doi:10.1093/bib/bbq067

Challenges in integrating *Escherichia coli* molecular biology data

Anália Lourenço[1], Sónia Carneiro[1], Miguel Rocha[2], Eugénio C. Ferreira[1] and Isabel Rocha[1]

Submitted: 10th June 2010; Received (in revised form): 5th September 2010

Abstract

One key challenge in Systems Biology is to provide mechanisms to collect and integrate the necessary data to be able to meet multiple analysis requirements. Typically, biological contents are scattered over multiple data sources and there is no easy way of comparing heterogeneous data contents. This work discusses ongoing standardisation and interoperability efforts and exposes integration challenges for the model organism *Escherichia coli* K-12. The goal is to analyse the major obstacles faced by integration processes, suggest ways to systematically identify them, and whenever possible, propose solutions or means to assist manual curation. Integration of gene, protein and compound data was evaluated by performing comparisons over EcoCyc, KEGG, BRENDA, ChEBI, Entrez Gene and UniProt contents. Cross-links, a number of standard nomenclatures and name information supported the comparisons. Except for the gene integration scenario, in no other scenario an element of integration performed well enough to support the process by itself. Indeed, both the integration of enzyme and compound records imply considerable curation. Results evidenced that, even for a well-studied model organism, source contents are still far from being as standardized as it would be desired and metadata varies considerably from source to source. Before designing any data integration pipeline, researchers should decide on the sources that best fit the purpose of analysis and be aware of existing conflicts/inconsistencies to be able to intervene in their resolution. Moreover, they should be aware of the limits of automatic integration such that they can define the extent of necessary manual curation for each application.

Keywords: *molecular biology; data integration; data standardization; data interoperability; semantic heterogeneity*

BACKGROUND

Life sciences research has been suffering a methodological shift with the emergence of high-throughput techniques [1, 2]. The amount of data generated by these experimental techniques has created new demands with respect to data management and accessibility. New systems biology studies require researchers to understand how interplay among a large number of biomolecular entities is orchestrated in order to achieve high-level cellular and physiological functions. Researchers need to compare and integrate a number of data, such as experimental data, data provided by different databases and additional information presented in literature. To do so, researchers face two problems: (i) to find and collect data scattered through multiple resources and (ii) to integrate data described in many different formats.

An outstanding number of public repositories of biological data has been developed to address these needs [3]. Many repositories engage particular types

Corresponding author. Anália Lourenço. Tel: +351 253 604 423; Fax: +351 253 678 986; E-mail: analia@deb.uminho.pt

Anália Lourenço is a Postdoctoral Research Associate in the IBB – Institute for Biotechnology and Bioengineering, Center of Biological Engineering, University of Minho. Her research interests include Genome-scale Model Reconstruction and Biological Network Analysis.

Sónia Carneiro is a PhD student in the IBB – Institute for Biotechnology and Bioengineering, Center of Biological Engineering, University of Minho, Braga, Portugal. Her research interests include Systems Biology analyses of metabolic and regulatory networks.

Miguel Rocha is an Assistant Professor at the Informatics Department of the University of Minho. His research interests include Bioinformatics, Natural Computation, Data Mining/Machine Learning.

Eugénio C. Ferreira is an Associate Professor at the IBB – Institute for Biotechnology and Bioengineering, Center of Biological Engineering, University of Minho, Braga, Portugal. His research interests include Bioprocess Systems Engineering.

Isabel Rocha is an Assistant Professor at the IBB – Institute for Biotechnology and Bioengineering, Center of Biological Engineering, University of Minho, Braga, Portugal. Her research interests include Systems Biology and Metabolic Engineering.

of data, such as biological sequences (e.g. GenBank [4], UniProt [5]), chemical compounds (e.g. Chemical Entities of Biological Interest (ChEBI) [6]), enzymatic information (e.g. BRaunschweig ENzyme DAtabase (BRENDA) [7] and SABIORK [8]), regulatory information (e.g. RegulonDB [9]) or 'omics' data (e.g. ArrayExpress [10], BioGrid [11], MINT [12] and IntAct [13]). Others depict the functional interactions of metabolic and regulatory pathways (e.g. BioCyc [14], Kyoto Encyclopedia of Genes and Genomes (KEGG) [15], and Reactome [16]) and integrate pathway data with high-throughput data (e.g. EcID [17]).

Data integration, however, is not trivial and requires data retrieval, parsing and pre-formatting. Syntactic and structural differences (differences related to data models such as relational databases, flat files and spreadsheets) lying in the data schemas that each source specifies are technical problems always present in data integration projects. In turn, semantic differences are expressed in the terminologies (vocabularies) recognized by the data schemas, which make it difficult to identify similar biomolecular entities across multiple sources.

Currently, to assist in data integration/interchange efforts, most databases maintain cross-links (also called cross-references or link-outs), i.e. links between their records and related records on external databases. For example, many sources keeping gene data associate to their records the corresponding Entrez Gene identifiers and a similar situation occurs with protein records and UniProt identifiers. Also, databases usually relate their records with a number of standard nomenclatures, thus providing controlled vocabulary. For example, often enzyme records include Enzyme Commission numbers (EC numbers) [18] and locus identifiers are associated to gene records. Notwithstanding, the set of standard nomenclatures and cross-references maintained by each source vary considerably from source to source, and each source has its own production and update cycles, differing in terms of actual contents.

Given that data source heterogeneity is unavoidable, and a single data model for all biomedical scenarios/problems is neither probable nor possible, much effort has been put on the development of data integration approaches/frameworks [19–21]. These include, among other, hypertext navigation and Web Services (e.g. SRS [22], Entrez [23] and BioMart [24]); data mediation and federation (e.g. KA-SB [25], TAMBIS [26], and BioMediator [27]); and data warehouses (e.g. Ondex [28], BNDB [29], Biowarehouse [30] and Columba [31]). An evaluation of the abilities of these approaches/frameworks bears many difficulties and is out of the scope of this article.

The aim of the present work is rather to evaluate the potential of basic elements of information (commonly present in most databases) for these integration approaches/frameworks (or any new ones) and thus, contribute to the discussion of (i) how challenging is it to integrate heterogeneous biological data and (ii) how straightforward is it to cross over similar contents from multiple data sources. In particular, the analysis is focused on the integration of data from the bacterium *E. coli* K-12 (whenever possible data were filtered for the sub-strain MG1655), for which public repositories keep considerable information, and which is at the heart of quite diverse studies (i.e. involving different biomolecular entities and/or demanding different levels of detail from the data). A project that requires such efforts in data gathering and integration is the reconstruction of the metabolic network of an organism [32] that, after validation, can be used for instance in metabolic engineering applications (e.g. [33]) and functional genomics. Clearly, any study that involves the generation and analysis of omics data also requires some level of data integration from public repositories.

In this study, and to facilitate the generalization of the approaches used, a very general view of the information flow from gene to protein to function was considered. In Figure 1 a scheme of that information flow is shown, together with the most important elements that were used to integrate the different biomolecular entities. Genes (identified mainly by numbers and names) can codify either for regulatory proteins (transcription factors) or enzymes. A Boolean rule might be needed to describe gene–protein encoding, since a single enzyme may be composed by two or more subunits that are codified by separate genes. Moreover, different genes can codify for enzymes with similar biochemical behaviour (isoenzymes), conferring redundancy to the systems. Proteins can be identified either by their CAS registry numbers or names. Enzymatic activities can be identified by EC numbers and have information associated with reaction reactants and products, among other data.

Commonly available information elements for the three main biomolecular entities, i.e. genes, proteins

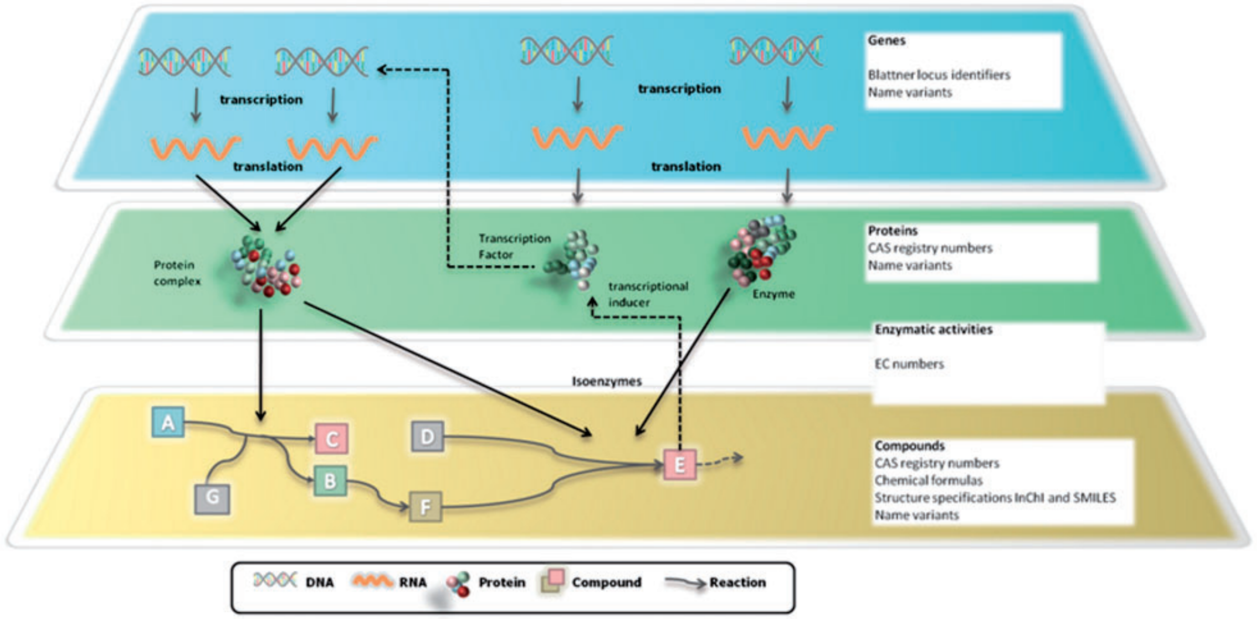


Figure 1: Information flow from gene to protein to function. The studied biomolecular entities (genes, proteins and compounds) are characterized by different information elements (in the white boxes on the right). Transitions between layers show the different interconnections between the entities.

and compounds, were evaluated by means of pair-wise source comparison. The circumstances leading to unsuccessful integration results (i.e. a high number of unresolved records) are then discussed. Following up, some strategies for the automatic identification and (whenever possible) resolution of data ambiguities are suggested. Finally, we outline the advantages of conciliating automatic data analysis and expert monitoring towards the maximisation of integration outcomes, in terms of the number of records and data quality, and the minimization of manual curation efforts.

MATERIALS AND METHODS

Data sources

In this work, different data sources containing information on *E. coli* K-12 (whenever possible for the sub-strain MG1655) were assessed. Overall, six data sources were covered by the study, including broad-scope, domain- and organism-specific data sources (Supplementary Table S2 provides an overview of the contents that were extracted from each source). The Chemical Entities of Biological Interest (ChEBI), EBI's freely available dictionary of chemical compounds, was included as an independent source of chemical data, namely terminology recommended by the IUPAC and the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) [6]. The NCBI's gene database, Entrez Gene, on fully-sequenced genomes [34] is a common database reference. Besides gene record cross-references, this source may assist on the identification of gene encoded proteins (the 'gene description' field). Likewise, the Universal Protein Resource (UniProt) Knowledgebase (UniProtKB), and particularly the UniProtKB/Swiss-Prot, provided for a fully curated protein sequence knowledgebase with extensive cross-references [5]. Besides protein details, UniProt records also provided the names and locus identifiers of the encoding genes, enabling additional gene integration.

The KEGG consists of several databases that encompass knowledge on molecular interaction networks (PATHWAY database), genes and proteins, generated by multiple genome sequencing projects (GENES databases), and the information about chemical compounds and chemical reactions that are relevant to cellular processes (LIGAND databases) [15]. Here, we inspected the data in: LIGAND/Compound (the chemical compound structures 'compound' file), GENES/Organisms (the 'E.coli.ent' file) and LIGAND/Enzyme (filtering the 'enzyme' file based on organism-specific gene coding information). EcoCyc differentiates from

KEGG since it is focused on the genome and biochemical machinery of *Escherichia coli* K-12 MG1655 [35]. It was expected that EcoCyc could provide additional metabolic information as well as extensive regulatory information (EcoCyc maintains a tight connection with RegulonDB, a database on transcriptional regulation in *E. coli* [9, 36]).

Finally, we chose a resource specialized in enzymatic data (e.g. kinetics, substrates/products, inhibitors/activators and cofactors). The BRaunschweig ENzyme DAtabase (BRENDA) is a manually curated and literature-based database that classifies reactions according to the EC system of the IUBMB Enzyme Nomenclature Committee [7]. We processed the available flat file to get *E. coli* related information. Also, given that no compound catalogue is publicly accessible, we collected compound names from substrate, product and cofactor fields in BRENDA records.

Source contents evaluation

We evaluated the data integration ability of cross-links and some standard nomenclatures, commonly present in most data sources (Supplementary Table S3), namely: the EC numbers [18] and the CAS registry numbers [37] for enzymes; the Blattner locus identifiers (commonly referred to as bnumbers) [38] for genes; and, the CAS registry numbers, the chemical formulas and the chemical structure specifications IUPAC International Chemical Identifiers (InChI) [39], and Simplified Molecular Input Line Entry Specification (SMILES) strings [40] for small molecules. Additionally, we also considered existing names, i.e. common names, synonyms, acronyms and abbreviations associated with biomolecular entity records. In particular, we evaluated a wide variety of names ranging from full systematic names and names following current recommendations of the IUPAC body on chemical nomenclature, conventional *E. coli* gene naming [41], and non-standard names commonly used by the biological community.

Gene, protein and compound data were analysed separately, classifying the elements of integration on the basis of their ability to unequivocally match biomolecular entity records across the different data sources. Specifically, the number of unmatched records was indicative of the degree of data discrepancy between sources, whereas the number of multiple match candidates evidenced the need for manual curation even when sources apparently have similar contents.

We tested the ability of each integration element by pairwise data source comparison, i.e. each data source was matched one-on-one with each of the other data sources. Given the comparison among two sources **A** and **B**, results were characterized as follows: (i) unique matches identified those elements that unambiguously related a record of source **A** with a record of source **B** and thus, can be automatically integrated; (ii) multiple matches indicated that the element used for integration could not unequivocally pair a record of source **A** with a record of source **B**, i.e. one record in source **A** was matched against more than one record in source **B** or one record in source **B** was matched against more than one record in source **A**; and (iii) non matches accounted for all records in **A** that could not be integrated. Two-way comparisons, i.e. matching records of source **A** to records of source **B** and records of source **B** to records of source **A**, evidenced that sources may keep similar but not exactly the same contents, highlighting the source matching direction that enhances data integration.

Furthermore, we complemented this evaluation by combining the elements that performed best. The idea was to assess if pairwise comparisons were able not only to characterize the interoperability of data sources, but also to support the design of the most adequate integration strategies. So, we followed a very simple procedure: (i) ranking the available elements by their integration ability, namely the number of unique matches, and (ii) applying element by element until there were not any more data to integrate or elements to support integration.

RESULTS

We considered a general systems-oriented scenario for *E. coli* K-12 MG1655, which requires the integration of gene, protein and compound data, across six data sources: an organism-specific data source keeping both metabolic and regulatory information—EcoCyc [35]; a data source that maintains metabolic information for multiple organisms—KEGG; and four domain-specific data sources—BRENDA for enzymatic activity, ChEBI for small molecule characterization, Entrez Gene [34] for genome data and UniProt for protein information. Several integration elements were explored. First, we evaluated the integration ability of the cross-links and then we assessed the ability of information elements coming from standard nomenclatures. Finally, we compared several alternative names, i.e. common names, synonyms, acronyms and abbreviations. Source contents are characterized in terms of the number of records, the type and extent of standard nomenclatures, the total and average number of name variants and the diversity and number of cross-links. A summary (Supplementary Table S1) and a detailed description of the contents of the data sources (at the section 'Data Sources' in the Web report) can be found in Supplementary Data.

The number of records per biomolecular entity is only indicative of the amount of source contents and should not determine integration outcomes by itself. For example, EcoCyc only keeps record of small molecules related to the metabolism and regulation of *E. coli* (1610 records) whereas ChEBI keeps a general (non-organism-dependent) repository of 17 445 small molecules. Likewise, KEGG is focused on enzyme information while UniProt and EcoCyc keep records of both regulatory and metabolic proteins. Still, as long as researchers are aware of the nature of source contents, they should easily interpret results and decide on the best strategy for a particular analysis.

In the next subsections, we outline the results of source pairwise comparison for the three types of biomolecular entities. The complete workflow is detailed in the Materials and methods section.

Source cross-linking

As illustrated in Table 1 and, in detail at the section 'Data Sources' in the Web report in Supplementary Data, even between major repositories, cross-linking varies considerably and it is often not bidirectional. In terms of gene records, KEGG supports extensive cross-linking to EcoCyc (98%), Entrez Gene (100%) and UniProt (94.76%), but only UniProt keeps links to KEGG (97%). Likewise, Entrez Gene is heavily linked (~98% of the records) to EcoCyc but EcoCyc linking to Entrez Gene is insignificant. In terms of proteins, KEGG records are fully linked to BRENDA and >95% of UniProt records are linked to EcoCyc records. Yet, there are no links (in any direction) between KEGG and UniProt, BRENDA and EcoCyc, or BRENDA and UniProt. In terms of compounds, KEGG and ChEBI organism-independent chemical repositories sustain similar low linking rates among themselves (~30%), EcoCyc is barely linked to ChEBI and 40% of EcoCyc records have no link to KEGG.

By only considering record cross-linking, we concluded that it is possible to almost fully (>98% of the records) integrate KEGG and EcoCyc gene information, KEGG and BRENDA enzyme information, and EcoCyc and UniProt protein information. However, metabolic data integration is hampered by the lack of an adequate number of cross-links between KEGG and EcoCyc for enzyme and compound information. Indeed, compound cross-linking was found insufficient for any of the analysed sources.

Gene and protein-specific elements

We considered Blattner standard identifiers (bnumbers), a special locus tag for *E. coli* genes, and gene names as information elements that might help in the integration of gene records. Gene integration results are presented in Figure 2 (first column) and, in detail, at the section 'Pairwise Evaluation->Genes' in the Web report and in Supplementary Figure S1. In most scenarios, locus tag identifiers performed extremely well, yielding >94% of unique matches, no

Table 1: Database cross-links for genes, proteins and compounds

	BRENDA	**ChEBI**	**EcoCyc**	**Entrez Gene**	**KEGG**	**UniProt**
BRENDA	1063 (E)	n.a.(E)	68% (E)	n.a. (E)	0 (E)	0 (E)
ChEBI	n.a. (C)	17445 (C)	0 (C)	n.a. (C)	32% (C)	n.a. (C)
EcoCyc	n.a.(G) 0 (E)	3% (C)	4477 (G) 5446 (P) 1610 (C)	13% (G)	0 (G) 0.09% (P) 52% (C)	0 (G) 78% (P)
Entrez Gene	n.a. (G)	n.a. (G)	98% (G)	4466 (G)	0 (G)	0 (G)
KEGG	100% (E)	34% (C)	98% (G) 0 (E) 0 (C)	100% (G) n.a. (E)	4466 (G) 726 (E) 15 403 (C)	95% (G) 0 (E)
UniProt	n.a. (G) 0 (E)	n.a. (C)	0 (G) 96% (P)	0 (G) n.a. (P)	97% (G) 0 (E)	4341 (G) 4342 (P)

The percentage of records with cross-links of a given source to the other sources under analysis are indicated. Each number represents pairwise percentage matches of sources **{**A, B**}**, i.e. the percentage of cross-links that each source in the *X*-axis maintains to the other sources per biomolecular entity type available (indicated in parenthesis as follows: C-compound, E-enzyme, G-gene and P-protein).

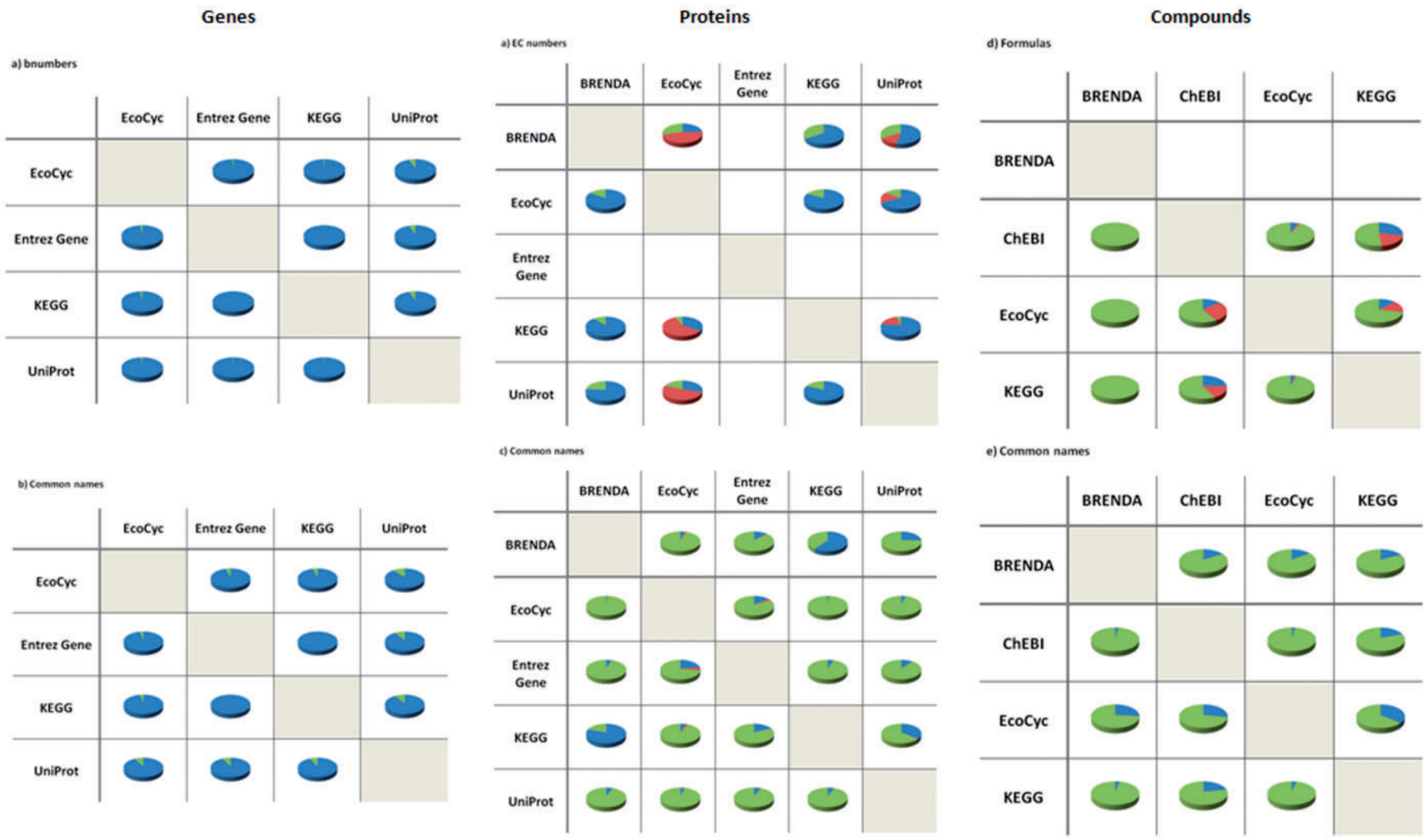


Figure 2: Record matches for genes, proteins and compounds. Two examples of the analysed integration elements are given for each biomolecular entity (results for the remaining elements are provided in Supplementary Data). Each table refers to the match results of an element of information under analysis. Results are plotted comparing the contribution of each match value to the total across categories, i.e. unique matches (in blue), multiple matches (in red) and non matches (in green). The absence of a pie indicates that the source does not keep that kind of contents whereas a pie of only non-matches (i.e. completely green) means that the source keeps those contents but the other source does not.

multiple match candidates (i.e. they unequivocally identify the genes) and <6% of non matches (mainly regulatory or predicted protein/pseudogene information).

Record matching based on gene names also presents good results with >90% of unique record matches. However, when compared to common names, the comparison of all name variants does not bring significant improvements (2–4% additional record matches) and depicts a higher number of elements to be curated. Indeed, genes often share synonyms and thus, multiple candidates are presented for record match (e.g. in EcoCyc and KEGG, the genes *arg*A and *arg*D share the name variant Arg1).

Next, we evaluated the concordance of the sources in terms of EC numbers, CAS registry numbers and different variants of protein names. Outcomes are reported in Figure 2 (second column) and, in detail, at the section 'Pairwise Evaluation->Proteins' in the Web report and in Supplementary Figure S2. Considering EC numbers, results suggest some discrepancy in terms of enzyme information. Over 15% of EC numbers are unfamiliar in most scenarios, either because one of the sources has not associated them to *E. coli* or they are incomplete (usual for predicted protein functions, e.g. 1.1.-.-). Additionally, since often EC numbers qualify the enzymatic activity of more than one protein, there is a considerable number of multiple matches (e.g. in EcoCyc-related scenarios, ~50% of the EC numbers under evaluation). Any attempt to automatically resolve these multiple matches would be recurring to gene coding information, but not always enzyme-related contents are associated with such information (e.g. there is no gene information associated with KEGG and BRENDA records).

Even though only two sources in the study, BRENDA and KEGG, keep record of CAS registry numbers for proteins, results show that this might be a valid element of integration. There is a good number of CAS-based matches (>57% and 75% of records respectively) and almost no multiple matches. In fact, the difference of record matches when integrating BRENDA and KEGG by CAS identifiers or EC numbers is ~6%. Thus, when available, CAS indexing might be considered an alternative to EC numbers. On the other hand, in most cases, record matching based on common protein names and name variants performed poorly (>95% of record non-matches). Names containing special characters, such as Greek letters, apostrophes, slashes and super/subscripts, are often encoded differently. For example, 'σ70' in 'EcoCyc:RPOD-MONOMER' and 'Sigma-70' in 'UniProt:P00579', and 'aminobutyraldehyde dehydrogenase' in 'BRENDA: 1.2.1.19' and 'γ-aminobutyraldehyde dehydrogenase' in 'EcoCyc: G6755-MONOMER'. Similarly, 'DsbC$_{\text{oxidized}}$' and 'disulfide interchange protein dsbC' (DSBCOX-MONOMER record in EcoCyc and P0AEG6 record in UniProt, respectively) and 'EntS MFS transporter' and 'Enterobactin exporter entS' (DSBCOX-MONOMER record in EcoCyc and P24077 record in UniProt, respectively) correspond to the same proteins. Also, there is the frequent use of general names for 'similar' proteins. For instance, in EcoCyc, it is common that a given complex and some of its monomers share names (e.g. 'Alanyl-tRNA synthetase' stands for both ALAS-CPLX and ALAS-MONOMER records).

Compound-specific elements

In terms of compound data integration, we inspected chemical formulas, SMILES and InChI chemical structure representations, CAS registry numbers and, once again, available biomolecular entity names. Outcomes are reported in Figure 2 (third column) and, in detail, at the section 'Pairwise Evaluation->Compounds' in the Web report and in Supplementary Figure S3.

Chemical formulas (>52% of record non-matches), SMILES and InChI-based record matching perform quite poorly (>96% and 75% of EcoCyc record non-matches, respectively). Similarly, we observed false positive matches in CAS-based matching scenarios. For example, the CAS number '2009-24-7' is associated with the compound 'xanthotoxol' in CHEBI:15709 and the compound dTDP-glucose in 'EcoCyc: DTDP-D-GLUCOSE'.

In most cases, we observed that the inclusion of name variants increases the number of record matches. Indeed, the number of unique matches is raised 20% in the EcoCyc-KEGG scenario (resolving almost 57% of the records) and improvements of >7% were achieved in BRENDA-KEGG and BRENDA-EcoCyc scenarios. However, the inclusion of a specialized repository such as ChEBI seems to be of value only when integrating KEGG data. In the case of BRENDA or EcoCyc, the number of unique matches is not greater than in the rest of scenarios, making this resource of little assistance in such cases.

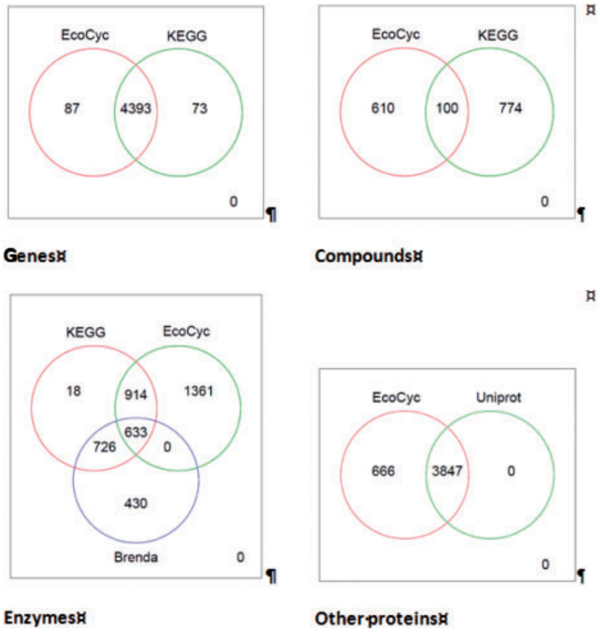
Combining elements and enhancing integration

Based on pairwise source comparisons, practitioners could easily identify the most adequate elements for integration and the main issues affecting automatic matching. Except for the gene scenario, in no other case an element of integration performs well enough to support the process by itself.

KEGG gene records are closely connected to Entrez Gene records (100% of the records) and thus, there is no apparent gain in their integration. In turn, the integration of KEGG and EcoCyc may be considered of interest since EcoCyc complements metabolic pathway data with regulatory data. By cross-linking, most gene records were resolved (4374 records), leaving 179 records to be matched. Locus identifiers could not provide additional matches, since all possible locus-based matches had already been solved by cross-linking. Unique gene names were used to resolve records (19 records) that either do not present link or locus data or do not agree on these elements, as it is the case of pseudo-genes (e.g. EcoCyc: G6211 and KEGG: b4579). Such strategy resulted in an overall of 4393 record matches, leaving 73 KEGG records and 87 EcoCyc records to be manually curated (Figure 3, left upper corner).

In the integration of protein records, data were divided in two sub-sets: enzymes (involving the EcoCyc, KEGG and BRENDA data sources) and other proteins involved in processes like gene regulation and cell signalling (included in EcoCyc and UniProt data sources). Data on non-enzymatic proteins was quite successful, relying basically on cross-linking (3252 record matches by cross-linking and 36 record names by unique name) (Figure 3, right bottom corner). All records in UniProt found a match whereas 666 EcoCyc records require manual curation (namely records reporting the presence of two component systems, lipoproteins and transporters in *E. coli*).

Enzyme data integration, however, represented a challenge. Available elements could not unequivocally identify enzyme entities, and therefore gene coding information was privileged in order to integrate these records. This information enabled 845 unique matches between KEGG and EcoCyc. Unique EC number information in EcoCyc (i.e. EC numbers relating to one enzyme acting on one particular reaction) resolved 21 records more and enzyme names contributed with 48 additional matches. The integration of BRENDA records was devised to enrich KEGG and EcoCyc characterisation of enzymatic activities (e.g. information on kinetic parameters, metabolic regulators or cofactors). It was possible to extend the data for 633 records out of the 914 records previously integrated (Figure 3, left bottom corner). The 430 BRENDA records left to be manually curated are associated with more than one reaction record in EcoCyc or relate to activities not yet documented in EcoCyc or KEGG.



Genes¤ Compounds¤

Enzymes¤ Other proteins¤

Figure 3: Comparison of source identities and specificities when combining elements to enhance integration.

By far, compound integration was the most challenging process. None of the inspected elements performed well enough (>50% of non-matches in any scenario) and there are major discrepancies in terms of the nomenclatures in use. Besides, the inclusion of ChEBI, a source rich in standard nomenclatures, does not improve results. Even when combining cross-links (830 records), CAS registry numbers (28 records) and unique names (142 records), which provide the best pairwise results, significant manual curation is required (>600 EcoCyc and 700 KEGG records) (Figure 3, right upper corner).

DISCUSSION

Given the diversity of source contents, the structural and semantic differences, it is not feasible to devise a single integration approach that will work for every possible scenario. Therefore, specific requirements and accomplishments need to be defined *a priori* for a particular problem before any integration attempts are initiated. In this work, we analysed the problem of integrating data on the bacterium *E. coli* K-12 using commonly used sources and biological entities. Although for some applications a complete analysis of unmatched records may not be necessary, this study provides some clues on the losses and gains of information obtained by including specific data sources and using different integration elements, helping in decision making when devising an integration strategy for a particular problem.

Even though existing data integration projects had certainly required substantial interaction between computer scientists and biology practitioners, there is little information on how they proceeded and what they have learned. In most cases, data integration papers are focused on discussing the strategies technologically (e.g. virtual versus physical integration, distributed versus centralized repository) and the benefits that the new repository will bring to the domain (e.g. access to new kinds of data or ability to query across multiple sources of data).

Although the proposed scenario is necessarily a simplification of systems-level scenarios (e.g. for genome-scale reconstruction or biological network modelling), we could still appreciate how data integration processes are hampered by structural inconsistencies and the lack of adequate means of standardisation. Regardless the biomolecular entity under evaluation, cross-links are considered the most reliable element of integration since they are maintained by database curators. Yet, cross-linking is not enforced by any organisation, body or initiative, i.e. database managers decide whether or not to support this effort and, if so, to which sources. Moreover, cross-link update is dependent of source-specific production cycles and release schedules and may be compromised when external repositories eliminate or recycle record identifiers as means of internal refactoring. Indeed, data interoperability is tighter between collaborating projects. For example, it is well-known that EcoCyc [35] and RegulonDB [9], two resources specialized on *E. coli*, sustain periodic cross-loads [9, 42]. In turn, the pairwise comparison of EcoCyc and KEGG contents has exposed significant discrepancies whilst the two resources aim to provide comprehensive metabolic pathway information on *E. coli*.

In the present study, the selected elements produced a high number of unique matches in gene scenarios, but in the rest of the scenarios no element performed well enough by itself and even the combination of several elements could not reduce manual curation reasonably. In particular, enzyme and compound data, i.e. the ground basis of the metabolic machinery, required the curation of >50% of the records.

Challenges at gene, protein and compound levels in *E. coli*

Regarding gene information, sources are quite consistent in terms of locus identifier information. However, it is important to bear in mind that this is not necessarily the case for all organisms or data sources. Locus tags are assigned to particular zones/genes in a genome and the same locus tag is used for all components of a single gene (e.g. all of the exons, mRNA and gene features for a particular gene share the same locus tag). While studying the machinery of an organism, locus tags can be modified and thus, some confusion may arise from the (temporary) use of deprecated tags. Furthermore, locus tag format may vary between different strains of the same organism (e.g. in *E. coli* the gene relA has 504 the locus tags ECK2778, b2784 and JW2755 for strains K-12, MG1655 and W3110, 505 respectively). Not to mention the fact that the nomenclature used for the genetics of eukaryotic organisms has not yet been as well formalized as that for bacteria and bacteriophages. On the other hand, the good performance of name matching (in many cases common name matching is almost as good as locus tag matching) is justified by the consistent use of the Demerec name format [41, 43], which uses a unique three-letter abbreviation intended to suggest a function, followed by a capital letter to distinguish different genes related to the same function.

Regarding compounds, the widespread use of non-standard complex nomenclature represents a major challenge to the integration. So, we investigated the potential of chemical structure representations and chemical formulas, which do not unequivocally identify compounds in the first place, to work around this lack of standardisation. Likewise, we considered the use of an additional broad-scope chemical source such as ChEBI as a means to enhance nomenclature mapping. However, none of the inspected elements performed well enough (>50% of non-matches in any scenario) showing that there are major discrepancies in terms of the nomenclatures in use.

SMILES and InChI specifications describe the structure of chemical molecules differently, but present a number of similar problems to data integration: (i) SMILES is unique for each structure, but it is dependent on the canonicalisation algorithm used (e.g. ethanol is represented as C(O)C in EcoCyc:ETOH and CCO in CHEBI:16236, and water is represented as 'O' in EcoCyc:WATER and '[H]O[H]' in CHEBI:15377); (ii) the flexible number of layers of information (e.g. the atoms and their bond connectivity) of the InChI representation may vary from source to source (e.g. the representation of xanthosine molecule in EcoCyc: XANTHOSINE and ChEBI:18107 records); (iii) compounds having akin structure share SMILES (e.g. the singlet diooxygen and dioxygen molecules in ChEBI, CHEBI:26689 and CHEBI:15379 records respectively) and InChI (e.g. L-RIBULOSE-5-P, RIBULOSE-5P and XYLULOSE-5-PHOSPHATE records in EcoCyc).

Chemical formulas identify the constituent elements (by the corresponding chemical symbol) and indicate the number of atoms of each element found in each molecule of a compound. Apparently, the sources that we analysed, i.e. ChEBI, EcoCyc, KEGG and BRENDA, use the same empirical formula representation, i.e. a simple expression of the relative number of each type of atom or ratio of the elements in the compound. Yet, at a closer look, formula conventions are quite different: the chemical formula of the compound 'nitrite' on 'KEGG:C00088' and 'EcoCyc:NITRITE' differs on the number of hydrogen atoms (HNO_2 and NO_2, respectively); and, the formula of 'Copper Sulfate' is 'O_4S' in 'EcoCyc: CUO_4S' and 'CuO_4S' in ChEBI (ChEBI:23414). Moreover, poor name-based record matching was expected given that it is widely recognized that compounds exhibit a proficiency of often non-standard names [6, 44, 45].

Regarding proteins, repositories usually include different elements of function/activity and structure characterisation as well as a number of alternative names. Here, it is of paramount importance to realize that EC numbers are a standard numerical classification for enzyme catalysed reactions, i.e. strictly speaking they do not specify enzymes, but rather the chemical reactions they catalyse. KEGG and BRENDA are quite consistent in terms of EC numbers, because these repositories index enzymatic activities rather than individual enzymes. However, EcoCyc keeps record of both reactions (with an EC number associated) and enzymes (with the corresponding gene coding) establishing their association at an intermediate level where an enzyme may be related to a number of reactions and a reaction may encompass the activity of several enzymes (isoenzymes). As such, to integrate both levels of information and unequivocally identify the interplaying enzymes, researchers must look for gene coding information, i.e. link the enzymatic reactions to the genetic coding.

Reaction stoichiometry and, in particular, the list of involved reagents and products represent another challenge to data integration. In some cases, an official EC reaction equation is attributed to various chemical reactions with alternative substrates (e.g. EcoCyc: GDPPYPHOSKIN-RXN). Also, some equations identify compound classes/families rather than actual compounds and may use the 'n/m' convention to show an unknown quantity. For instance, the equation 'an alcohol + NAD+ = an aldehyde or ketone + NADH + H+' is associated to the alcohol dehydrogenase (EC number 1.1.1.1) and 'NAD^+ + (deoxyribonucleotide)n + (deoxyribonucleotide)m = AMP + nicotinamide nucleotide + (deoxyribonucleotide)n+m' is associated to the DNA ligase (NAD^+) (EC number 6.5.1.2). On top of all this, there is the identification of each of the compounds. Often enough, sources keep a descriptive field, where the reaction is described in terms of the common names of the involved compounds, and/or a detailed field that breaks down the equation into compounds fully linked to the corresponding records. This linkage is an internal procedure, i.e. source identifiers are in use rather than any standard, which implies that it is absolutely necessary to map compounds between sources before integrating reactions. However, none of the evaluated standards seems to perform adequately enough for this purpose. Chemical structure representations and formulas may provide intuitive support to human curation but they are unable to unequivocally identify compounds in automatic processes.

In general, biomolecular entity names may not be considered a reliable element for data integration due to the proficiency in synonyms and homonyms and the coding of special characters. Genes/proteins closely related usually share a number of common names (e.g. the name 'arg1' is common to two genes in both EcoCyc and KEGG), and a similar situation happens with compounds belonging to the same chemical family (e.g. the name 'alanine' is shared by three compounds in ChEBI). Furthermore, protein and compound names are often composed by hyphen- and apostrophe-based long-forms and/or special characters, such as Greek letters, italics and superscripts/subscripts. Their encoding varies from source to source, ranging from plain 'flat' text (e.g. 'NAD+' and 'UDP-alpha-D-glucose' in KEGG: C00003 and KEGG: C00029, respectively) to HTML-alike formatting (e.g. 'L-&alpha;-alanine' and 'NAD<sup>+</sup>'in EcoCyc: L-ALPHA-ALANINE and EcoCyc: NAD, respectively).

The need for systematic approaches to source comparison

Usually, the development of a data integration framework is motivated by a particular problem that, for some reason, cannot be addressed conveniently by existing approaches. Technological options are varied, ranging from source record mapping to a common data model (i.e. data is not dissociated from original sources) or to full record integration.

Frameworks such as Biowarehouse [30] (currently supporting EcoliHub repository), BNDB [29] or Ondex [46] are freely available to anyone in need of data integration. However, it is not easy for biologists that are not familiar with integration approaches to assess the implications of using a given framework/approach to meet their analysis. Framework development is focused on implementing novel integration heuristics (e.g. using cross-links, processing names or comparing sequence similarity) and providing enhanced means of visualisation. It is unusual for frameworks to interact with biologists towards the examination of challenges and the assessment of alternative strategies. So, often biologists end up picking the most familiar data sources without considering whether they are in fact the best for their particular analysis.

The proposed systematic pairwise source comparisons are a very simple, yet quite practical and highly extensible means of bridging this gap. Computationally speaking, the approach is inexpensive and can be easily integrated in any framework. Available integration frameworks deal with several source metadata (implementing specific loaders) and they are able to identify the elements shared by the data sources. Therefore, it is feasible for those frameworks to evaluate different integration scenarios before committing with a particular strategy.

Regarding the analysis of data, results expressed in terms of unique, candidate and inexistent matches provide immediate insights, without requiring any computational abilities or technological knowledge from biologists. By pointing out the number of record matches and, in particular, differentiating between unique and multiple match candidates, biologists will become aware of structural heterogeneity and nomenclature challenges. Also, they will be able to estimate the information losses/inaccuracies and integration costs (automatic integration versus manual curation) associated to each potential element of integration.

CONCLUSIONS

Considering the wide scope of applications that benefit from the analysis of large amounts of data, many have been the efforts focused on developing new and comprehensible ways of data integration. Currently, a number of general purpose frameworks are available to support the design and implementation of workflows for the integration and visualization of complex datasets. Yet, most works fail to debate a previous, crucial step of the process: the selection of the most adequate data sources for the analysis and the elements of integration across sources.

The purpose of this work has not been to review all the available technologies and strategies for integration, but to illustrate, using a familiar set of data sources, why the selection and integration of the most adequate data sources are not trivial tasks, as well as to raise awareness of some of the challenges involved. We explored the automatic integration of contents from several well-known repositories that keep genome and biochemical information for the bacterium *E. coli* K-12. Our aim was to present what we see as a systematic discussion of the strengths and weaknesses of common integration elements, many of which have not been discussed previously. Our results reflect the lack of standardisation of common biological contents even for well-studied organisms. It is acknowledged that data standardisation and interoperability efforts are in action, but they lag behind what is expected from them. For some entities (enzymes and compounds), none of the elements of integration performed well enough by itself and even the combination of several elements could not be considered satisfactory.

Besides the clear problems for systems biology applications such as metabolic and regulatory reconstructions, the challenges exposed in this article are already posing significant difficulties when analysing data originated from the several omics technologies. In fact, being non-biased techniques, the results obtained need to be analysed in the scope of the corresponding metabolic or regulatory pathways. While transcriptomic and proteomic experiments are easily linked with existing databases, data originated from the emergent field of metabolomics face the problems discussed above for metabolic compounds. For example, in GC-MS experiments, compound identification is performed using dedicated commercial databases in which non-standard complex nomenclature is used, making it quite difficult to integrate these results with data available in databases such as KEGG.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- We explored the automatic integration of contents from several well-known repositories that keep genome and biochemical information for the bacterium *E. coli* K-12.
- Our aim was 2-fold: to present what we see as a systematic discussion of the strengths and weaknesses of common integration elements, many of which have not been discussed previously and to suggest some integration measures that would enable biologists to have an active intervention in the definition of new pipelines.

Acknowledgements

Both A.L. and S.C. have made equal substantive contributions to this manuscript. A.L. and S.C. designed the experiments and defined the integration strategies to be studied. A.L. and M.R. carried out computational work. S.C., I.R. and E.C.F. were responsible for manual curation. All authors participated in the preparation of the manuscript. All authors read and approved the final manuscript.

FUNDING

SYSINBIO, an European Coordination and Support action (call FP7-KBBE-2007-1) in the field of model driven metabolic engineering; Portuguese FCT (Fundação para a Ciência e Tecnologia) funded MIT-Portugal Program in Bioengineering (MIT-Pt/BS-BB/0082/2008); PhD grant from FCT (ref. SFRH/BD/22863/2005) to S.C.

References

1. Ge H, Walhout AJM, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* 2003;**19**:551–60.
2. Kitano H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr Genet* 2002;**41**:1–10.
3. Cochrane GR, Galperin MY. The 2010 nucleic acids research database issue and online database collection: a community of data resources. *Nucleic Acids Res* 2010;**38**:D1–4.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al*. GenBank. *Nucleic Acids Res* 2008;**36**:D25–30.
5. Bairoch A, Apweiler R, Wu CH, *et al*. The universal protein resource (UniProt). *Nucleic Acids Res* 2005;**33**:D154–9.
6. Degtyarenko K, De Matos P, Ennis M, *et al*. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;**36**:D344–50.
7. Barthelmes J, Ebeling C, Chang A, *et al*. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* 2007;**35**:D511–4.
8. Krebs O, Golebiewski M, Kania R, *et al*. SABIO-RK: a data warehouse for biochemical reactions and their kinetics. *J Integr Bioinform* 2007;**4**:49–58.
9. Salgado H, Gama-Castro S, Peralta-Gil M, *et al*. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 2006a;**34**:D394–7.
10. Parkinson H, Kapushesky M, Kolesnikov N, *et al*. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 2009;**37**:D868–72.
11. Breitkreutz BJ, Stark C, Reguly T, *et al*. The BioGRID interaction database: 2008 update. *Nucleic Acids Res* 2008;**36**:D637–40.
12. Ceol A, Chatr AA, Licata L, *et al*. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* 2010;**38**:D532–9.
13. Aranda B, Achuthan P, Alam-Faruque Y, *et al*. The IntAct molecular interaction database in 2010. *Nucleic Acids Res* 2010;**38**:D525–31.
14. Caspi R, Foerster H, Fulcher CA, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2008;**36**:D623–31.
15. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
16. Matthews L, Gopinath G, Gillespie M, *et al*. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;**37**:D619–22.
17. Andres LE, Ezkurdia I, Garcia B, *et al*. EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Res* 2009;**37**:D629–35.

18. Webb EC. Nomenclature Committee of the International-Union-Of-Biochemistry (Nc-Iub) - Enzyme Nomenclature - Recommendations 1984 - Supplement-2 - Corrections and Additions. *Eur J Biochem* 1989;**179**:489–533.

19. Köhler J. Integration of life science databases. *Drug Discovery Today: BIOSILICO* 2004;**2**:61–9.

20. Philippi S, Kohler J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet* 2006;**7**:482–8.

21. Stein LD. Integrating biological databases. *Nat Rev Genet* 2003;**4**:337–45.

22. Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* 1996;**266**:114–28.

23. Geer RC, Sayers EW. Entrez: making use of its power. *Brief Bioinform* 2003;**4**:179–84.

24. Haider S, Ballester B, Smedley D, *et al*. BioMart Central Portal–unified access to biological data. *Nucleic Acids Res* 2009;**37**:W23–7.

25. Roldan-Garcia MD, Navas-Delgado I, Kerzazi A, *et al*. KA-SB: from data integration to large scale reasoning. *BMC Bioinform* 2009;**10**(Suppl 1):S5.

26. Stevens R, Baker P, Bechhofer S, *et al*. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 2000;**16**:184–5.

27. Donelson L, Tarczy-Hornoch P, Mork P, *et al*. The BioMediator system as a data integration tool to answer diverse biologic queries. *Stud Health Technol Inform* 2004; **107**:768–72.

28. Lysenko A, Hindle MM, Taubert J, *et al*. Data integration for plant genomics – exemplars from the integration of Arabidopsis thaliana databases. *Brief Bioinform* 2009;**10**: 676–93.

29. Kuntzer J, Backes C, Blum T, *et al*. BNDB - the biochemical network database. *BMC Bioinform* 2007;**8**:367.

30. Lee TJ, Pouliot Y, Wagner V, *et al*. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinform* 2006;**7**:170.

31. Rother K, Muller H, Trissl S, *et al*. COLUMBA: Multidimensional data integration of protein annotations. *Data Integr Life Sci Proc* 2004;**2994**:156–71.

32. Rocha I, Forster J, Nielsen J. Design and application of genome-scale reconstructed metabolic models. In: Osterman AL, Gerdes S, (eds). *Microbial Gene Essentiality: Protocols and Bioinformatics. Methods in Molecular Biology*;**vol. 416**. New York: Humana Press, 2008:409–31.

33. Stephanopoulos G, Aristidou A, Nielsen J. *Metabolic Engineering*. San Diego, CA: Academic Press, 1998.

34. Maglott D, Ostell J, Pruitt KD, *et al*. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005;**3**:D54–8.

35. Keseler IM, Collado-Vides J, Gama-Castro S, *et al*. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* 2005;**33**:D334–7.

36. Huerta AM, Salgado H, Thieffry D, *et al*. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res* 1998;**26**:55–9.

37. Odette RE. Cas data-base. *Pure Appl Chem* 1977;**49**: 1781–92.

38. Blattner FR, Plunkett G, Bloch CA, *et al*. The complete genome sequence of Escherichia coli K-12. *Science* 1997; **277**:1453–74.

39. Heller SR, Stein SE, Tchekhovskoi DV. InChI: open access/open source and the IUPAC international chemical identifier. *Abstr Papers Am Chem Soc* 2005;**230**:U1025–6.

40. Weininger D. Smiles, A chemical language and information-system. 1. Introduction to methodology and encoding rules. *J Chem Inform Comput Sci* 1988;**28**:31–6.

41. Seringhaus MR, Cayting PD, Gerstein MB. Uncovering trends in gene naming. *Genome Biol* 2008;**9**:401.

42. Salgado H, Santos-Zavaleta A, Gama-Castro S, *et al*. The comprehensive updated regulatory network of Escherichia coli K-12. *BMC Bioinform* 2006b;**7**:5.

43. Riley M, Abe T, Arnaud MB, *et al*. Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res* 2006;**34**:1–9.

44. Klinger R, Kolarik C, Fluck J, *et al*. Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics* 2008;**24**: 268–76.

45. Kolarik C, Klinger R, Friedrich CM, Hofmann-Apitius M, Fluck J. Chemical names: terminological resources and corpora annotation. *Workshop on Building and Evaluating Resources for Biomedical Text Mining* (6th edition of the Language Resources and Evaluation Conference) 2008.

46. Kohler J, Baumbach J, Taubert J, *et al*. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* 2006;**22**:1383–90.