribute this finding to the different sizes of databases, with MINT and HIPPIE representing the smallest and largest datasets analyzed (both in terms of numbers of proteins and interactions, Table 1).

3.3 Neighborhood characteristics of datasets

The topological characteristics of a protein in a database are not only defined by the *number* of interaction partners, but, perhaps even more importantly, by the *identity* of interaction partners. We therefore assessed whether proteins have similar or different interaction partners in the databases analyzed. For our analysis, we focused on the 'shared' proteins, i.e., those listed in the two databases being compared. For these proteins, we identified their interaction partners in each of the databases, and then compared the interaction partners between the databases (Figure 2; see also Supplementary Figure 3 for the absolute numbers).

As expected, the highest percentage of shared neighbors was detected for the comparison of 'derived resources (STRING and HIPPIE) to 'primary resources (MINT, BIOGRID and INTACT). However, for comparisons that do not involve the HIPPIE database, no more than 40% of interaction partners are shared. As described earlier, STRING

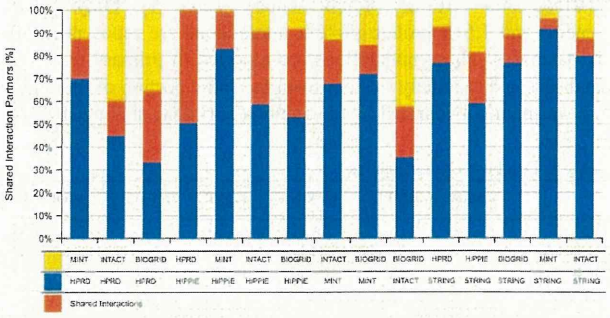


Fig. 2. Shared and exclusive interaction partners in a pair-wise comparison of PPI databases. For proteins shared between two databases, we identified their interaction partners in each of the databases, and then compared the interaction partners. Yellow and blue represent the indicated databases. Shown in red are the interaction partners predicted in both databases.

comprises not only protein interactions but also other functional associations originating, for example, from pathway databases (Jensen, et al., 2009; von Mering, et al., 2005). This results in a large number of interactions that are not covered by the other databases and transforms the interactions of the other databases into a subset of those reported by STRING.

Collectively, our analysis revealed considerable differences in predicted interaction partners between the databases. These differences likely stem from differences in the size of databases, algorithms used, and differences in the portion of the literature used by primary database curators. Researchers should take these issues into account when attempting to identify critical interaction partners of their protein(s) of interest.

3.4 Quality of interactions of key protein sets

STRING, HIPPIE and MINT assign quality scores to each interaction and this is used to assess the confidence level of an analysis; HIPPIE and MINT calculate the confidence score based on accumulated experimental evidence of protein interactions (M.S. et al., submitted)(Ceol, et al., 2010). This stringent approach leads to scores below 0.5 for more than 75% of the interactions reported in these databases (Figure 3A). STRING calculates its confidence score based on the likelihood that two proteins have a functional association that is as specific as the association between an average pair of proteins present in the same KEGG pathway (Kanehisa, et al., 2010; Szklarczyk, et al., 2011). In addition, higher scores are assigned to associations supported by several sources of evidence. Consequently, intensively-studied interactions are more likely to be supported by higher confidence scores. Indeed, we find that more than 80% of the STRING interactions have scores above the acceptable cut-off of 400 (defined by the authors in the program web-site).

Next, we asked whether heavily studied proteins are correspondingly covered by good quality interactions in the PPI databases. To address this question, we selected the 10% most popular human genes/proteins from the literature (that is, 2921 genes/proteins), and ranked them by popularity based on the number of PubMed entries mentioning these genes (Supplementary Table 2); of those, 2,790 were present in HIPPIE, 2,460 in STRING and 1,653 in MINT database.

We performed pair-wise comparisons of the confidence levels of the interactions shared between databases, and that involve the 10% most intensively studied proteins (Figure 3B; Supplementary Table 3). We observed a lack of agreement between the scores calculated in the databases, i.e. several interactions reported as high confidence in one database are reported as low confidence interactions in the other. In the comparison between STRING and HIPPIE, ~70% of the interactions involving the 10% most studied proteins have a high confidence score in STRING but low confidence score in HIPPIE. On the other hand, we observed that 14% of shared interactions had a score above cut-off in both databases. An example is the interaction between TP53 and HMGB1 (Jayaraman, et al., 1998), with a score of 0.83 in HIPPIE and 932 in STRING.

As mentioned before, STRING and HIPPIE are derived databases, thus several interactions shared between them were originally reported in MINT. However, each database assign different scores to those interactions, resulting in no correspondence between the scores of different databases. Therefore, to search for tendencies or biases of each scoring scheme, we considered interactions involving at least one popular protein and with conflicting scores between the databases. With these interactions, we created four groups with distinct characteristics (Table 2) and evaluated a sample of 100 interactions (25 from each group), by manually searching experimental evidence supporting these interactions in the scientific literature (Supplementary Table 4).

We observed that a protein association had high confidence score only in STRING (and low scores in the other two databases), the experimental evidence supporting an association could not be readily identified, reflecting that the scoring scheme used by

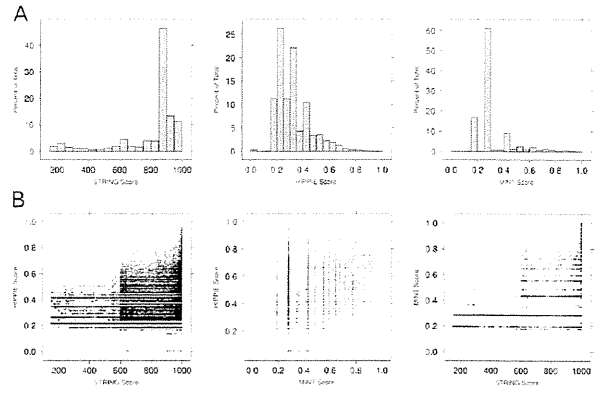


Fig. 3. (A) Three databases assign quality scores for protein interactions (HIPPIE, MINT) or functional associations (STRING). MINT and HIPPIE have a stringent quality score based on cumulative evidence from multiple sources and therefore the majority of its interactions are below 0.5. STRING on the other hand assigns a high score for proteins that are reported in pathway databases (Szklarczyk, et al., 2011). (B) Confidence scores of interactions that involve intensively studied proteins. We observed that in general there is no agreement between the database scores, with the exception that among the 31,229 interactions shared between STRING and HIPPIE, 4,539 have high confidence score in both databases. In addition, in both comparisons involving STRING, no proteins had low confidence score in MINT or HIPPIE and high confidence score in STRING.

STRING - assigning a high score to proteins belonging to the same pathway - may be difficult to validate. In contrast, when either MINT or HIPPIE assigned high scores to an interaction, the supporting evidence could be confirmed in one or more publications; although HIPPIE has a very strict scoring scheme: occasionally more than one publication reported an interaction but it still received a low score. Lastly, as part of the iMEX curation guidelines (Orchard, et al., 2007), the scoring scheme used by MINT was very accurate: interactions with scores greater than 0.5 could be readily confirmed by manuscripts often containing the identity of both interacting partners in its title and specifically investigating that interaction.

Table 2. Groups of interactions

High-Score[1]	Low-Score[2]	Interactions[3]
STRING	HIPPIE	22,177
STRING	HIPPIE and MINT	2,225
STRING and HIPPIE	MINT	448
STRING and MINT	HIPPIE	353

[1]High-scores considered for STRING, MINT and HIPPIE were values greater than 400, 0.5 and 0.5. [2]Low-scores for STRING, MINT and HIPPIE were values lower than or equal to 400, 0.5 and 0.5. [3]All interactions included at least one popular protein.

Summarizing, we observed that although there are differences in the calculations of the quality score, interactions that are highly trustable are those that supported by different experimental systems (especially low-throughput methods), and are manually curated from literature. Ideally, interaction studies should be

carried out in different experimental systems to overcome technique-specific bias (Braun, et al., 2009; Chen, et al., 2010; von Mering, et al., 2002).

3.5 Subnetworks based on organ- and cell type-specific expression data

Protein-protein interaction databases are used to address a wide range of questions that span different organisms, cell types, developmental stages, and/or phases of the cell cycle. To date, no public PPI database takes these issues into account, with the exception of the HPRD team, which in the long-term may also incorporate tissue-specific expression information. Some private companies, e.g. Ingenuity, provide tissue specific network construction, but as they limit the size of the PPI networks to be on the order of hundreds of nodes, these are not the most suitable tools for whole network studies. Here, we assessed how the incorporation of organ- and cell-type-specific expression data influence network analysis.

Using a gene expression dataset of 84 human organs and cell types (Su, et al., 2004; Wu, et al., 2009), we first selected all genes with moderate to high expression levels in each cell type (see Methods). Next, we evaluated the coverage of each database for the proteins expressed from these genes. STRING and HIPPIE cover about 60% of the organ/cell type-specific proteins, whereas the coverage reaches about 40-50% in the other databases (Supplementary Figure 4). It is also interesting to note that all databases have a relatively even coverage of all organs and cell types, although the number of genes expressed varies significantly between the different organs/cell types (Supplementary Figure 5). For example, ten times more genes are expressed in liver and heart as compared to the ovary; yet, the percent coverage in the PPI databases is comparable for these three organs.

To create organ/cell type-specific PPI networks, we then identified in the PPI database interactions for which both partners are expressed in the same organ/cell type (while eliminating interactions between proteins that are expressed in different organs/cell types). Each organ/cell type subnetwork was then built from the resulting dataset and we included 570 housekeeping proteins that are believed to be expressed in all tissues (Eisenberg and Levanon, 2003). As expected, the resulting organ/cell type-specific subnetworks possess significantly fewer interactions than the original PPI databases (between 1-25%) (Supplementary Figure 6). In addition, these subnetworks are considerably more fragmented than the parent networks, resulting in several smaller connected components (Supplementary Figure 7). We observed significant differences between the numbers of interactions for organ/cell type-specific subnetworks, which strongly correlated with the number of genes expressed in the respective organ/cell type (Supplementary Figure 8). For example, more than 6,000 different genes are expressed in BDCA dendritic cells, resulting in a subnetworks that retained 20% of the interactions found in the respective parental PPI databases. By contrast, fewer than 700 genes are expressed in ovary or skin, which reduced the specific subnetworks to just 0.4% of interactions reported in the parental networks (Supplementary Figure 6).

To assess the potential value of organ/cell type-specific subnetworks, we analyzed the interaction of cellular proteins with two medically relevant human viruses, hepatitis C virus (HCV) and human immunodeficiency virus (HIV). First, we obtained a list of 481 human proteins that interact with HCV proteins (de Chassey, et al., 2008) and compared these to the HIPPIE subnetwork created for liver. The HIPPIE database was chosen because it contains a relatively large number of interactions and covers most of the other databases; we focused on the liver subnetwork because of the relevance of this organ in HCV infection (Patrick, 1999).

From the original list of 481 HCV interactors, 98 proteins were present in the liver-specific subnetwork and they interacted with 394 different host proteins (Supplementary Table 5). Comparing the pathway membership of these 492 proteins (interactors and neighbors) with proteins specifically expressed in the liver as a background set, we observed appreciable enrichment in complement and coagulation cascades (p-value: 0.04), apoptosis (p-value: 2.94e-4), Chemokine signaling pathway (p-value: 0.0009) and focal adhesion (p-value: 1.03e-7). By contrast, when we used the complete HIPPIE database, 372 of 481 HCV interactors mapped to the database and were involved in 8,489 interactions with 3,317 different proteins. Using the same analysis that we used for the subnetwork analysis, the HCV interactors and their neighbors fell into many different categories, and no specific pathways or Gene Ontology categories, were significantly enriched, making it very difficult to identify critical pathways for the HCV pathogenesis. Hence, organ/cell type-specific subnetworks may aid in the identification of nodes that are critical in specific biological processes.

As a second example of subnetwork analysis, we studied the interaction of HIV with host cells. From the HIV-1 Human Protein Database (Ptak, et al., 2008), we obtained a dataset of 1,432 host proteins that interact with viral proteins. Next, we created subnetworks containing housekeeping genes and genes expressed in BDCA dendritic cells (DC), CD14+ monocytes, and CD4+ T-cells (all datasets were derived from the HIPPIE database). These datasets were chosen since these cell types play critical roles in HIV infections (Dragic, et al., 1996; McDonald, et al., 2003; Zhu, et al., 2002).

From the original list of 1,432 cellular proteins that interact with HIV proteins, 72 were exclusively found in the DC subnetwork and had 55 neighbors not present in the other two subnetworks. According to the pathway databases, these proteins are present in the systemic lupus erythematosus pathway (p-value: 0.001) and in the B-cell receptor signaling pathway (p-value: 0.01). By contrast, 65 cellular HIV interactors were restricted to the CD14+ monocyte subnetwork (interacting with 31 exclusive neighbors), and showed an enrichment for the apoptosis pathway (p-value: 0.08), Focal Adhesion (p-value: 0.007) and Fc gamma R-mediated phagocytosis (p-value: 0.04). Finally, 58 cellular HIV interactors (and 39 neighbors) were only detected in the CD4+ T-cell subnetwork, with an enrichment for T-cell receptor signaling (p-value: 6.8e-5) and primary immunodeficiency pathway (p-value: 0.05). These analyses demonstrate cell-type-specific interactions between HIV and cellular proteins that may be critical for the infection process. The complete list of cell-specific HIV interactors and neighbors is available in Supplementary Table 6 .

4 DISCUSSION

In this study, we compared six widely used public PPI databases for their basic characteristics, their neighborhood features, and

their overlap with the other databases analyzed. In addition, we demonstrated that predictions could be significantly improved by the analysis of cell/tissue specific subnetworks, and by obtaining additional experimental verification for the interaction partners of the most intensively studied genes from literature.

The six databases compared here have different levels of coverage, in regard to both the number of proteins and the number of protein-protein interactions. Nonetheless, they assign similar topological positions to particular proteins within the network; hence, proteins with few or many interaction partners in one database are likely to have few or many interaction partners in the other databases analyzed. However, the identity of these interaction partners may differ between the databases, resulting in great uncertainty in model building. These differences reflect the differences in the algorithms, portion of literature curated by the different groups (Turinsky, et al., 2010), and the experimental techniques used to build the databases.

Many protein-protein interaction datasets are generated by expressing the two proteins of interest in one cell (for example, in the yeast two-hybrid system). In such in vitro assays, proteins may be co-expressed and interact, but in reality their expression may be dependent on cell type, different experimental stages, and/or during different phases of the cell cycle/organism development. As a result, the currently available PPI databases are believed to contain a significant percentage of false-positive entries (Deane, et al., 2002). To address this weakness, PPI databases could be combined with the increasing number of transcriptomics or proteomics datasets that assess the expression of genes or proteins in a specific organ, cell type, developmental or cell cycle stage. We here provide two examples that demonstrate the potential of this approach.

In one example, we show that the host cellular interaction partners of HCV proteins are not enriched for particular gene ontology categories or pathways in an analysis based on the entire HIPPIE database; in contrast, three KEGG pathways (apoptosis, focal adhesion, complement and coagulation cascades) are highly enriched when the HIPPIE database was analyzed in combination with a liver-specific gene expression dataset. Regulation of apoptosis may play a critical role in HCV infection to establish chronic or persistent infections (Bantel and Schulze-Osthoff, 2003). Activation of the complement and coagulation pathways has been described for HCV infections (Ueda, et al., 1993), and it was verified that hepatic inflammation can be reduced by administering CD55, a regulator of the complement pathway (Chang, et al., 2009). However, the significance of proteins involved in focal adhesion for HCV infections is currently not known, which may be addressed in further investigations. This example demonstrates how the generation of subnetworks may help in the prioritization of pathways for future studies.

In the second example, we show that each cell type subnetwork has exclusive proteins that interact with HIV. Among the exclusive proteins from each cell type are some representing critical processes studied and validated experimentally. Apoptosis induced by HIV proteins was reported to be a critical aspect of its pathogenicity (Castedo, et al., 2002; Rasola, et al., 2001; Zheng, et al., 2007). Cases of patients with concomitant systemic lupus erythematosus and HIV have been reported (Calza, et al., 2003; Gould and Tikly, 2004) and the interplay between autoimmune diseases and retroviruses is an active topic of research (Balada, et al., 2010). In addition, it was observed the association between HIV-infection and the down regulation of Fc-gammaR-mediated phagocytosis in HIV infected macrophages (Kedzierska, et al., 2002).

Some studies have generated subnetworks to address medical questions. In one example, subnetworks from normal and cancer cells have been established to identify protein-protein interactions that are characteristic of cancer development and could be targeted to 'rewire' these cells (Quayle, et al., 2007). In the context of a metabolic study, the creation of tissue-specific subnetworks helped to elucidate post-transcriptional regulation of genes from 10 different tissues that are involved in metabolic diseases (Shlomi, et al., 2008). Collectively, these and our own analyses demonstrate that cell/tissue specific subnetworks can be used to increment the biological relevance of PPI datasets.

Our analysis also revealed that current databases possess many interactions that are characterized by low confidence scores, a finding that is of particular concern for intensively studied proteins. While it is not feasible to verify all predicted interactions with different techniques, we suggest here focusing PPI evaluation efforts on the verification of low-confidence interactions of selected proteins widely used in research models but lacking high-confidence interactions. Towards this goal, we created a priority list of interactions that include highly investigated proteins such as TP53 (described earlier), MAPK1 (mitogen-activated protein kinase 1), BCL2 (B-cell CLL/lymphoma 2), or TNF (tumor necrosis factor F), among many others. Additional experimental data confirming or revealing new interactions of these 'key players' with their predicted cellular interaction partners will push PPI databases a step closer to becoming a reliable, daily-use tool for researchers, in the same way sequence analysis and protein structure databases already are.

ACKNOWLEDGEMENTS

We are grateful to the skillful and critical reading of the manuscript by Amie Eisfeld-Fenney and fruitful discussions with Takeshi Hase. We also acknowledge the Japanese Science and Technology Agency (JST), project ERATO Kawaoka.

REFERENCES

Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks, *Nature*, **406**, 378-382.

Aranda, B., *et al.* (2010) The IntAct molecular interaction database in 2010, *Nucleic Acids Res*, **38**, D525-531.

Balada, E., Vilardell-Tarrés, M. and Ordi-Ros, J. (2010) Implication of Human Endogenous Retroviruses in the Development of Autoimmune Diseases, *International Reviews of Immunology*, **29**, 351-370.

Bantel, H. and Schulze-Osthoff, K. (2003) Apoptosis in hepatitis C virus infection, *Cell Death Differ*, **10**, S48-S58.

Bhattacharya, B., *et al.* (2004) Gene expression in human embryonic stem cell lines: unique molecular signature, *Blood*, **103**, 2956-2964.

Braaksma, M., *et al.* (2011) Metabolomics as a tool for target identification in strain improvement: the influence of phenotype definition, *Microbiology*, **157**, 147-159.

Braun, P., *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions, *Nat Methods*, **6**, 91-97.

Breitkreutz, B.J., *et al.* (2008) The BioGRID Interaction Database: 2008 update, *Nucleic Acids Res*, **36**, D637-640.

Calza, L., *et al.* (2003) Systemic and discoid lupus erythematosus in HIV-infected patients treated with highly active antiretroviral therapy, *Int J STD AIDS*, **14**, 356-359.

Castedo, M., *et al.* (2002) Sequential involvement of Cdk1, mTOR and p53 in apoptosis induced by the HIV-1 envelope, *EMBO J*, **21**, 4070-4080.

Ceol, A., *et al.* (2010) MINT, the molecular interaction database: 2009 update, *Nucleic Acids Res*, **38**, D532-539.

Chang, M.-L., *et al.* (2009) Hepatic inflammation mediated by hepatitis C virus core protein is ameliorated by blocking complement activation, *BMC Medical Genomics*, **2**, 51.

Chen, Y.C., *et al.* (2010) Exhaustive benchmarking of the yeast two-hybrid system, *Nat Methods*, **7**, 667-668; author reply 668.

Coulomb, S., *et al.* (2005) Gene essentiality and the topology of protein interaction networks, *Proc Biol Sci*, **272**, 1721-1725.

de Chassey, B., *et al.* (2008) Hepatitis C virus infection protein network, *Mol Syst Biol*, **4**, 230.

Deane, C.M., *et al.* (2002) Protein Interactions, *Molecular & Cellular Proteomics*, **1**, 349-356.

Dragic, T., *et al.* (1996) HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5, *Nature*, **381**, 667-673.

Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact, *Trends Genet*, **19**, 362-365.

Ewing, R.M., *et al.* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry, *Mol Syst Biol*, **3**, 89.

Gentleman, R., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology*, **5**, R80.

Gould, T. and Tikly, M. (2004) Systemic lupus erythematosus in a patient with human immunodeficiency virus infection – challenges in diagnosis and management, *Clinical Rheumatology*, **23**, 166-169.

Han, J.D., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, **430**, 88-93.

Hase, T., *et al.* (2009) Structure of protein interaction networks and their implications on drug design, *PLoS Comput Biol*, **5**, e1000550.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protocols*, **4**, 44-57.

Jayaraman, L., *et al.* (1998) High mobility group protein-1 (HMG-1) is a unique activator of p53, *Genes & Development*, **12**, 462-472.

Jensen, L.J., *et al.* (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms, *Nucleic Acids Res*, **37**, D412-416.

Jeong, H., *et al.* (2001) Lethality and centrality in protein networks, *Nature*, **411**, 41-42.

Joy, M.P., *et al.* (2005) High-betweenness proteins in the yeast protein interaction network, *J Biomed Biotechnol*, **2005**, 96-103.

Kamburov, A., *et al.* (2011) ConsensusPathDB: toward a more complete picture of cell biology, *Nucleic Acids Research*, **39**, D712-D717.

Kanehisa, M., *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Research*, **38**, D355-D360.

Kedzierska, K., *et al.* (2002) HIV-1 Down-Modulates γ Signaling Chain of FcγR in Human Macrophages: A Possible Mechanism for Inhibition of Phagocytosis, *The Journal of Immunology*, **168**, 2895-2903.

Krogan, N.J., *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature*, **440**, 637-643.

Mathivanan, S., *et al.* (2006) An evaluation of human protein-protein interaction data in the public domain, *BMC Bioinformatics*, **7 Suppl 5**, S19.

McDonald, D., *et al.* (2003) Recruitment of HIV and Its Receptors to Dendritic Cell-T Cell Junctions, *Science*, **300**, 1295-1297.

Orchard, S., *et al.* (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition, *Proteomics*, **7 Suppl 1**, 28-34.

Patil, A. and Nakamura, H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks, *FEBS Lett*, **580**, 2041-2045.

Patrick, M. (1999) Hepatitis C: the clinical spectrum of the disease, *Journal of hepatology*, **31**, 9-16.

Prasad, T.S., Kandasamy, K. and Pandey, A. (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology, *Methods Mol Biol*, **577**, 67-79.

Ptak, R.G., *et al.* (2008) Cataloguing the HIV type 1 human protein interaction network, *AIDS Res Hum Retroviruses*, **24**, 1497-1502.

Quayle, A.P., Siddiqui, A.S. and Jones, S.J. (2007) Perturbation of interaction networks for application to cancer therapy, *Cancer Inform*, **5**, 45-65.

Ramirez, F., *et al.* (2007) Computational analysis of human protein interaction networks, *Proteomics*, **7**, 2541-2552.

Rasola, A., *et al.* (2001) Apoptosis enhancement by the HIV-1 Nef protein, *J Immunol*, **166**, 81-88.

Rual, J.F., *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, **437**, 1173-1178.

Shlomi, T., *et al.* (2008) Network-based prediction of human tissue-specific metabolism, *Nat Biotechnol*, **26**, 1003-1010.

Stumpf, M.P.H., *et al.* (2008) Estimating the size of the human interactome, *Proceedings of the National Academy of Sciences*, **105**, 6959-6964.

Su, A.I., *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A*, **101**, 6062-6067.

Szklarczyk, D., *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, *Nucleic Acids Res*, **39**, D561-568.

Turinsky, A.L., *et al.* (2010) Literature curation of protein interactions: measuring agreement across major public databases, *Database*, **2010**.

Ueda, K., *et al.* (1993) The association between hepatitis C virus infection and in vitro activation of the complement system, *Ann Clin Biochem*, **30 (Pt 6)**, 565-569.

Venkatesan, K., *et al.* (2009) An empirical framework for binary interactome mapping, *Nat Meth*, **6**, 83-90.

von Mering, C., *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucleic Acids Res*, **33**, D433-437.

von Mering, C., *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

Wilhelm, B.T., *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, *Nature*, **453**, 1239-1243.

Wu, C., *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources, *Genome Biol*, **10**, R130.

Yildirim, M.A., *et al.* (2007) Drug-target network, *Nat Biotechnol*, **25**, 1119-1126.

Zheng, L., *et al.* (2007) HIV Tat protein increases Bcl-2 expression in monocytes which inhibits monocyte apoptosis induced by tumor necrosis factor-alpha-related apoptosis-induced ligand, *Intervirology*, **50**, 224-228.

Zhu, T., *et al.* (2002) Evidence for Human Immunodeficiency Virus Type 1 Replication In Vivo in CD14+ Monocytes and Its Potential Role as a Source of Virus in Patients on Highly Active Antiretroviral Therapy, *J. Virol.*, **76**, 707-716.

STUDY DESIGNS

Software for systems biology: from tools to integrated platforms

Samik Ghosh, Yukiko Matsuoka*‡, Yoshiyuki Asai§, Kun-Yi Hsin§ and Hiroaki Kitano*§‖*

Abstract | Understanding complex biological systems requires extensive support from software tools. Such tools are needed at each step of a systems biology computational workflow, which typically consists of data handling, network inference, deep curation, dynamical simulation and model analysis. In addition, there are now efforts to develop integrated software platforms, so that tools that are used at different stages of the workflow and by different researchers can easily be used together. This Review describes the types of software tools that are required at different stages of systems biology research and the current options that are available for systems biology researchers. We also discuss the challenges and prospects for modelling the effects of genetic changes on physiology and the concept of an integrated platform.

**The Systems Biology Institute, 5F Falcon Building, 5-6-9 Shirokanedai, Minato, Tokyo 108-0071, Japan.*
‡JST ERATO Kawaoka Infection-induced Host Response Project, 4-6-1 Shirokanedai, Minato, Tokyo 108-8639, Japan.
§Okinawa Institute of Science and Technology, 1919-1, Tancha, Onna-son, Kunigami, Okinawa 904-0412, Japan.
‖Sony Computer Science Laboratories, Inc., 3-14-13 Higashi-Gotanda, Shinagawa, Tokyo 141-0022, Japan.
Correspondence to S.G. and H.K.
e-mails: ghosh@sbi.jp; kitano@sbi.jp
doi:10.1038/nrg3096
Published online 3 November 2011

Systems biology emerged in the mid-1990s with the aim of achieving a system-level understanding of living organisms and applying this knowledge in various fields, including medicine and biotechnology[1–4]. Early applications included modelling cell cycle dynamics[5–7], such as a computational model that explained the effects of over 120 knockout mutations on cell cycle dynamics in yeast[7]. Significant progress has also been made in the analysis of signalling pathways — for example, in understanding the dynamics of mitogen-activated protein kinase (MAPK) signalling[8] — and in cancer drug discovery applications, in which a reagent that was developed using model-based computational analysis is now in clinical trials[9,10].

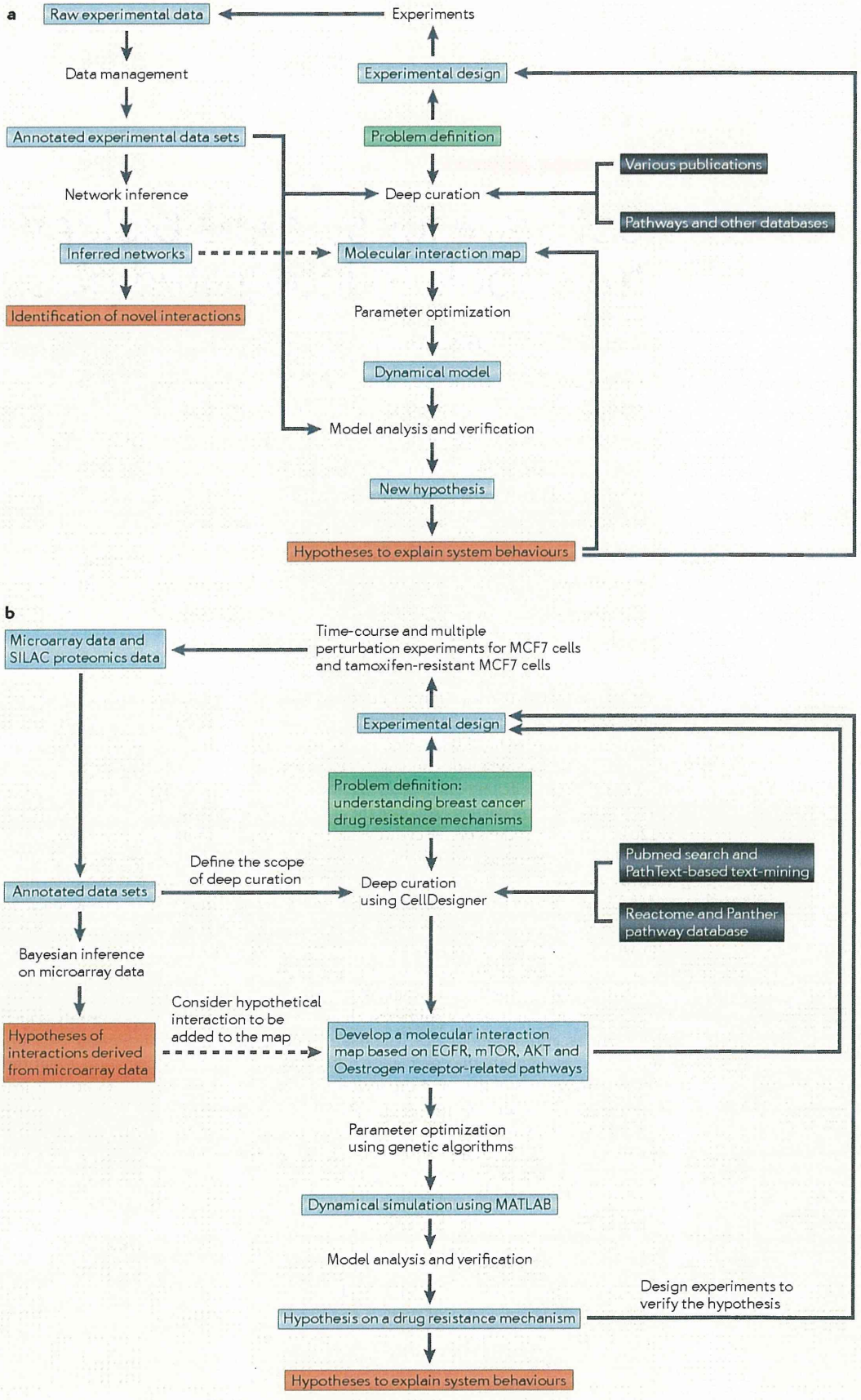
System-level studies are often built on molecular and genetic findings and 'omics' studies, such as genomics, proteomics, and metabolomics. The main challenges in systems biology are the complexity of the systems, the vast quantities of data and the scattered pieces of knowledge; these all have to be integrated; therefore, systematic, computational tools are crucially important in systems biology. Software platforms have transformed industries — such as aviation, entertainment and electronics — by drastically improving productivity and by offering new capabilities[11]. Biological sciences are no different. In particular, the success of systems biology, and its application in areas such as systems drug design, requires sophisticated data handling, modelling, integrated computational analysis and knowledge integration. For example, the creation of computational models enables us to predict the behaviours of biological systems, thereby helping us to understand the underlying molecular mechanisms and to predict the impact of perturbations, such as drug treatments, on these biological systems.

Software tools and resources for systems biology need to be tailored to their intended applications in order to achieve the objectives of novel biological discoveries, drug design and answers to life-science research questions. A typical workflow for computational analysis is a cyclical process involving data acquisition, modelling and analysis. Prediction and explanation capabilities are associated with this cycle, and the integration and sharing of knowledge help to sustain these capabilities (FIG. 1).

Here we describe the principles of each stage in this workflow and some examples of current tools. Links to the tools and resources mentioned in this Review are provided in Supplementary information S1 (table), along with information about their type and access policy. TABLE 1 provides a matrix to help users choose appropriate tools and resources. We provide a perspective on the current challenges facing systems biology software tools, and we describe our view that integrated software platforms will help to address future research problems in biology and medicine.

Data management

The proper acquisition and handling of data is crucially important for both the generation and verification of hypotheses. The rapid development of high-throughput experimental techniques is transforming life-science research into 'big data' science[12], and although numerous data-management systems exist[13–16], the heterogeneity of


a
Raw experimental data
Experiments
Data management
Experimental design
Annotated experimental data sets
Problem definition
Network inference
Deep curation
Various publications
Pathways and other databases
Inferred networks
Molecular interaction map
Identification of novel interactions
Parameter optimization
Dynamical model
Model analysis and verification
New hypothesis
Hypotheses to explain system behaviours
b
Microarray data and SILAC proteomics data
Time-course and multiple perturbation experiments for MCF7 cells and tamoxifen-resistant MCF7 cells
Experimental design
Problem definition: understanding breast cancer drug resistance mechanisms
Define the scope of deep curation
Annotated data sets
Deep curation using CellDesigner
Pubmed search and PathText-based text-mining
Reactome and Panther pathway database
Bayesian inference on microarray data
Consider hypothetical interaction to be added to the map
Hypotheses of interactions derived from microarray data
Develop a molecular interaction map based on EGFR, mTOR, AKT and Oestrogen receptor-related pathways
Parameter optimization using genetic algorithms
Dynamical simulation using MATLAB
Model analysis and verification
Design experiments to verify the hypothesis
Hypothesis on a drug resistance mechanism
Hypotheses to explain system behaviours

◀ Figure 1 | **Workflow of computational tasks in systems biology.** A research cycle showing the computational modelling and analyses that are involved in the workflow. **a** | The workflow starts from the 'problem definition' of the research project (shown in the green box). One stream of the workflow starts with experimental design, followed by the execution of experiments, data management and network inference. A parallel stream of the workflow consists of deep curation, parameter optimization, dynamical model analysis and model verification using experimental data. Outputs are shown in red boxes. Discrepancies between simulation results from the computational model and experimental data indicates that some of the underlying hypotheses need to be modified; the simulation should then be tested again when these new hypotheses are incorporated into the model. Transformation of a network that is inferred from large-scale data into a precise, mechanism-based model is an important step. However, this step is not yet fully achievable in practice, as indicated by the dotted arrow in the figure. **b** | An example biological application of the workflow from part **a**; in this case, research aiming to understand mechanisms of drug resistance in breast cancer. After the definition of the problem, time-series, multiple perturbation experiments would be designed, followed by data annotation, data analysis and network inference. Results from the data analysis would be used to define the scope of deep curation. However, in some cases, a molecular interaction map would be created before the experiment is designed, so that the experiments could be designed based on existing knowledge. When moving from the molecular interaction map to dynamical simulation, often only a part of the deep-curation-based molecular interaction map would be used for dynamical modelling, by which possible hypotheses for drug resistance mechanisms could be generated. This is an iterative process involving both 'dry' and 'wet' research. EGFR, epidermal growth factor receptor; mTOR, mammalian target of rapamycin; SILAC, stable isotope labelling with amino acids in cell culture.

formats, identifiers and data schema pose serious challenges. In this context, data-management systems need standardized formats for data exchange, globally unique identifiers for data mapping[17] and common interfaces that allow the integration of disparate software tools in a computational workflow.

Data-management standards. The development of data representation and communication standards for systems biology and bioinformatics has become a distinct field of work[18]. Standards for data management have focused on three core aspects: minimum information, file formats and ontologies.

Minimum information is a checklist of required supporting information for data sets from different experiments. Examples include: Minimum Information About a Microarray Experiment (MIAME)[19], Minimum Information About a Proteomic Experiment (MIAPE)[20,21] and the Minimum Information for Biological and Biomedical Investigation (MIBBI) project[22]. An important element of these standardization efforts is the incorporation of metadata (that is, data about data), which has led to the definition of standards such as the International Organization for Standardization metadata registry (ISO–MDR) standard and the Dublin Core Metadata Initiative (DCMI) standard. Standards for file formats define how the minimum information should be stored. These formats are generally Extensible Markup Language (XML)-based, which facilitates automatic processing by computers. Organizations that have defined standards include the Microarray Gene Expression Data (MGED) Society, the Proteomics Standards Initiative (PSI) and the Metabolomics Standards Initiative (MSI).

Ontologies define the relationships and hierarchy between different terms and allow the unique, semantic annotation of data. Various specialized ontologies for biology are in development; for example, the Gene Ontology (GO) and the Systems Biology Ontology (SBO) (see Supplementary Information S1 (table) for a comprehensive list of biomedical ontologies).

Data-management and data-analysis tools. Current data-management systems can be broadly classified as spreadsheet-based or Web-based, or as laboratory information management systems (LIMS). Spreadsheet programs have historically been the most popular mode of data storage and communication in the life-science community, owing mainly to the ease of use and sharing; for example, template-based spreadsheets like MAGE-TAB (a spreadsheet-based, MIAME-supportive format for microarray data) and the Investigation–Study–Assay (ISA)-TAB formats. However, their integration with analysis tools and computational workflows requires custom-built interfaces that are not supported on all software platforms. In addition, a standardized practice for filling the spreadsheet is required.

More recently, online wiki-based document and project management has become a popular mode of exchange for different laboratories, and these formats now provide security and privacy options for data protection. Other alternatives are custom-built information systems for laboratory data storage and management, such as electronic lab notebooks (ELN). These are routinely deployed in large research laboratories. While providing various features and functionalities, they are usually associated with steep learning curves for users, which, together with the cost of deployment, creates a substantial barrier to the adoption of these systems across the scientific community.

A different option, which integrates data management and analysis, is the use of workflow-management systems (WMSs). These systems harness the power of the Web to integrate different tools and services in a computational pipeline. Systems like Konstanz Information Miner (KNIME), caGrid[23], Taverna[24], Bio-STEER[25] and Galaxy[26], allow the construction, execution and sharing of specialized workflows. A comprehensive catalogue of biological Web services is available at BioCatalogue. WMSs provide the first step in building a computational pipeline by enabling data exchange, data integration and inter-tool communication. However, most current systems are tailored for specific research workflows (for example, KNIME for bioinformatics tools and Galaxy for genomic data analysis), and they support only specific sets of tools and standards; this forces researchers to use several different WMSs for a holistic understanding of their biological system of interest.

There are emerging efforts that focus on data management, such as Sage Bionetworks and ELIXIR. Sage Bionetworks is currently focused on establishing a platform for data acquisition and curation. The future aim of this platform is for modelling, using an open collaborative approach for gathering expression profile and protein interaction data, with the specific aim of using these data for drug discovery. ELIXIR is a European effort that plans to build a biological data-management infrastructure.

Table 1 | **A resource matrix of software tools and data resources**

	Tools		Standards			Projects
	Software	**Resources**	**Ontologies**	**File format**	**Minimum information**	
Data and knowledge management	MAGE-TAB, ISA-TAB KNIME, caGrid, Taverna, Bio-STEER	BioCatalogue	SBO, OBO, NCBO	MGED (MAGE), PSI, MSI	MIAME, MIAPE, MIBBI, ISO MDR, DCMI	
Data-driven network inference	R, MATLAB, BANJO					DREAM Initiative, Sage Bionetworks
Deep curation	CellDesigner, EPE, Jdesigner, PathVISIO	KEGG, Reactome, Panther pathway database, BioModels.net, WikiPathways		SBML, SBGN, CellML, BioPAX, PSI-MI	MIRIAM	
***In silico* simulation**	COPASI, SBW, JSim, Neuron, GENESIS, MATLAB, ANSYS, FreeFEM, ePNK, ina, WoPeD, Petri nets, OpenCell, CellDesigner + COPASI, CellDesigner + SOSlib, PhysioDesigner (formerly *insilico*IDE)			SED-ML, SBRML, PNML, SBML	MIASE	
Model analysis	MATLAB, Auto, XPPAut, BUNKI, ManLab, ByoDyn, SenSB, COBRA, MetNetMaker, DBSolve Optimum, Kintecus, NetBuilder, BooleanNet, SimBoolNet					
Physiological modelling	JSim, PhysioDesigner (formerly *insilico*IDE), CellDesigner (cellular modelling), FLAME, OpenCell, Virtual Physiology (produced by cLabs), GENESIS, Neuron, Heart Simulator, AnyBody			CellML, SBML, NeuroML, MML		IUPS Physiome Project, Virtual Physiological Human, High-Definition Physiology
Molecular interaction modelling	AutoDock Vina, GOLD, eHiTS	RCSB PDB, ZINC, PubChem, PDBbind				

This table summarizes the tools and resources that correspond to each step in a systems biology workflow; please refer to FIG. 1 for an overview of the workflow and to Supplementary information S1 (table) for additional information and Weblinks to these resources.

Data-driven network inference

A specific kind of modelling from large-scale data, known as data-driven network-based modelling, has been developed over the last decade[27]. Data-driven network-based modelling approaches use computational algorithms to infer causal relationships among molecular entities (such as genes, transcription factors, proteins and metabolites) from high-throughput and time-course experimental data that has been collected under various perturbations. The models that result from this kind of modelling from large-scale data sets are variously known as inference networks, co-expression networks or association networks. Early studies focused on finding patterns in gene expression profiles to distinguish disease states from healthy states; for example, in breast cancer prognosis[28]. Further studies have integrated multi-dimensional data — including genome-scale DNA variation data[29–31], gene expression data[32–34], protein–protein interaction data, DNA–protein binding data and complex binding data — to construct probabilistic, causal gene networks[35–37]. The advent of next-generation sequencing technologies provides new opportunities to incorporate the knowledge of splicing variation and SNPs into network inference models.

Approaches to network inference models. Network inference models have been predominantly based on Bayesian inference techniques; that is, computing the probability of a hypothesis (in this case, the relationship between two molecular entities) based on some kind of evidence or observations (known as priors). However, several alternative techniques have also been applied[38–45], including regression, correlation methods and mutual information approaches. Mutual information approaches compute the relationship between two genes or proteins based on mutual information (a quantity that measures the mutual dependence of two variables) to infer statistically significant associations between these variables[38,39].

The current focus of the research community is on the development of novel algorithms and techniques for reconstructing molecular interaction networks from large-scale experimental data sets. In this regard, standard tools and exchange formats are not yet well established, and most research groups develop their own implementation of network reconstruction algorithms. Common software tools for implementing network reconstruction algorithms include R, MATLAB and BANJO.

Mutual information
A dimensionless quantity that measures the extent to which one random variable is informative about another variable. Zero mutual information between two random variables means that they are independent.

Standards in data-driven inference. One of the key challenges in network inference techniques is the problem of underdetermination[46], in which the number of possible inferred interactions far exceeds the number of independent measurements. The number of experiments and the systematic selection of perturbations and time points play an important part in the reliability of inferred networks. Also, there are no true benchmarking standards for biological data and networks, and most techniques currently have their accuracy evaluated using simulated data, which do not always capture the reality in biological systems. Recent efforts towards community-driven standardization and systematic, rigorous assessment have been initiated through Sage Bionetworks (see above), and the Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative. The DREAM project attempts to evaluate and benchmark different algorithms that influence network inference. Analysis of DREAM results (from the DREAM2 and DREAM3 challenges) reveal that algorithms complement each other in a highly context-specific manner, and that a community-based, consensus-driven reverse-engineering approach can lead to high-quality network inference[46]. One of the explanations for why such a community-based approach performs better than the best algorithm in a pool of algorithms is the compensatory effects from multiple algorithms on the strength and weaknesses of each individual algorithm. This is an interesting observation and it is consistent with the proposed explanation for why IBM's DeepQA system (an open-domain, automatic question-answering computing system) was successful in a 'Jeopardy!' challenge[47], based on a US quiz show that requires participants to have a wide range of topical knowledge and to interpret nuances in subtle clues that are provided to them.

Deep curation

An alternative to data-driven network inference is the deep curation approach. The deep curation approach creates a detailed molecular interaction map by the large-scale integration of knowledge, such as information from publications, databases and high-throughput data[48–51]. Unlike the data-driven approach, in which hypotheses about interactions are generated automatically, the deep curation approach constructs the model manually or semi-manually, thus making it easier for researchers to add their own hypotheses into it. Users can explicitly add unknown interactions to deep curation pathways as 'hypotheses', but it would be helpful if these interactions were made distinct from the evidence-based interactions and if they also included a rationale to support the hypothesis. Although the data-driven approach, depending on observed data, might generate networks that represent inferred causality or the association of behaviours at the transcriptional or protein–protein interaction level, they do not provide mechanistic details nor confirm causality. By contrast, the deep curation approach can provide mechanistic details of each interaction because curators will look into the details of the reported molecular mechanisms and experiments in the literature and will read them critically. Precise and in-depth mechanistic-level models are essential not only for precise computer simulations and an understanding of biological mechanisms, but also for the proper evaluation of potential drug targets. In both basic research and drug discovery, a deep curation approach is essential when the priority is to understand the details of molecular mechanisms, rather than to identify novel molecules and novel interactions.

It would be ideal to combine deep curation and data-driven approaches, but this will require further work. For example, some of the interactions that have been inferred by data-driven approaches are likely to be confirmed by deep curation approaches, and some can be clearly rejected. The remaining inferred interactions can be prioritized for further studies, and resources can be focused on these hypotheses.

Resources, standards and software for deep curation. Deep curation requires an open-ended assembly of knowledge from diverse literature and data sources and is tailored for specific purposes. Therefore, if required, the scope of the model can span multiple pathways. A variety of pathway databases — such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[52], Reactome[53], Panther pathway database[54], Pathway Commons[55], BioCyc[56] — provide information that can be used to create an initial draft of the pathway model. There also are meta-databases, such as the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) and ConsensusPathDB, which integrate diverse knowledge resources and provide a broader context for pathway curation.

There are several machine-readable model-representation standards, which have been developed for different purposes; two widely used standards are the Systems Biology Markup Language (SBML)[57] and the Biological Pathways exchange (BioPAX)[58] format, both of which were designed to represent biomolecular networks from different perspectives. The Systems Biology Graphical Notation (SBGN)[59] was designed to standardize a human-readable pathway notation. This notation defines the graphical representation of networks so that users can interpret the diagrams consistently. Minimum Information Required in the Annotation of Models (MIRIAM)[60] defines the rules for model annotation. Workshops, such as the Computational Modelling in Biology Network (COMBINE) workshop, occur regularly and provide a forum for such standardization efforts. The establishment of standards enables data and models to be re-used across multiple software tools, promotes healthy competition among these tools and helps to build a pipeline of tools for efficient analysis.

Several tools and model databases are currently available to support deep curation efforts. CellDesigner[61] is one of the most widely used software tools[62] — it enables users to visually define a model of biological interactions and to comply with SBML and SBGN. A plug-in application programming interface (API) for CellDesigner enables users to develop various additional functionalities, including the conversion of models to other formats, such as BioPAX. Several other tools

Meta-database
A database for storing metadata, which was originally defined as 'data about data', such as tags and keywords. The database is used for integrating independent distributed databases.

provide graphical editing and visualization capabilities; for example, the Edinburgh Pathway Editor (EPE), JDesigner[63], PathVISIO[64] (which is for pathway curation) and Cytoscape[65] (which is a widely used tool for the visualization of molecular networks).

Challenges of deep curation. The quality of pathways in existing pathway databases is often compromised by fragmentation and inaccuracy because these databases cover a broad range of pathways and hence little time can be spent on curating each pathway. The current 'gold standard' is manually curated maps that have been carefully built by a small group of people who spend months studying a pathway, such that they would be familiar with almost every publication on that pathway[66]. Several such maps have been reported, including for the epidermal growth factor receptor (EGFR) pathway[49], the Toll-like receptor pathway[48], the mammalian target of rapamycin (mTOR) pathway[50], the yeast cell cycle[51] and the E2F pathway[67]. In addition, the community-based reconstruction of metabolic networks for several species has been accomplished through the systematic use of various omics databases and publications[68–71].

Another consideration is that pathways reflect a specific context, such as a tissue, a disease status or a species. Pathway databases do not always identify the tissues in which interactions have been identified, thus the context of interactions should be carefully noted during the curation process. In addition, tissue-specific proteomic and gene-expression data can be used to ascertain which parts of generic pathways actually exist in the tissue of interest. This is an important practice, especially when computational models are used to explain and predict cell-line-based drug-screening experiments[10]. An additional point to consider is that there can be crosstalk among pathways.

One of the main challenges of the deep curation approach is to keep the pathways up-to-date and to validate them. This is particularly important in view of the context-specificity of molecular maps. Several disease-specific maps have been curated — for example, for rheumatoid arthritis[72] and for cardiovascular pathways — but manually creating large-scale network maps from the literature is extremely labour-intensive and requires specific quality-control procedures. Also, it is challenging for curators to maintain the motivation to continuously update the map with new discoveries. There is a need to develop techniques that automate knowledge discovery, the aggregation of pathway components and the addition of context-specific control mechanisms to pathway maps. Automated literature mining has also been investigated extensively, but is not yet close to being ready to replace human curators. Pathway validation requires an expert knowledge of the underlying biology and the ability to transform literature evidence into pathway diagrams. Recruiting experts, assigning them to pathway curation and coordinating their efforts to build integrated pathways is a major challenge.

Another option is collaborative curation, and several approaches are being developed to enable community-driven pathway updates. An example is the Payao[73] system, which has been used to promote pathway development and annotation in large and geographically distributed teams. An alternative is the community-based development and refinement of pathways, as is used in WikiPathways[74]. However, insufficient participation from active users remains a challenge for such approaches, and it is not yet clear how the widespread engagement of the biological community can be incentivized.

In silico simulation models

Molecular interaction maps provide a static picture, but the dynamics of molecular interactions in time and space have a central role in the behaviour of cells and organisms. Dynamical simulations are mostly based on models that have been created by the deep curation approach, rather than by the data-driven approach. This is because deep curation captures causality, stoichiometry and mechanisms of interactions, which are mandatory in dynamical simulations. Here we provide a brief overview for readers who are unfamiliar with the subject; for further details we recommend reading reviews that are focused on simulation and analysis[62,75,76].

Simulations have an important role in the computational verification of biological models and the computational prediction of behaviours. After the initial model is created as a set of hypotheses, dynamical simulations examine whether the model behaves like the real biological system. When some observed behaviours are not reproduced by the model, this indicates that some hypotheses are inaccurate or missing, and alternative hypotheses should be incorporated into the model and verified. Thus, the proper identification of discrepancies between experimental results and model predictions is the key for successful computational research. Dynamical modelling of complex biological systems has been applied with varying degrees of success[10,77]. Ordinary differential equations (ODEs) have been used widely as a standard numerical method in many successful cases of biological modelling[5,6,9,10]. Dynamical models that capture the stochastic (random) behaviour of molecular interactions have successfully elucidated gene transcription and translation processes[78,79] or *Escherichia coli* fate decisions during phage infections[80]. Physiological models of systems also use partial differential equations (PDEs) and a different set of tools (see below). Other techniques, such as agent-based modelling[81], process algebra (for example, the Petri net[82] system) and rule-based modelling[83], have also been applied to study the behaviour of specific biological systems.

Reaction constants and other parameters are required for simulations, and the proper calibration of models remains a major bottleneck for biological systems. Researchers can consider using rate constants that have been measured using biochemical assays, but in many cases these differ from the rate constants within cells and have not been collected in a high-throughput manner. Thus, parameters must be measured *in vivo* or be estimated through parameter-optimization techniques that are supported by various simulation and model-analysis

Ordinary differential equations
(ODEs). A type of differential equation involving functions of one independent variable, such as time, and derivatives of the functions with respect to the variable.

Partial differential equations
A type of differential equation involving functions of several independent variables, such as time and spatial axes (that is, x, y and z), and partial derivatives of the functions with respect to those variables.

Agent-based modelling
A class of computational models that simulate the interaction of agents to study their effects on a system. Agents are autonomous, decision-making entities that have heterogeneous characteristics; examples of agents are molecules or cells.

Process algebra
A mathematical modelling language for describing the behaviour of distributed systems.

Rule-based modelling
When used in biochemical science, this term refers to a way to model molecules and proteins as objects that interact with each other. The interactions are described as rules that define how the objects transform their attributes and the relationships between the objects.

Box 1 | **Parameter optimization: stochastic search methods and gradient descent methods**

There are several methods to estimate parameters for models. The stochastic search approach generates a set of parameters randomly, but often following certain rules to make the search more efficient. Each parameter set is tested in the model to see whether it generates results that are consistent with the experimental results or other defined criteria. The best set is selected and parameter values are generated again, usually close in value to the selected set, to see if there are better parameter sets. Eventually, a parameter set that can be considered optimal will be found.

The gradient descent approach has a defined algorithm that tunes parameters. It depends on error gradients that can be calculated from the difference in error values between two parameter sets. The parameter value is chosen that is estimated to have a smaller error value. Such algorithms can quickly find the optimal parameters for simple problems in which there is only one optimal point and the parameter sets near this optimal point only gradually become suboptimal. However, it may only find a local optimal parameter set for highly nonlinear and multi-peak problems.

tools and reaction databases, such as the System for the Analysis of Biochemical Pathways — Reaction Kinetics (SABIO–RK)[84] database. Sophisticated parameter-estimation algorithms, and data to calibrate them, are essential. Algorithms for optimization include stochastic methods and gradient-descent methods (BOX 1). Nevertheless, there are limitations in the current technologies and resources for creating large-scale dynamical models; it may be more practical to select part of the pathways for precise dynamical modelling, rather than to try to use an entire pathway map that inevitability contains uncertain parameters.

Standards and tools for simulations. Several standardization efforts empower the modelling community. Examples include SBML[57], SBGN[59] and MIRIAM[60] for model representation and annotation. Minimum Information About a Simulation Experiment (MIASE)[85] is used to define the minimum set of information that is required to reproduce numerical simulations, and the Simulation Experiment Description Markup Language (SED-ML) is an XML-based specification for encoding configurations for simulations, for defining models to be used, for setting up numerical calculations and for formatting outputs. In addition, the Systems Biology Results Markup Language (SBRML)[86] is a complementary language to SBML that specifies the format of results of simulations carried out on models.

Based on these standards, a series of simulation tools and software has been developed, with tools such as MATLAB and the Complex Pathway Simulator (COPASI)[87] being widely used for model simulation and analysis. The Systems Biology Workbench (SBW) is a software platform that allows multiple applications — such as software packages for modelling, analysis or visualization — to communicate with each other; this aims to enhance model exchange and simulation efficacy. Several tools support process algebra and Petri net modelling. For example: ePNK, a modelling platform for Petri nets that is based on the Petri net Markup Language (PNML); Time Petri Net Analyser (TINA), a toolbox for the editing and analysis of Petri nets; and WoPeD, a tool for modelling, simulation and analyses of Petri nets that also supports PNML. BioModels.net provides a database portal for curated, validated dynamical models that can be used to kick-start a modelling effort by re-using well-known components.

Phase-space analysis
A way to analyse the dynamics of a system in a space (the phase-space), in which each of the possible states of the system is represented as a unique point.

Bifurcation analysis
A way to analyse the qualitative changes in the dynamics of a system that are caused by varying one or several parameter values continuously.

Model analysis. The next step is to analyse models for insights into the intrinsic and dynamical nature of the system (FIG. 1). A conventional time-course simulation from a defined initial state gives an indication of how the system behaves under a specific condition; more in-depth insight is provided by systematic analyses of the system under different conditions. Different mathematical techniques have been developed to analyse the behaviour of complex biological models and are supported by specific software tools[88,89] (BOX 2).

Many model-analysis techniques focus on dynamical systems that are represented as set of ODEs (BOX 2), but alternative analyses have also been developed that are based on statistical network analysis[82]. In particular, Boolean network modelling of genetic regulatory networks has gained wide acceptance in the modelling community, based on pioneering work by Kauffman[90]. Several Boolean network simulators for biological systems have been developed, including NetBuilder, BooleanNet and SimBoolNet[91]. In addition, a series of tools is available for phase-space analysis and bifurcation analysis, such as XPPaut and BUNKI. We refer readers elsewhere for details of using these analysis approaches[5,76,88,92].

Multi-scale physiological modelling

The next level, in which there is an increasing interest, is to develop physiological models that are linked with underlying molecular networks and genetic polymorphisms. Developing these models is a substantial challenge, but such models should have important applications because genetic polymorphisms and the associated differences in network dynamics can influence many diseases. For example, mutations in the voltage-gated sodium channel SCN5A disrupt the flow of sodium ions into cardiac muscle cells, which affects heart electrophysiology and leads to clinical syndromes[93]. Understanding how genetic differences affect protein structure, ion channel function, molecular network dynamics and cellular behaviours (such as electrophysiology and cardiac events) would lead to a better understanding of diseases but requires well-integrated, multi-scale modelling.

Efforts are underway to achieve integrated multi-scale modelling that links molecules and genetics to physiology, especially for models of the heart [94], and large, community-driven projects have been

Homeodynamics
A concept that views an organism as a dynamical system; this concept emerged after the concept of homeostasis. Biological systems can be considered as homeodynamic: they can lose stability and show diverse behaviours, such as bi-stability, periodicity and chaotic dynamics.

launched. The long-running International Union of Physiological Sciences (IUPS) Physiome Project aims to promote basic science and to provide a technological foundation for integrated physiological models. Two new initiatives that started in 2010 are the Virtual Physiological Human (VPH) project in Europe and the High-Definition Physiology (HD-Physiology) project in Japan. The HD-Physiology Project, funded by the Japanese government, is trying to develop a comprehensive platform for the virtual integration of models from the molecular to whole body levels. It focuses on developing a combined model of whole-heart electrophysiology that is interconnected with cellular-, pathway- and molecular-level models and a whole-body metabolism model (FIG. 2).

Box 2 | **Model-analysis methods and tools**

Several different mathematical techniques have been developed to analyse the behaviour of complex biological models[88,89]. Here we describe the basic principles of some of the options: sensitivity analysis, phase-space analysis and metabolic control analysis.

Sensitivity analysis
The sensitivity of a system against various parameter changes is one of the properties that affects the robustness and fragility of a system. Sensitivity analysis can reveal not only the stability of a system against various perturbations, but can also provide information about the controllability of a system.

Phase-space analysis
As living systems operate under conditions of cellular homeostasis and homeodynamics, it is highly informative to study complex biological models to discover possible steady state and dynamical behavioural tendencies. Bifurcation analysis (the analysis of a system of ordinary differential equations (ODEs) under parameter variation) and phase-plane analysis (for example, the analysis of null-clines and local stability) help to predict systems behaviour (such as equilibrium or oscillations) when parameters are perturbed. (For details, please consult dedicated textbooks and papers[5,76,88,92].)

Metabolic control analysis
Metabolic control analysis (MCA) is a powerful quantitative framework for understanding the relationship between the properties of a metabolic network (at steady state) that is characterized by its stoichiometric structure and component reactions. MCA has been widely applied for the analysis of cellular metabolism, particularly for the analysis of the regulation of cellular metabolism. An alternative to MCA is flux-balance analysis (FBA); this a constraint-based modelling technique that has been applied in metabolic engineering[108,109]. FBA does not require details of enzyme kinetics or metabolite concentrations. It aims to compute metabolic fluxes across a network that maximizes certain system properties (such as growth rates) under conditions of constraint. Notably, FBA has been shown to accurately predict the growth rates of *Escherichia coli* under different culture conditions[109].

Model analysis is supported by many ODE solver systems (such as MATLAB), but more specialized tools are widely used in the community. Some examples are AUTO (a software package for bifurcation analysis) and XPPAut (a tool for solving ODEs that is capable of showing an orbit on the phase plane and that provides a user-friendly interface on AUTO). BUNKI and ManLab are MATLAB-based bifurcation analysis toolkits. Several tools support sensitivity analysis and parameter estimation; these include SBML-SAT, MATLAB SimBiology, ByoDyn and SensSB. SensSB is a MATLAB-based toolbox for the sensitivity analysis of systems biology models.

A related set of tools allows the study of metabolic networks. For example, DBSolve Optimum can be used for MCA computations and Kintecus is a software tool for simulating chemical kinetics, for MCA and for sensitivity analysis. These techniques fall into the category of constraint-based reconstruction and analysis (COBRA) methods, and several tools exist to support them. The COBRA Toolbox is a MATLAB-based toolbox that can be used to perform a variety of COBRA methods, including many FBA-based methods. MetNetMaker is a software tool that can create metabolic networks ready for FBA based on the KEGG LIGAND database.

Physiological modelling tools and standards. Currently there is no agreed standard for modelling physiological functions and for performing simulations at all levels of physiology. Indeed, more research is probably needed before these standards can be fully established. A hindrance to the development of standards in this field is the diversity of biological processes that operate at different spatiotemporal scales (such as in cells, tissues or organs); these processes require diverse modelling and numerical computation techniques[95]. CellML is a pioneering effort to define a markup language to describe mathematical models of physiology. Modelling languages are also available for specific fields, such as NeuroML[96] and NineML for describing models in computational neuroscience. Several tools that are based on these standards have been developed for physiological modelling (BOX 3). For example, the HD-Physiology project uses both CellDesigner (for cellular-level modelling) and PhysioDesigner, which is a software tool for modelling physiology from multicellular to whole-body levels. PhysioDesigner supports the *in silico* Markup Language (ISML)[97], which is an emerging standard XML-based language for multi-level physiological modelling, and is partially compatible with CellML and SBML. Both CellDesigner and PhysioDesigner can interface with other software platforms, and these tools are envisaged to be able to communicate with other tools through the Garuda platform (see below).

There also are publicly accessible resources that provide molecular structure and bioactivity data and that can be used for physiological modelling. These include RCSB PDB, ZINC, PubChem and PDBbind, the latter of which has had several of its commonly used programs comprehensively evaluated[98]. *In silico* simulation of protein–ligand interactions can be considered as an option for predicting the activity of small molecules, such as drugs[98,99]. This type of simulation can be performed using 'virtual docking' software, such as AutoDock Vina, GOLD or eHiTS.

Although integrating multiple levels of simulation has advantages, how this integration can be accomplished and how standards should be defined require further investigation. Some working standards are useful for clarifying the issues that need to be resolved and for outlining what can be achieved based on our current understanding; however, the introduction of obligatory standards may hamper the progress of the field.

An integrated software platform

Integrated software platforms have been a driving force of productivity, quality improvement and innovation in industries[11], and we can expect the same in systems biology. The concept is of an integrated software platform that enables users to access data and knowledge from any stage in the workflow, that allows the adaptation of the workflow to best fit the user's needs and that provides consistent user experiences and high levels of interoperability. All of these features can reduce the time costs that are associated with using independent and incompatible software. In principle, integrated platforms would significantly improve productivity and would reduce errors in the handling and analysis of complex data and models.

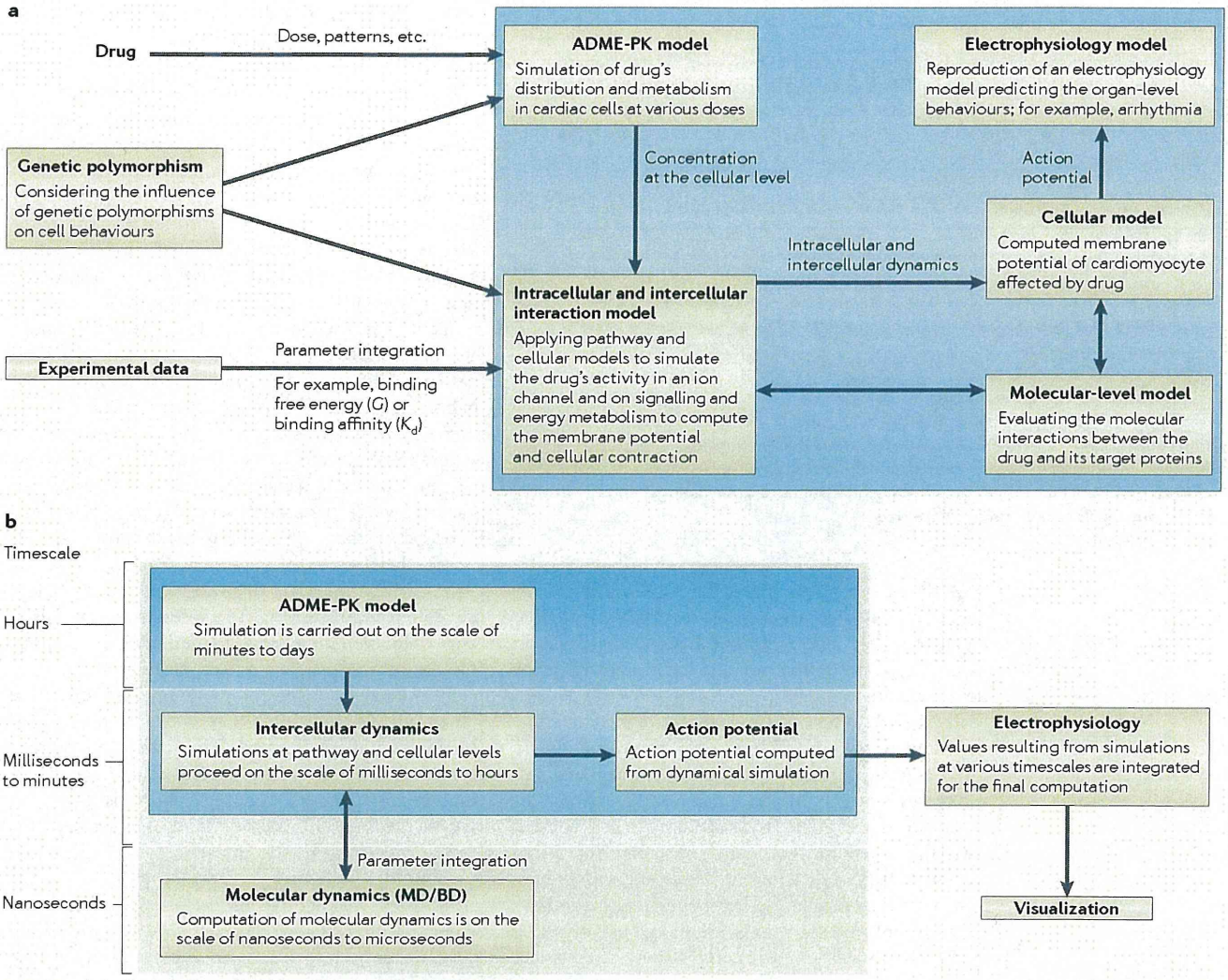


Figure 2 | **An example application of the High-Definition Physiology Project. a** | A possible use of an integrated multi-scale model is to evaluate the effects of a drug on cardiac events. A simulation condition can be set that consists of a specific drug dose and its temporal pattern of administration. Absorption, distribution, metabolism and excretion pharmacokinetics (ADME-PK) models that are built based on various molecular properties can compute drug distribution and metabolism, so that a change in the drug dose that a cardiomyocyte is exposed to can be simulated. The molecular properties of the drug can also be calculated using *in silico* methods[110], such as quantitative structure–activity relationship (QSAR) modelling, and can be applied as a parametric component to a specific cell model. Pathway- and cellular-level models use the computed drug dose as an environmental factor in the simulation of ion channel activity, signalling and energy metabolism and then compute the membrane potential and cellular contraction. In some cases, genetic polymorphisms may change the behaviours of the cell. For novel protein structures of ion channels or other important molecules, *in silico* simulations of molecular interactions may be used to better estimate the interaction parameters that are not experimentally known. The computed membrane potential can be used to reproduce the organ-level electrophysiology of arrhythmia. **b** | Three different timescales have to be coupled for the simulations that are outlined in part **a**, and the methods that are relevant to each simulation are computationally intensive. ADME-PK are simulated on the scale from minutes to days. Cellular- and pathway-level simulations are mostly on the scale of milliseconds to hours. Molecular dynamics is computed on the scale of nanoseconds to microseconds. Owing to these large differences in timescales, loosely coupled, dynamically measured simulations and precomputed values are used for the final integrated computation. Inevitably, different numerical solution methods need to be used, but they must function coherently. For example, fluid dynamics of the blood in a heart can be described by partial differential equations (PDEs). An electrocardiogram that is derived from the cardiac electrical activity can also be computed using PDEs, but most of the intracellular signalling and the whole-body ADME-PK model will be calculated by ordinary differential equations (ODEs). Close linkage of ODEs and PDEs is crucial in such a model. In those cases in which the stochastic behaviour of molecules has a crucial role, stochastic computation may also need to be used. MD/BD, molecular dynamics or Brownian dynamics.

Constraint-based reconstruction and analysis (COBRA). A suite of methods to simulate, analyse and predict various phenotypes using genome-scale models. These methods are used particularly for metabolic networks.