58. Lim, L. P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
59. Chan, S. W. *et al.* RNA silencing genes control *de novo* DNA methylation. *Science* **303**, 1336 (2004).
60. Aravin, A. A. *et al.* The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5**, 337–350 (2003).
61. Reinhart, B. J. & Bartel, D. P. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297**, 1831 (2002).
62. Cam, H. P. *et al.* Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome. *Nature Genet.* **37**, 809–819 (2005).
63. Decatur, W. A. & Fournier, M. J. RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.* **278**, 695–698 (2003).
64. Dominski, Z. & Marzluff, W. F. Formation of the 3′ end of histone mRNA. *Gene* **239**, 1–14 (1999).
65. Marzluff, W. F. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr. Opin. Cell Biol.* **17**, 274–280 (2005).
66. Nusinow, D. A. & Panning, B. Recognition and modification of sex chromosomes. *Curr. Opin. Genet. Dev.* **15**, 206–213 (2005).
67. Chow, J. C., Yen, Z., Ziesche, S. M. & Brown, C. J. Silencing of the mammalian X chromosome. *Annu. Rev. Genomics Hum. Genet.* **6**, 69–92 (2005).
68. Guthrie, C. & Patterson, B. Spliceosomal snRNAs. *Annu. Rev. Genet.* **22**, 387–419 (1998).
69. Sharp, P. A. The discovery of split genes and RNA splicing. *Trends Biochem. Sci.* **30**, 279–281 (2005).
70. Ofengand, J. & Fournier, M. J. in *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) Ch. 12 (American Society for Microbiology Press, Washington DC, 1998).

Acknowledgements
We would like to thank S. Eddy for helpful comments and suggestions and D. Bartel, V. Ambros, R. Bock, M. Terns, L.A. Huber, P. Loidl, N. Polacek and laboratory members from the Division of Genomics and RNomics for critical reading of the manuscript. The work discussed here was supported by an Austrian FWF (Fonds zur Förderung der wissenschaftlichen Forschung) and a German DFG (Deutsche Forschungsgemeinschaft) grant to A.H.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Division of Genomics and RNomics, Innsbruck Biocenter: http://genomics.i-med.ac.at
Access to this links box is available online.

OPINION

Addressing the problems with life-science databases for traditional uses and systems biology

Stephan Philippi and Jacob Köhler

Abstract | A prerequisite to systems biology is the integration of heterogeneous experimental data, which are stored in numerous life-science databases. However, a wide range of obstacles that relate to access, handling and integration impede the efficient use of the contents of these databases. Addressing these issues will not only be essential for progress in systems biology, it will also be crucial for sustaining the more traditional uses of life-science databases.

Several decades ago, scientists started to set up biological data collections for the centralized management of and easy access to experimental results, and to ensure long-term data availability (FIG. 1a). Many early data collections were initially administered using word processing or spreadsheet applications. Owing to the limited amount of data that could be stored in this way, and the reductionist viewpoint that characterized most biological research at that time, this approach to data collection seemed reasonable, and was sufficient for occasional exchanges with colleagues.

However, with the exponential growth of experimental data that is taking place owing to rapid biotechnological advances and high-throughput technologies, as well as the advent of the World Wide Web as a new means for data exchange, the world dramatically changed. The huge amounts of data that are now produced on a daily basis require more sophisticated management solutions, and the availability of the internet as a modern infrastructure for scientific exchange has created new demands with respect to data accessibility. Furthermore, the relatively new field of systems biology has further increased the requirements that are demanded of life-science databases. The general vision of systems biology is to move out of the era of reductionist studies of isolated parts of interest — for example, individual proteins and genes — and to develop a molecular understanding of more complex structures and their dynamics, such as regulatory networks, cells, organs and, ultimately, whole organisms[1].

The most important tool for reaching an understanding of biology at the level of systems is the analysis of biological models (FIG. 1b). The basic building blocks for these models are existing experimental data, which are stored in literally thousands of databases[2–4]. As a result, database integration is a fundamental prerequisite for any study in systems biology[5,6]. Because database integration has long been recognized as a key technology in the life sciences, research in this area also has a long tradition. However, although many approaches exist, database integration in the life sciences is still far from being trivial.

A common misconception is that the main problems of database integration are related to the technology that is used for these purposes. Here we argue that although the mastering of such technology can be challenging, the main problems are actually related to the databases themselves. There are many issues with life-science databases that prevent the effective use of integration technology. These problems not only have adverse effects on the quintessential task of ensuring data availability to the general research community, but present an even greater obstacle to systems biology. Here we provide a systematic analysis of the common problems that relate to life-science databases — which are technical, social and political — and suggest solutions for how they could be overcome.

Technical problems

As a prerequisite for the discussion of technical problems with life-science databases it is important to understand the general principles of database integration. Life-science databases have experienced an exponential growth in numbers in recent years and contain information of many types[7]. To bridge the gap between these often unconnected islands of biological knowledge, and between the different types of experimental data that they contain, various approaches to data integration have been pursued over the past decade. These range from basic hypertext linking to more advanced approaches that involve the use of federated databases and data warehouses (BOX 1). It is on the advanced approaches that we focus here, as they provide the best illustration of the diverse problems with life-science databases that affect data integration, particularly with respect to the goals of systems biology.

Although there are many variants of the more advanced applications, the problems with life-science databases that affect integration using federated database technology or data warehouses are almost identical.

Here we use the popular data warehouse approach as an example. The typical steps in database integration that are followed in this approach are illustrated in FIG. 2. First, the databases to be integrated must be identified, and data must be extracted from the identified sources. The extracted data are usually then preprocessed; for example, this can involve the conversion of data sources into more accessible formats such as XML. Source database structures are then mapped to a so-called integrated schema, which defines an integrated structure over all the data sources to be integrated (BOX 1). Finally, the data are imported into the data warehouse, where integrated access is made possible, for example, through a 'browsable' search interface.

Here we provide a detailed discussion of the problems with life-science databases that can occur at each stage of the data-integration process outlined above.

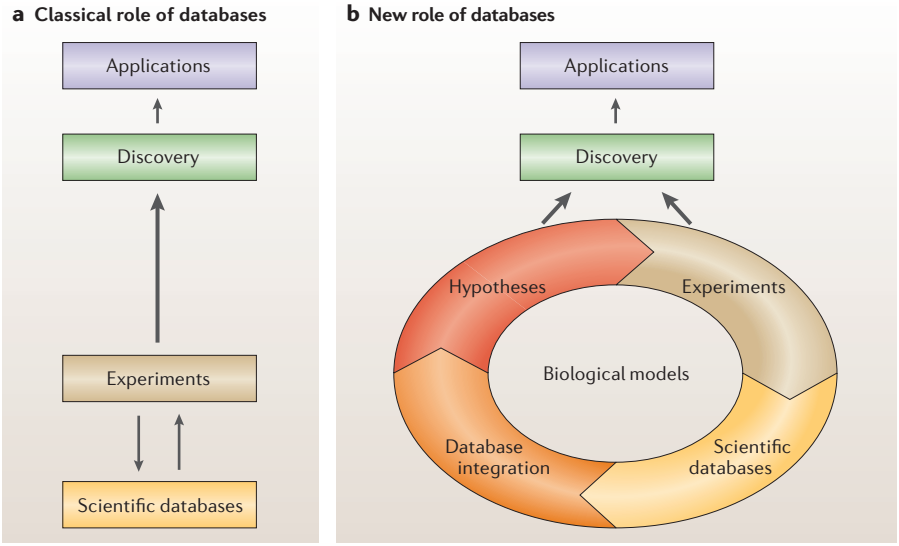

Figure 1 | **Classical and systems biology roles of life-science databases.** The classical role of life-science databases is to provide easy access to and long-term storage of experimental results, with centralized data management. By contrast, more recent systems biological approaches exploit the information in databases to generate hypotheses for *in silico* discovery, which, after experimental verification, can be used to populate other databases.

Web access problems. The first step in building models for systems biology is the identification of suitable data sources. Therefore, a description of at least the database contents and the way in which the data are produced and/or derived from other data sources is mandatory. Unfortunately, not every life-science database provides such meta-information. For example, it is often assumed that online visitors to a database already know what type of data are stored there. Given the number of life-science databases, however, this is unrealistic. As a prerequisite for identifying and using data, each database should be described with appropriate meta-information, such as the type of data that are stored, the way these data have been produced, the guidelines for data curation, the structures used to store the data, and information about release management and database versions.

To examine available data in more detail, browsing the contents of a database with a searchable interface that is accessible using a web browser is usually the preferred method of access. However, if a search interface has been successfully located, which can sometimes be difficult, it is not guaranteed that the data can be appropriately searched. Common problems are that interfaces do not allow all fields in a database to be searched, search modes such as 'and', 'or' and 'not' are either not supported at all or only in a rudimentary way, and query results cannot be downloaded for further processing. For large data collections, just one of these obstacles can render a search interface useless.

In systems biology, the examination of isolated data entries usually provides little insight, as it is the identification and discovery of relationships between entries that is most important. Web-based browsing interfaces would therefore ideally offer as many links as possible between individual entries in different databases. However, not all database providers support such direct links from external sources. Biologists are consequently forced to make use of search interfaces, which renders discovery by browsing across different databases an awkward approach at best.

Problems with data extraction and lack of software interfaces. If a biologist is interested not only in a few entries of a database but in a potentially larger number, which are to be further processed, web-based searching and browsing is clearly no longer a viable option. To support large-scale data processing, the collection of relevant data has to be automated. Therefore, each database would ideally be equipped with programming interfaces that enable software developers to query and search databases from within their own programs.

Although modern database management systems support mature standard interfaces for this purpose, such as ODBC (Open Database Connectivity) and JDBC (Java Database Connectivity), public access to these interfaces is only rarely granted by database providers. The reasons for this access restriction range from security concerns to political issues, as discussed below. There are, however, a few databases that allow access to life-science data with 'canned queries' through the use of web services[8] as a more recent technical standard. With the help of web services, predefined queries can be used to automatically access a remote database. Prominent examples in this context are the DDBJ (DNA Data Bank of Japan)[9], KEGG (Kyoto Encyclopedia of Genes and Genomes)[10] and the EBI (European Bioinformatics Institute)[11].

As web-based access is not suitable for bulk queries and programming interfaces are only rarely available, the complete download of life-science data collections provides a last resort for large-scale data processing. If a database is not available for download its contents have to be extracted from the web interface. For these purposes, specifically tailored data-extraction software is needed for each data source. However, there are technical drawbacks to this approach. First, every change in the web layout of a source database usually necessitates corresponding changes to the data-extraction software. In addition, data extraction from large databases takes considerable time, in some cases in the range of several days.

Many of the technical problems discussed above have their roots in the fact that database management systems, such as PostgreSQL and mySQL, are infrequently used to store data; word processors and spreadsheet applications are used instead.

Box 1 | **Integration methods for life-science databases**

Database integration has long been recognized as one of the most important fields in bioinformatics, and several technologies and approaches are used for this purpose.

Hypertext links

The most basic way of 'integrating' life-science data from different sources is to provide simple hypertext links between related entries in different databases. By doing this, the user is supported in browsing and exploring the contents of different data sources. However, although web links are important, in comparison to the total number of existing databases, such links are usually provided to only relatively few other databases. Furthermore, web links do not provide much support beyond explorative browsing. For example, they do not allow bulk queries over several data sources.

Full-text indexing

A popular approach that overcomes this problem is based on full-text indexing, which, from a technological point of view, uses similar technology to current search engines. Systems such as the SRS (Sequence Retrieval System)[58] locally mirror the contents of databases as flat files and create a full-text index over all locally mirrored data sources. This way a user can address multiple databases with a single query. The drawbacks are that there is no integration beyond the shared full-text index and also semantic queries that make use of ontologies are usually not supported.

Federated databases and data warehouses

More recent developments in data integration for the life sciences are based on federated databases and data warehouses. Federated database systems, such as DiscoveryLink[59], make use of an application-specific integrated schema, which specifies an integrated conceptualization over the databases to be integrated; that is, it describes how data are structured in the integrated database. The remote data sources that are to be integrated are mapped to such an integrated schema by means of rules that define in which way entries of a data source are to be inserted into the schema. A user can then pose queries against the integrated schema, which exploits the mapping definitions in order to integrate the requested data on demand. Although the advantage of this approach is that the returned data entries are always the most recent ones, the disadvantage is that the response times for complex queries can be long, in some cases in the range of several days.

Data warehouse approaches, such as ONDEX[60] and others, also make use of an integrated schema, but locally mirror the data sources to be integrated. Although this approach considerably improves the efficiency of query processing, the drawback is that the returned data might be out of date if there have been remote updates that have not yet been locally mirrored.

For a more detailed discussion on the current 'state of the art' in data integration for the life sciences see REFS 61,62.

For example, the use of database management systems greatly simplifies the provision of powerful web search interfaces and programming interfaces (such as JDBC and ODBC) for effective data access. Therefore, the use of such systems should be a standard for data providers.

Problems with data preprocessing. If a provider supports the download of a database, which is the case for only about 50% of databases (Information Systems and Databases Group, Univ. Koblenz, unpublished observations), flat files are still used as a *de facto* standard for data exchange in the life sciences. Because there is no standardized format for flat files, there are in effect many formats for the thousands of biological data collections. To process a flat file, appropriate parsing software is needed. For some of the more prominent databases such parsers are developed by open source projects such as BioJava, BioPerl[12] and BioRuby (reviewed in REF. 13). However, despite these efforts, not every parser supports all of the fields in a specific flat-file format and there is often a long delay between the modification of a flat-file format and the availability of an updated parser. Furthermore, there are many databases for which no free parser is available. As a consequence, the development of parsers is an integral part of almost any project for the integration of life-science data. Because flat-file formats are mostly non-trivial in their structure, not always well documented and change their structures over time, the development and maintenance of parsers requires considerable effort.

The distribution of life-science data as self-describing XML files would solve most of these problems, as generic XML parsers are available for almost every platform and programming language. However, because XML is a relatively new technology, currently only about 5% of the publicly available life-science databases, such as some EMBL (European Molecular Biology Laboratory) databases[14], KEGG[10] and UniProt (the Universal Protein Resource)[15], are provided in an XML format. As the importance of XML has been recognized for systems biology, several initiatives (such as BioPAX[16], CellML[17], MAGE[18], PSI-MI (Proteomics Standards Initiative — molecular interaction)[19] and SBML (systems biology markup language[20]) are working towards the standardization of XML-based data-exchange formats.

Inappropriate conceptualizations. Even if databases can be accessed through programming interfaces or parsers, their integration is still far from being trivial, as the conceptual structure of each source database has to be mapped to a unifying target schema. However, source databases often use inappropriate conceptualizations — that is, the underlying data structures do not allow for the correct representation of all relevant information.

The same type of data is occasionally represented in different ways, sometimes even within a single database. In one prominent case, data on enzymatic reactions were not stored appropriately, as the database lacked the ability to represent more than one educt and product per enzymatic reaction (FIG. 3). Consequently, a more appropriate representation was introduced, but because only newly inserted entries are represented in the more sophisticated way, data integration using this resource is a considerable challenge. Similar problems in pathway databases are reported in REF. 21.

Ontologies and controlled vocabularies are frequently used in the life sciences as semantic references, which define commonly agreed definitions of real-world entities (concepts) and the relationships between them. At the level of data entries such semantic references are ideally used for encoding fields in life-science databases — such as the NCBI (National Center for Biotechnology Information) Taxonomy IDs for species names and EC (enzyme class) numbers of the enzyme nomenclature for enzymatic functions — instead of manually created, and therefore potentially wrongly typed, free text descriptions. At the schema level, ontologies and controlled vocabularies can be used for semantic data integration[22–25]. For example, if two sources store data about proteins and one database structure is named 'protein_entries' and the other 'p_data', a reference from both to the semantically defined ontological concept 'protein' can be exploited to link entries between the two sources, even if there are syntactical differences at the schema level.

Although the use of ontologies and controlled vocabularies such as the Gene Ontology[24], the EC numbers[26] and the NCBI

taxonomy[27] is generally a valuable concept[28], semantic references are not always built in a meaningful way. For example, functions and processes should not be assigned to genes in ontologies, but rather to gene products — owing to splicing events, not all the products of a given gene have the same function, nor are they expressed in the same cell types. More importantly, it is inappropriate to assign functions to individual genes, as a function is not inherent to a gene, but to the interactions between gene products and to the coordinated expression of genes in time and space (see for example REF. 29). Clearly, in many cases more complex data structures are needed to better reflect biological reality (for example, see FIG. 3c).

Another important problem is the representation of facts in semantic references themselves. The information provided by some of these references is, unfortunately, not fine-grained enough to appropriately capture the complexity of biological knowledge. In one prominent example, different types of relationship — such as 'is a' and 'is part of' — are not distinguished and only a single type of relationship is used. Consequently, the use of sophisticated data-integration methods, which rely on fine-grained information about the types of relationship between concepts in a semantic reference, is impossible. A more detailed account of conceptual problems in ontologies is given in REFS 30,31.

Inappropriate conceptualizations as described above often lead to considerable problems, as the early loss of information in source databases and ontologies can only rarely be compensated for at a later stage. Pragmatic guidelines to avoid some of the above problems are given in REFS 32–36.

Problems with the contents of databases. Although it is a widespread belief that we live in the 'post-genomic era', most of the data produced still come from new sequencing projects. For systems biology, many types of information are often missing, including functional annotations of genes and proteins, genotype–phenotype relationships, kinetic values for enzymes and detailed pathway information. Even when information that is based on sequence similarity is taken into consideration, only about 50% of all reaction steps in metabolic pathways can be linked to the genes and proteins that catalyse them[37]. In consequence, the parameterization of systems biological models is partly based on guesswork, which undermines reliability, credibility and predictive potential. More projects should therefore be financed to provide the missing data. Furthermore, funding agencies should demand the submission of results from experiments that they have financed to public databases as a standard, which is not yet the case for projects in every domain.

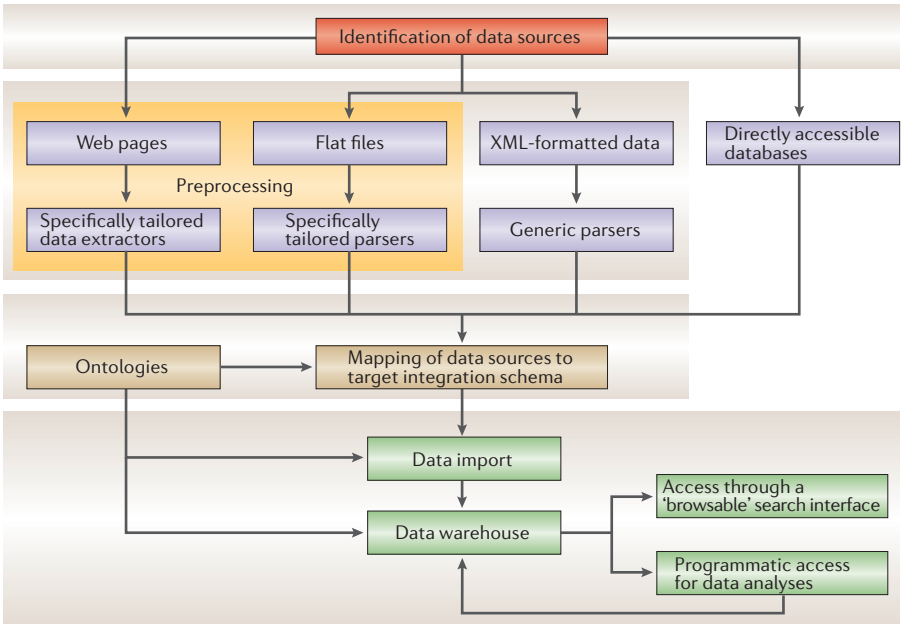

Figure 2 | **The database integration process: a database warehouse as an example.** The first step for any approach to database integration is the identification of suitable data sources for a given application. For example, in the case that a database is needed to integrate all known data about mammalian transcription factors, some of the data sources to be integrated would be from the TransFac[63] and TRRD[64] databases. The next step is the extraction of the data that are to be integrated; depending on the accessibility of each data source, the specific actions to be carried out differ. Data can be extracted from web pages, which takes considerable effort, or can be downloaded as traditional or XML-formatted flat files. Direct access to a database by means of standardized programming interfaces is rarely possible. During the subsequent preprocessing stage, data are often converted into more accessible formats — for example, from a traditional flat-file format into an XML representation. Individual data sources are then mapped to a so-called integrated schema, which defines a structure in which the data to be integrated are stored. Ideally, such mapping makes use of suitable ontologies, which are commonly agreed definitions of real-world concepts that can be used as references for semantic data integration. After mapping to the integrated schema, the data sources can be imported into the warehouse. Depending on the needs of the particular application, this stage also often includes data cleansing, merging of entries from different sources and deletion of duplicates. When the data sources are finally integrated, access to the resulting data warehouse can be provided through different types of interface.

Another problem is that biological data are usually 'entry-centric' for historical reasons — that is, information about biological links between data entries is often unavailable[38]. Data stored in this way need to be more complete for systems biological analyses, where relationships between biological entities are important. Furthermore, it is often impossible to clearly identify entries in databases. Many databases do not provide accession numbers, and even if accession numbers are available, they might not be stable across different versions of the same database. As a consequence it is often impossible to accurately reproduce search results.

Even if data are available, there is no guarantee that they are valid. A common source of problems is error propagation in sequence annotations through the use of automated annotation mechanisms[39]. These problems can be avoided by using evidence codes to keep track of how annotations were created, and by making sure that annotations are automatically inferred only from manual annotations and not from other automatically inferred ones[40,41]. Another problem is that EST data are usually stored without related trace files from the sequencing experiment and therefore valuable information is lost. Primary nucleotide sequence databases only apply basic quality checks, partly because more extensive ones would be unmanageable, and partly to enable users of the submitted data to decide which data are sufficiently reliable. However, users of these data have no means of assessing their correctness and quality, despite the fact

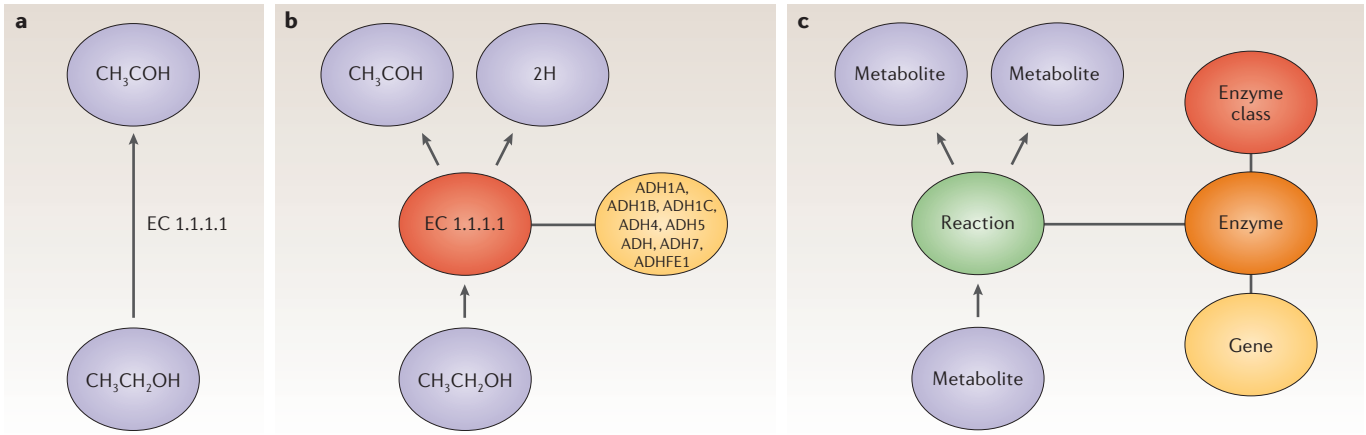

Figure 3 | **Alternative representations of metabolic pathways: alcohol dehydrogenase as an example.** In the simplest representation (**a**), two metabolites are linked by an arc that represents the enzyme class EC 1.1.1.1 of the enzyme nomenclature[26]. This representation disregards the stoichiometry of the reaction and is also incomplete, as it does not allow the actual genes and enzymes that catalyse a specific biochemical reaction to be represented. Although the representation shown in panel **b** overcomes this issue it introduces new problems: in contrast to what is suggested by the representation, not all alcohol dehydrogenases (ADHs) catalyse the specific reaction that is shown. Furthermore, this representation does not capture situations where only certain splicing variants of the enzyme have a given catalytic function. The representation in panel **c** removes this ambiguity by directly linking enzymes to the reactions that they catalyse. By associating the enzymes with genes and enzyme classes, no information is lost in comparison to the previous two representations. Most pathway databases use representations **a** and/or **b**, whereas only **c** overcomes most of the conceptual problems.

that even basic uses, such as the prediction of ORFs, would benefit from the availability of sequence trace files. The recent development of the Trace Archive at the Sanger Institute (see link to the Ensembl Trace Server) is a significant step in the right direction.

Several recent surveys also reveal problems with the statistical methods that are used in publications[42], incorrect data that are due to flaws in platform technologies[43] and a lack of independent replication and validation of experimental results[44,45]. Furthermore, human error is a frequent source of mistakes in life-science databases. Although manual curation is important for quality assurance in most areas, there is also a high potential for the introduction of errors. Some examples include: specific release versions of databases from which whole sections with hundreds of entries are missing; the absence of entries to which references within the same database exist; multiple variants of spellings for the same species in a single database; duplicate lines in flat files; the occurrence of numbers in sequence information; and typing errors in every imaginable field of databases. Further problems with biological data have been reported, for example in REFS 21,46–48.

To minimize the risk of human error at the data level, appropriate curation protocols and supporting software, as well as suitable ontologies and controlled vocabularies, have to be developed and used wherever possible. As errors in databases are common, data providers should also implement appropriate means for reporting, tracking and correcting errors.

Social issues

Communication problems. The problems discussed above make it clear that there should be a close two-way communication between database providers and database users in order to collectively address these issues. Although this would seem to be an obvious course of action, there are, in fact, serious problems. Although many database providers are open and freely offer information about their underlying conceptual schema, data-curation processes and changes of schema and database contents, this is not yet standard practice. Furthermore, although many database providers are responsive with respect to reported errors and suggestions for improvements, this unfortunately does not hold in every case. In fact, hardly any provider of a life-science data collection offers clearly visible means for error reporting and tracking.

Educational problems. Many problems with life-science databases have their origins in the fact that biologists often lack even a basic understanding of data management, and bioinformaticians are often not aware of biological needs. Because the communication difficulties that arise from these problems clearly have educational roots, the interfaces in the curricula for both bioinformaticians and biologists should be better defined so that future students are better equipped for the challenges ahead.

Political obstacles

Licensing and access problems. Some problems with life-science databases clearly do not have technical or social roots, but have 'political' causes. Perhaps the most important problem in this context is the question of free or commercially limited access to life-science data, which is a highly controversial area (see for example REFS 49,50). It might seem obvious that free access for all to life-science data would be beneficial, but the reality is different. Because data curation is labour intensive and requires highly qualified personnel, financing questions naturally arise[51,52]. As most funding agencies do not provide long-term support for data curation, alternative funding models have to be used.

The consequences of this depend on the funding models of the databases. Some important databases and ontologies are not publicly available at all — that is, even academic institutions have to acquire licences for their use. Other databases are freely accessible through a web interface, but cannot be downloaded. Although some providers offer the download of their data with a commercial licence, which academic institutions can purchase, others do not. Automated parsing of data from web interfaces does not necessarily help this situation, as there are reported cases where database providers block requests from whole domains where it is suspected that someone attempts to 'steal' their data using such an approach.

Even if data are freely available, it still does not mean that all non-technical

Box 2 | **Suggested solutions to problems with life-science databases**

Actions for funding agencies
- Support more projects to deliver data for systems biology approaches.
- Support projects for the long-term curation of life-science databases.
- Demand the submission of experimental results to public databases.

Actions for educational organizations
- Educate biologists and bioinformaticians to use better defined interfaces between the respective curricula.

Actions for database users
- Report errors in databases.
- Submit new findings to public databases.

Actions for database providers
- Make use of database management systems.
- Use appropriate conceptualizations.
- Make use of appropriate means for data curation.
- Provide proper documentation about: the kind of data that are stored; the way data are produced; the guidelines for data curation; the data structures that are used to store the data; and information about release management and versioning.
- Offer proper access to data through: a powerful web interface; programming interfaces; and/or download facilities.
- Use XML for the exchange of biological flat files.
- Store data from other groups who work in the same area.
- Correct reported errors.
- Use stable accession numbers for identifying individual database entries across database versions.
- Remove inappropriate clauses from licences.

Actions for publishers of scientific articles
- Publish articles about life-science databases only if their providers fulfil the requirements listed above.

problems that are related to data integration are solved. Obeying the conditions of 'free' licences imposes considerable obstacles. For example, many databases demand that the origin of the data is transparent to the user and often provide detailed instructions on how this should be achieved. Developing a data-integration system that fulfils all the requirements of the different database providers is not only technically challenging, but inevitably results in software that is not user friendly. Therefore, even popular database-integration systems often ignore these requirements in practice.

Another licensing problem is that the redistribution of data, and therefore of the biological models that are developed through data integration, is often not allowed. One of the major pathway databases even demands co-authorship in any publication that makes use of the database in any way. If the developers of the BLAST algorithm had insisted on such a licensing condition, they probably would have a publication list consisting of millions of papers!

It is clear from the above that a universal legal framework as a prerequisite for database interoperability is needed. A suggestion for how such a framework might operate is detailed in REF. 53.

Funding issues. As life-science databases are fundamental to a whole discipline, their importance cannot be overestimated. However, because long-term curation projects are only rarely supported, only a few providers of publicly available databases do not have funding problems. Support for the long-term curation of life-science databases is therefore urgently needed.

Requirements for publications. To support a broad implementation of standards for life-science databases, publishers and peer reviewers of database-related articles should ensure the fulfilment of the points that we have suggested for database providers (BOX 2). The fulfilment of preconditions before the publication of scientific articles is not a new idea. For example, articles about newly sequenced genes and microarray experiments are only published if the findings have already been submitted to a public database in order to guarantee that certain standards with respect to quality and availability are met, such as the minimum information about a microarray experiment (MIAME)[54] (see also REF. 55 for further discussion).

Conclusions

The problems that we have outlined clearly affect the traditional roles of life-science databases. For example, a badly designed web interface can have an effect on the availability of data that can have as much impact as a lack of funding to further maintain a database. The additive effects of the problems that we have discussed are even greater when different databases are integrated, as is required for progress in systems biology.

Some providers have already identified the difficulties with their databases and have started the transition from *ad hoc* data collection to the adoption of higher standards. The Protein Data Bank[56] is a good example; the recent overhaul of this database solves

Glossary

Controlled vocabulary
A standardized set of terms that can be used in a given application domain. A prominent example is the enzyme class nomenclature, which describes classes of biochemical reaction.

Database management system
A system that provides a means of storing, modifying and extracting data from a database.

Evidence code
A controlled vocabulary that is used to track the types of evidence that support a gene annotation.

Flat file
Human readable, non-standardized files that can be used to exchange the contents of life-science databases.

Ontology
A commonly agreed definition of real-world concepts, such as 'protein' and 'enzyme', and their particular relationships, for example, an enzyme 'is a' protein.

Parser
Software that reads a given input, such as a flat file, for further processing.

Web service
A standardized way to allow for interoperable machine-to-machine interaction over a network.

XML
The extensible markup language (XML) is a standard for the creation of application-specific, self-descriptive markup languages, which, for example, can be used for the definition of data-exchange formats.

many of the issues that have been criticized in the past (for example, see REF. 57). However, many providers are unaware of all the difficulties with their databases. We hope that the actions that we have suggested (which are summarized in BOX 2) provide a starting point for further discussions and evolve into standards, which will in the long run contribute to an increased quality and usability of life-science databases, for both traditional and systems biological uses.

Stephan Philippi is at the Department of Computer Science, University of Koblenz, PO Box 201602, 56016 Koblenz, Germany.
Jacob Köhler is at the Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK.
Correspondence to S.P.
e-mail: stephan.philippi@uni-koblenz.de

doi:10.1038/nrg1872
Published online 9 May 2006

1. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
2. Pennisi, E. How will big pictures emerge from a sea of biological data? *Science* **309**, 94 (2005).
3. Roos, D. S. Computational biology. Bioinformatics — trying to swim in a sea of data. *Science* **291**, 1260–1261 (2001).
4. Augen, J. Information technology to the rescue! *Nature Biotechnol.* **19**, BE39–BE40 (2001).
5. Ge, H., Walhout, A. J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19**, 551–560 (2003).
6. Carel, R. Practical data integration in biopharmaceutical research and development. *PharmaGenomics* 22–35 (June 2003).
7. Galperin, M. Y. The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res.* **34**, D3–D5 (2006).
8. Cerami, E. *Web services essentials* (O'Reilly, Beijing; Sebastopol, California, 2002).
9. Sugawara, H. & Miyazaki, S. Biological SOAP servers and web services provided by the public sequence data bank. *Nucleic Acids Res.* **31**, 3836–3839 (2003).
10. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
11. Pillai, S. *et al.* SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res.* **33**, W25–W28 (2005).
12. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611–1618 (2002).
13. Mangalam, H. The Bio* toolkits — a brief overview. *Brief. Bioinformatics* **3**, 296–302 (2002).
14. Wang, L., Riethoven, J. J. & Robinson, A. XEMBL: distributing EMBL data in XML format. *Bioinformatics* **18**, 1147–1148 (2002).
15. Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).
16. Luciano, J. S. PAX of mind for pathway researchers. *Drug Discov. Today* **10**, 937–942 (2005).
17. Lloyd, C. M., Halstead, M. D. & Nielsen, P. F. CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* **85**, 433–450 (2004).
18. Spellman, P. T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, RESEARCH0046 (2002).
19. Orchard, S. *et al.* Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17–20th April 2005). *Proteomics* **5**, 3552–3555 (2005).
20. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
21. Green, M. L. & Karp, P. D. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res.* **33**, 4035–4039 (2005).
22. Stevens, R. *et al.* TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* **16**, 184–185 (2000).
23. Köhler, J., Philippi, S. & Lange, M. SEMEDA: ontology based semantic integration of biological databases. *Bioinformatics* **19**, 2420–2427 (2003).
24. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
25. Philippi, S. & Köhler, J. Using XML technology for the ontology-based semantic integration of life science databases. *IEEE Trans. Inf. Technol. Biomed.* **8**, 154–160 (2004).
26. NC-IUBMB. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes* (Academic Press, San Diego, 1992).
27. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32**, D35–D40 (2004).
28. Hendler, J. Communication. Science and the semantic web. *Science* **299**, 520–521 (2003).
29. Noble, D. Will genomics revolutionise pharmaceutical R&D? *Trends Biotechnol.* **21**, 333–337 (2003).
30. Smith, B., Köhler, J. & Kumar, A. On the application of formal principles to life science data: a case study in the gene ontology. *Proc. Data Integr. Life Sci. First Int. Workshop* 79–94 (2004).
31. Zhang, S. & Bodenreider, O. Law and order: assessing and enforcing compliance with ontological modeling principles in the Foundational Model of Anatomy. *Comput. Biol. Med.* 6 Sep 2005 (doi:10.1016/j.compbiomed.2005.04.007).
32. van Helden, J. *et al.* Representing and analysing molecular and cellular function using the computer. *Biol. Chem.* **381**, 921–935 (2000).
33. Bornberg-Bauer, E. & Paton, N. W. Conceptual data modelling for bioinformatics. *Brief. Bioinformatics* **3**, 166–180 (2002).
34. Nelson, M. R., Reisinger, S. J. & Henry, S. G. Designing databases to store biological information. *BioSilico* **1**, 134–142 (2003).
35. Taylor, C. F. *et al.* A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnol.* **21**, 247–254 (2003).
36. Ma, Z. & Chen, J. (eds) *Database Modeling in Biology: Practices and Challenges* (Springer, in the press).
37. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
38. Searls, D. B. Data integration — connecting the dots. *Nature Biotechnol.* **21**, 844–845 (2003).
39. Karp, P. D. What we do not know about sequence analysis and sequence databases. *Bioinformatics* **14**, 753–754 (1998).
40. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **32**, D262–D266 (2004).
41. Gattiker, A. *et al.* Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.* **27**, 49–58 (2003).
42. Garcia-Berthou, E. & Alcaraz, C. Incongruence between test statistics and P values in medical papers. *BMC Med. Res. Methodol.* **4**, 13 (2004).
43. Mecham, B. H. *et al.* Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics* **18**, 308–315 (2004).
44. Ntzani, E. E. & Ioannidis, J. P. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439–1444 (2003).
45. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
46. Müller, H., Naumann, F. & Freytag, J.-C. Data quality in genome databases. *Proc. Conf. Inf. Qual. (IQ 03)* 269–284 (2003).
47. Iliopoulos, I. *et al.* Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* **19**, 717–726 (2003).
48. Leser, U. & Hakenberg, J. What makes a gene name? Named entity recognition in the biomedical literature. *Brief. Bioinformatics* **6**, 357–369 (2005).
49. Resnik, D. B. Strengthening the United States' database protection laws: balancing public access and private control. *Sci. Eng. Ethics* **9**, 301–318 (2003).
50. Maurer, S. M., Hugenholtz, P. B. & Onsrud, H. J. Intellectual property. Europe's database experiment. *Science* **294**, 789–790 (2001).
51. Merali, Z. & Giles, J. Databases in peril. *Nature* **435**, 1010–1011 (2005).
52. Ellis, L. B. & Kalumbi, D. The demise of public data on the web? *Nature Biotechnol.* **16**, 1323–1324 (1998).
53. Greenbaum, D. & Gerstein, M. A universal legal framework as a prerequisite for database interoperability. *Nature Biotechnol.* **21**, 979–982 (2003).
54. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genet.* **29**, 365–371 (2001).
55. Bourne, P. Will a biological database be different from a biological journal? *PLoS Comput. Biol.* **1**, 179–181 (2005).
56. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
57. Rother, K. *et al.* Columba: multidimensional data integration of protein annotations. *Proc. Data Integr. Life Sci. First Int. Workshop* 156–171 (2004).
58. Zdobnov, E. M., Lopez, R., Apweiler, R. & Etzold, T. The EBI SRS server — recent developments. *Bioinformatics* **18**, 368–373 (2002).
59. Haas, L. M. *et al.* DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Syst. J.* **40**, 489–511 (2001).
60. Köhler, J. *et al.* Linking experimental results, biological networks and sequence analysis methods using Ontologies and Generalised Data Structures. *In Silico Biol.* **5**, 33–44 (2004).
61. Stein, L. D. Integrating biological databases. *Nature Rev. Genet.* **4**, 337–345 (2003).
62. Köhler, J. Integration of life science databases. *Drug Discov. Today* **2**, 61–69 (2004).
63. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
64. Kolchanov, N. A. *et al.* Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.* **30**, 312–317 (2002).

Acknowledgements
The authors would like to thank C. Rawlings and P. Verrier for commenting on an earlier version of this article. Furthermore we would like to thank the following individuals for exploring with us the pitfalls of life-science databases over the past years: J. Baumbach, J. Butz, E. Kirchem, F. Klingert, S. Knop, B. Kormeier, I. Kupp, A. Neu, A. Rüegg, A. Skusa, B. Steuernagel, J. Taubert, P. Verrier and R. Winnenburg. S.P. gratefully acknowledges funding by the European Science Foundation. Rothamsted Research receives grant-aided support from the UK Biotechnological and Biological Science Research Council.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
BioJava: http://biojava.org
BioPAX — Biological Pathways Exchange: http://www.biopax.org
BioPerl: http://bioperl.org
BioRuby: http://bioruby.org
CellML: http://www.cellml.org
DiscoveryLink: http://www.developer.ibm.com/university/scholars/products/lifesciences/discoverylink
DNA Data Bank of Japan: http://www.ddbj.nig.ac.jp
EC (enzyme class) numbers of the enzyme nomenclature: http://www.chem.qmul.ac.uk/iubmb/enzyme
Ensembl Trace Server: http://trace.ensembl.org
European Bioinformatics Institute SRS server: http://srs.ebi.ac.uk
European Bioinformatics Institute: http://www.ebi.ac.uk
Extensible markup language (XML): http://www.w3.org/XML
Gene Ontology homepage: http://www.geneontology.org
Kyoto Encyclopedia of Genes and Genomes: http://www.genome.jp/kegg
Microarray Gene Expression Data Society: http://www.mged.org
mySQL: http://www.mysql.com
NCBI taxonomy: http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html
Nucleic Acids Research Database Categories List: http://www3.oup.co.uk/nar/database/c
ONDEX: http://ondex.sourceforge.net
Open Biomedical Ontologies: http://obo.sourceforge.net
Open Source Initiative License Index: http://www.opensource.org/licenses
PostgreSQL: http://www.postgresql.org
Proteomics Standards Initiative — molecular interaction: http://psidev.sourceforge.net
Systems biology markup language: http://sbml.org
Universal Protein Resource: http://www.ebi.uniprot.org/index.shtml
Web Services Activity: http://www.w3.org/2002/ws
Access to this links box is available online.