Graph-based sequence annotation using a data integration approach

Robert Pesch[1], Artem Lysenko[2], Matthew Hindle[2], Keywan Hassani-Pak[2], Ralf Thiele[1], Christopher Rawlings[2], Jacob Köhler[3] and Jan Taubert[2,4]

[1]Department of Computer Science, Bonn-Rhein-Sieg University of Applied Sciences, Germany

[2]Department of Biomathematics and Bioinformatics, Rothamsted Research, UK

[3]Protein Research Group, University of Tromsø, Norway

Summary

The automated annotation of data from high throughput sequencing and genomics experiments is a significant challenge for bioinformatics. Most current approaches rely on sequential pipelines of gene finding and gene function prediction methods that annotate a gene with information from different reference data sources. Each function prediction method contributes evidence supporting a functional assignment. Such approaches generally ignore the links between the information in the reference datasets. These links, however, are valuable for assessing the plausibility of a function assignment and can be used to evaluate the confidence in a prediction. We are working towards a novel annotation system that uses the network of information supporting the function assignment to enrich the annotation process for use by expert curators and predicting the function of previously unannotated genes. In this paper we describe our success in the first stages of this development. We present the data integration steps that are needed to create the core database of integrated reference databases (UniProt, PFAM, PDB, GO and the pathway database AraCyc) which has been established in the ONDEX data integration system. We also present a comparison between different methods for integration of GO terms as part of the function assignment pipeline and discuss the consequences of this analysis for improving the accuracy of gene function annotation.
The methods and algorithms presented in this publication are an integral part of the ONDEX system which is freely available from `http://ondex.sf.net/`.

1 Introduction

A good summary of the problems inherent in the reliable annotation and re-annotation of genome sequence data has been recently provided by (Salzberg, 2007). He describes the two key challenges as being the initial prediction of an accurate gene model from the raw genomic sequence and then the assignment of an annotation by sequence comparison with public databanks (e.g. GenBank). Salzberg highlights the difficulty of avoiding false gene function assignments inferred from incorrect reference database annotations. The topic of this paper is the development of an alternative approach to the assignment of reliable gene function based on reference databases. Our particular focus is on alleviating the problem caused by the propagation of false functional inferences, which is a major source of the many anecdotal accounts of

[4]To whom correspondence should be addressed. E-mail: `jan.taubert@bbsrc.ac.uk`

incorrect annotations in the sequence databases. Our proposed solution is to make explicit the interactions between the annotations from various reference databases by using semantic data integration to create networks of links among the related concepts and entities from the reference databases. These networks are then linked to the genes using sequence analysis methods. The annotation attached to a gene then becomes a network or graph whose structure and content can be analysed or visualised to explore the consistency of the biological information supporting the overall annotation.

An important factor, that will determine the long term success of this approach, is selecting the best sources of functional data from among reference databases and the best methods for integrating these data. We have therefore begun to make quantitative comparisons between annotation steps and we present some preliminary results. Before this analysis can take place, it is important to set out the principles behind our approach to semantic data integration and the methods we have used to achieve it.

1.1 Motivation

Bringing biological data together coherently to extract additional meaning is a major undertaking for any systems biology project. The development of biological thesauri and classification systems (ontologies) continue to make it easier to link between components of different databases. For example, by exploiting more consistent nomenclatures and using accepted lists of synonyms for biological processes and structures. This, however, only solves part of the problem of data integration because biological components can be related in many different ways. For example by taking part in a particular reaction, performing a certain function within a specific location or being part of a more complex structure. This information needs to be captured and classified accurately for it to be useful in data integration. Similarly, information about the provenance of data can be important in subsequent interpretations of any results. New types of information, such as descriptions of biological processes and pathways for metabolism and information flow, are also emerging in databases that are valuable for linking among databases. Many of these have been created by extracting information from the scientific literature to form as the basis of predictive dynamic models and simulations of biological systems. They also use complex representations that challenge traditional database systems.

1.2 Approaches to Data Integration

Different approaches to database integration have been proposed and can be characterised as being based on principles such as warehousing, federation, flat-file indexing or frameworks for data collection (Köhler, 2004). In addition to standard database approaches, graph and ontology oriented approaches are being used with increasing success. While some of these graph based frameworks can be found in commercial products such as VTT (Gopalacharyulu, et al., 2005), ChipInspector / Bibliosphere Pathway Edition (from Genomatix), Phylosopher (from Genedata), ExPlain (from BIOBASE) and PathwayStudio (Nikitin, et al., 2003) (from Ariadne Genomics), some non-commercial systems are also available free of charge such as PathSys (Baitaluk, et al., 2006), BN++ (Küntzer, et al., 2006) and the ONDEX system (Köhler, et al., 2006; Köhler, et al., 2004). Several projects have also followed a similar approach using the principles established for the semantic web such as SWEDI (Post, et al., 2007) or BioDASH (Neumann and Quan, 2006).

For data integration we use the ONDEX system which takes an ontology graph-based approach. ONDEX uses a range of algorithms and mapping methods, suitable for identification and linking of equivalent and related data entries from a wide variety of data sources.

In this study, the ONDEX system was used to integrate protein sequence data from UniProt (Apweiler, et al., 2004) with protein structures from the Protein Data Bank (PDB) (Sussman, et al., 1998), protein family assignments derived from the use of the PFAM database (Sonnhammer, et al., 1997), biological pathway data from AraCyc (Mueller, et al., 2003) and terms from the Gene Ontology (GO) (Ashburner, et al., 2000). These data sources were chosen as they contain some of the richest sources of protein function information and therefore are a good basis for evaluating the potential of a graph-based approach to gene annotation. The use of the GO annotation data was considered key to the later evaluation of the different approaches for mapping protein sequence to functional annotations.

2 Data Integration

A typical data integration pipeline for the ONDEX system is summarised in Figure 1 and consists of the following three steps: 1) parsing different data sources into the generalized object data model of ONDEX; 2) identifying equivalent and related entries and creating new relations between them using integration methods; 3) analysing the integrated data using client tools, for example, the ONDEX Visualisation Tool Kit (OVTK).

2.1 Data import

The first step in the development of an integrated data resource using ONDEX consists of parsing data sources into the ontology graph-based data model. Parsers play an important role in (re)modelling heterogeneous data for ONDEX. We have created custom parsers for biological databases including: AraCyc (Mueller, et al., 2003), GO (Ashburner, et al., 2000), GOA (Camon, et al., 2003), UniProt (Apweiler, et al., 2004) and PFAM (Sonnhammer, et al., 1997). Furthermore, parsers have been developed for a number of general-purpose file formats, including FASTA (Pearson, 1990), OBO (Smith, et al., 2007) and the flat files describing the cross references between GO terms and database entries made available from the GOA database.

Imported data are represented as a graph of concepts (nodes) and relations (edges). By analogy with the use of ontologies for knowledge representation in computer science, concepts correspond to real world entities. Relations are used to represent the way in which concepts are semantically linked to each other. Furthermore, concepts and relations may have attributes and optional characteristics attached to them. A formal description of the data structure of the ONDEX system is presented by (Köhler, et al., 2006). During data import, consistency checks on the data are performed, e.g. concept names are lexicographically normalized.

The current ONDEX graph based data structure is implemented using an object based data model which makes use of the Berkley DB Java edition (Oracle-Corp., 2006) for persistent data storage. ONDEX uses Lucene (http://lucene.apache.org) for full text search of integrated data sets. Each step in the process of data integration presented in Figure 1 is controlled by an ONDEX workflow enactor, which processes user-defined scripts written in XML.

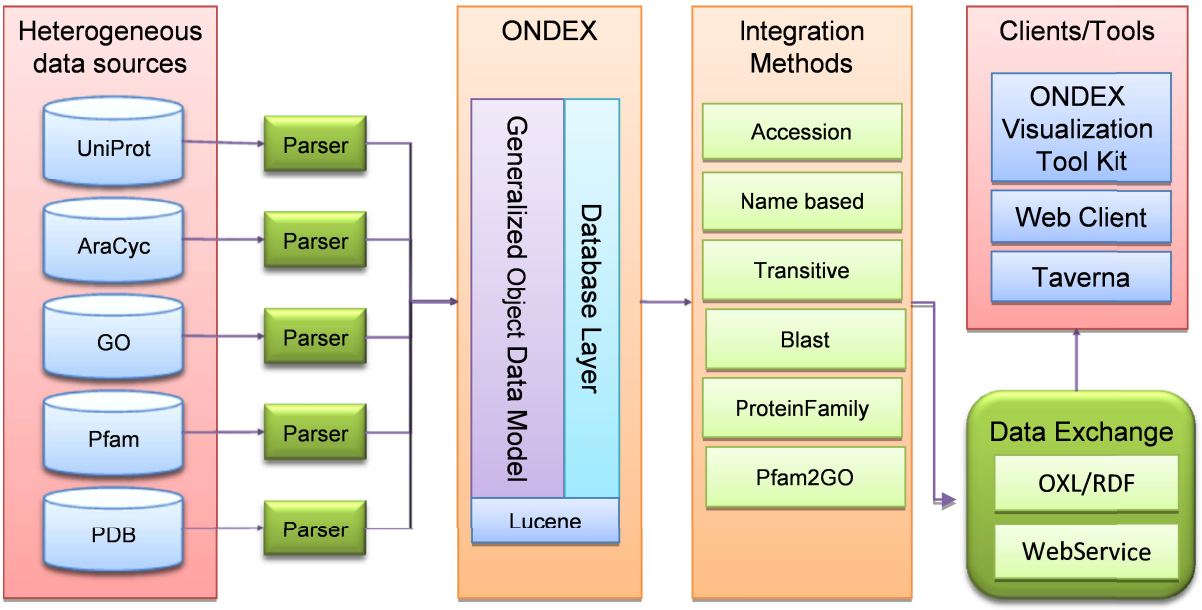

Figure 1: Overview of a typical ONDEX pipeline consisting of three parts: 1) parsing heterogeneous data sources into the generalized object data model of ONDEX; 2) identifying equivalent and related entries and creating new relations between them using integration methods; 3) analysing the integrated data using client tools, e.g. the ONDEX Visualisation Tool Kit (OVTK).

2.2 Data mapping

Data mapping methods create new relationships between equivalent or related entries from integrated data sets. Rather than merging elements which are found to be equivalent, data mapping creates a new equivalence relation between such concepts. Mapping methods assign scores and other parameters to the created relations, including provenance information. It is therefore possible to track the evidence for two concepts that are mapped and the method that created that mapping. The mapping methods used in this study include general purpose accession based mapping and transitive mapping methods. They also includes more specific sequence based mapping methods such as those used to link protein sequences to PFAM functional domains and ontology based mapping between GO terms and other databases.

Accession based mapping: Concepts in the ONDEX graph can have accession data attached. Accessions are extracted from references to external databases in the imported databases. These references may not always present a one-to-one relationship between entries of different databases. References presenting one-to-many relationships are termed "ambiguous" within the ONDEX system. The accession based mapping uses only non-ambiguous accessions to create links between equivalent concepts, i.e. concepts that share the same references in a one-to-one relationship.

Transitive mapping: Transitive relationships between concepts in the ONDEX graph are inferred from existing relations, e.g. if concept A is identified to be equivalent to concept B and concept B is known to be equivalent to concept C, then a new equivalent relationship between concept A and concept C is created by this mapping method.

Sequence2pfam mapping: The assignment of protein domain functional information to proteins sequences in ONDEX is achieved by exporting the sequence data into a FASTA (Pearson, 1990) file and matching against the consensus sequences from a local PFAM database (Sonnhammer, et al., 1997) using BLAST (Altschul, et al., 1997) or HMMER (Durbin, et al., 1998). The results are used to create relations between concepts representing proteins and relevant entries in the PFAM database.

External2go mapping: The GO consortium provides reference lists of GO terms that map terms to other classification systems, e.g. EC (Bairoch, 2000) enzymes or PFAM domains. The external2go mapping parses these lists and creates relations between entries of the GO database and entries of the other classification system.

2.3 Data filtering and knowledge extraction

Data mapping methods create a large number of new relations – edges on the ONDEX data graph. To be able to explore large and densely connected graphs, methods for data filtering and knowledge extraction are essential in order to reduce overall complexity. As shown in Figure 1, the final stage in the data integration process is to use filters to allow the extraction of sub-graphs according to certain criteria. These criteria are defined with respect to ONDEX metadata, e.g. concept classes or types of relations, the graph structure, e.g. the degree of a node, and context information associated with concepts and relations.

A new feature in the most recent version of the ONDEX system is the support for contexts which allow relations and concepts to be annotated or qualified with other concepts in the graph. Contexts permit a finer level of classification than would have been possible with metadata alone. For example, concepts that are components of biological pathway databases, such as proteins or reactions belong to certain pathways and are linked to certain cellular locations. These concepts are therefore qualified by having pathways or cellular location included in their lists of contexts. It is therefore possible during knowledge extraction to restrict the results returned to the corresponding sub-graph of a pathway or cellular location and thus reduce the number of nodes and edges significantly.

For equivalent or related entries that were identified by the mapping methods presented in Section 2.2 it is necessary, to copy contexts information across from different data sources, in order to include them during the process of knowledge extraction. This is achieved using the copycontext transformer, which extends contexts annotation across relations. New insights may emerge by transferring context information across data sources; For example, a protein is identified as belonging to a certain pathway having a known small molecule inhibitor within one data source and is characterized as being expressed in a certain tissue by another data source. By combining the context information from these two sources it is possible to infer that a pathway may be inhibited in this particular tissue type by the inhibitor.

A special kind of filter is the “relation collapse” filter. This filter processes the ONDEX graph, merging together concepts that satisfy the defined semantic constraints, e.g. concept class. Hence new super concepts are created that represent clusters of nodes. The redundant concepts are subsequently removed. The “relation collapse” filter is described in detail in (Taubert, et al., 2008). By collapsing concepts identified to be equivalent, the number of nodes and edges in the graph can be reduced.

In addition to the two methods outlined above which perform graph transformation and filtering (transfer of context information and collapsing equivalent concepts) several other filter and transformer methods have been implemented in the ONDEX system:

Concept class filter: Concepts of a given concept class and their corresponding relations are removed from the ONDEX graph.

Relation type filter: Relations of a given relation type are removed from the ONDEX graph. This might result in some previously connected concepts becoming unconnected.

Unconnected filter: Removes concepts (nodes) with a degree of zero from the ONDEX graph. Unconnected concepts do not usually have any value to the information in the graph.

All of these methods can be flexibly linked together to create a workflow for particular user applications. The resulting graphs can be exported in an ONDEX specific XML or RDF dialect for which generic exporters are available (Taubert, et al., 2007) and loaded into the ONDEX Visualisation Tool Kit (OVTK) (Köhler, et al., 2006) for further analysis. In the latest ONDEX release, workflows are not restricted to the example given in this publication and can easily be extended with user-defined functionality through a plug-in architecture.

2.4 Data Integration Exemplar

Figure 2 shows a graph of the meta data for the AraCyc database and illustrates how data integration in ONDEX can provide an elegant overview of the information captured during integration. A meta-graph shows all concepts classes and relation types currently in the graph; much like a database schema does for a relational database. Every entity type is represented as a concept class with all relevant relations. Further information such as cross reference accessions or synonyms are stored with the actual concepts. The meta-graph representation provides a useful high level overview of the data and helps to understand the structure of a loaded graph.

AraCyc is a biochemical pathway database for Arabidopsis thaliana containing information about enzymes, proteins, reactions, compounds and genes and how they are related to each other. For each pathway component (e.g. protein, reaction), additional annotations such as cross database references, publication references, or in the case of enzymes, GO-terms and EC-numbers, are also available.

The OVTK user interface uses context relationships to facilitate a pathway-oriented view of the integrated dataset. For example, any given pathway concept provides a context for all constituent elements of that pathway. By selecting a particular pathway from a list, all of the associated data for that pathway is displayed without the need for additional filtering. .The result is similar to querying the web interface of the AraCyc database.

2.5 Data Integration Pipeline

A simple data integration workflow is presented schematically in Figure 3. Information from the AraCyc database, UniProt, and the PDB is combined with GO-terms and PFAM protein family information to provide richer annotations. Structural information from PDB was mapped to protein families; GO terms annotated to these families were used to infer protein function.

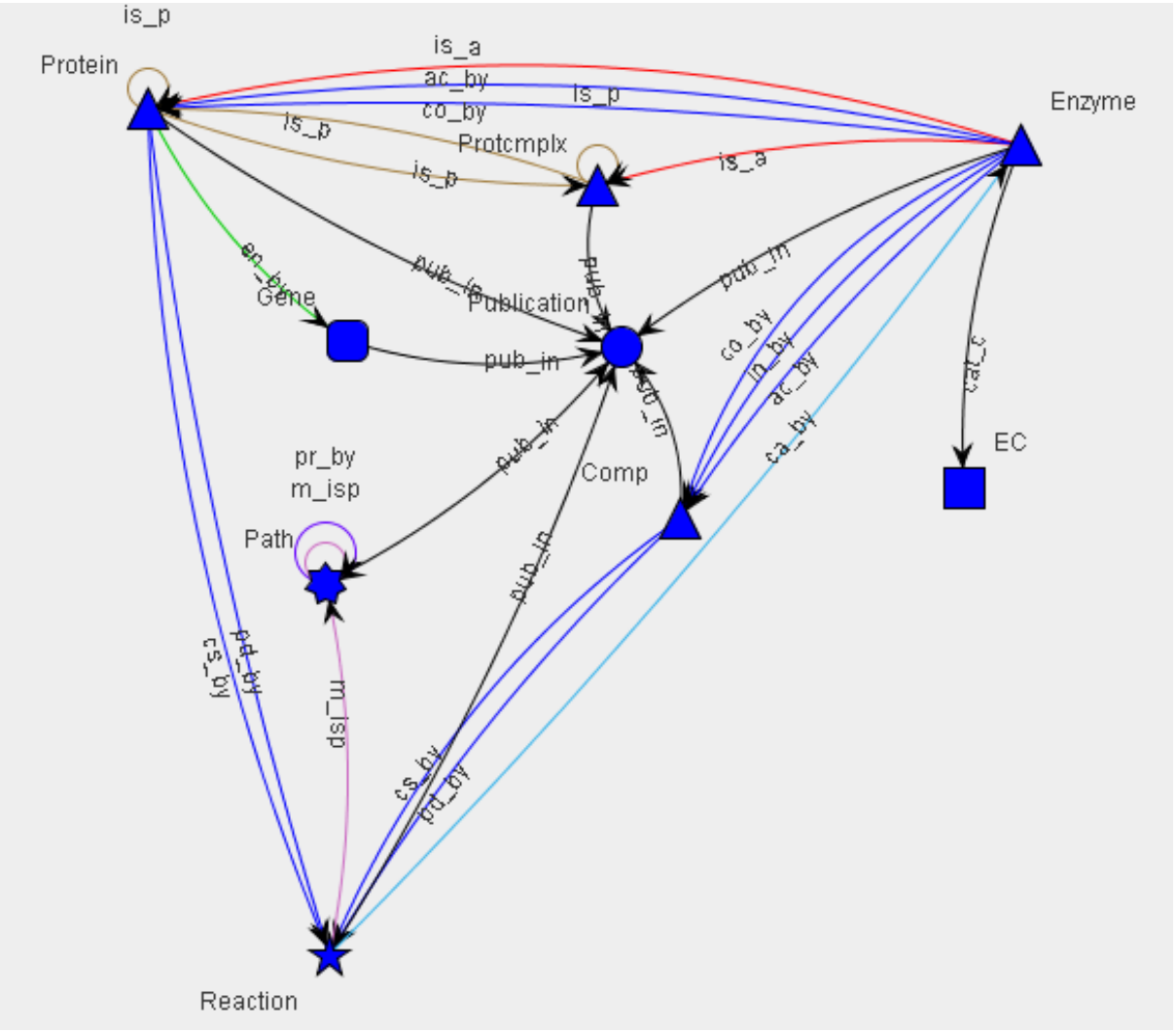

Figure 2: ONDEX MetaGraph for the AraCyc database

The pipeline consists of 16 steps grouped into four blocks that can be separated into four distinct stages, shown by colour and explained in the following:

Step 1 to 5: Integrating AraCyc, GOA and UniProtKB. The current release of AraCyc contains information for about 6025 proteins and a number of protein complexes. In the second pipeline step, protein sequences for all of these entries were obtained from the UniProt database.

At the time of writing, the UniProt database contained approx. 5.5 million entries, of these 349,480 were manually curated (UniProtKB/Swiss-Prot) and the remaining 5,329,119 were automatically annotated (UniProtKB/TrEMBL). Both Swiss-Prot and TrEMBL were used to create the integrated dataset.

GO annotations for the protein entries in UniProt were taken from the GOA database cross reference files. These files provide links from each protein to their manually curated GO-terms. The GOA parser creates concepts of class protein connected to concepts for GO terms.

After the parsing process, two equivalent protein concepts exist for each protein in AraCyc. These were mapped using accession based mapping to combine the data from AraCyc, Uniprot

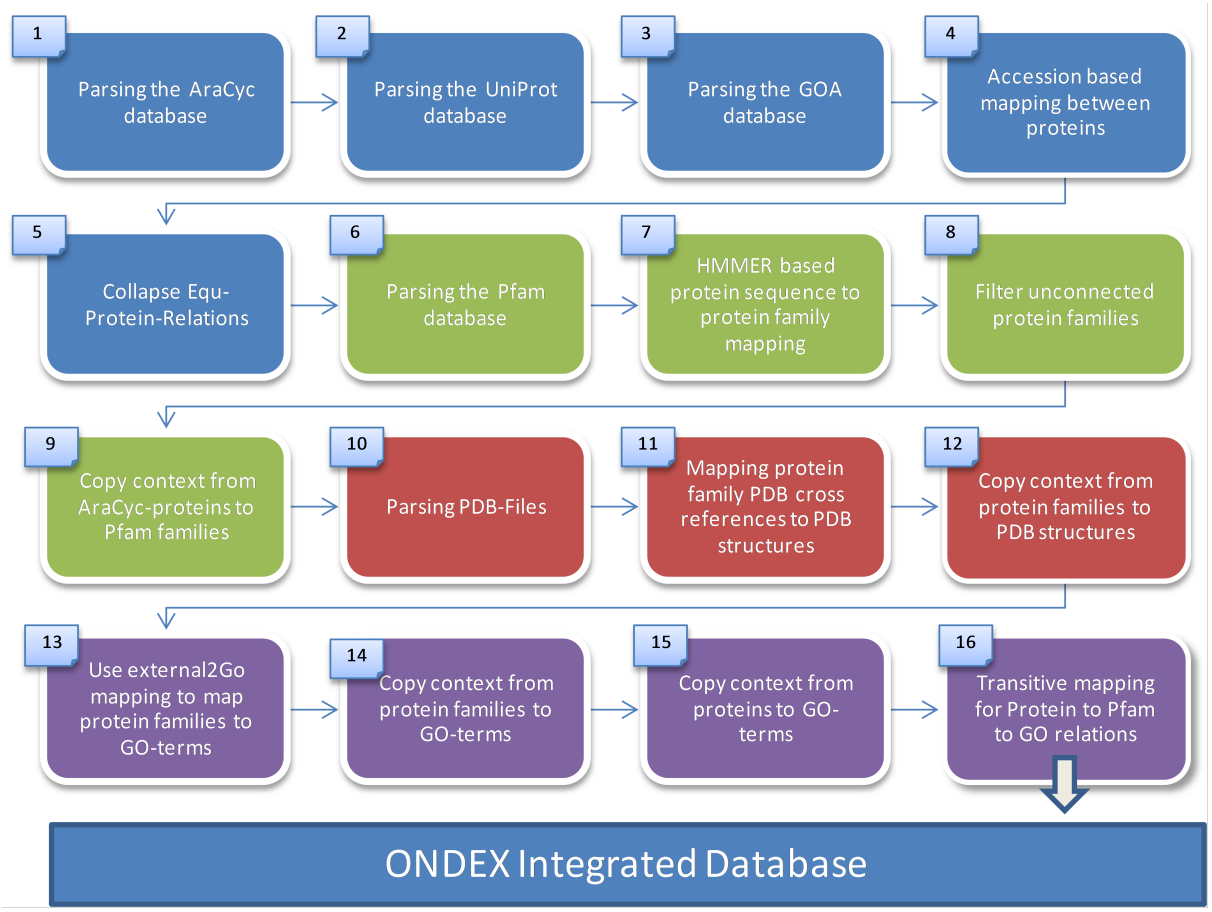

Figure 3: Pipeline for annotating protein sequences with GO terms, protein family information and PDB structures. Blue: Integrating AraCyc, GOA and UniProtKB; Green: Adding PFAM-family information to proteins; Red: Mapping structural information; Purple: Mapping GO terms to proteins.

and GOA. Where there were multiple matches in UniProt for the same entry in AraCyc, the manually curated one was preferred. Where there were no manually curated annotations, automatic annotation was used instead. In all cases only one annotation per sequence was used in order to avoid redundancy in the test set. Afterwards the "collapse filter" was used to merge the information from these three sources into one super concept.

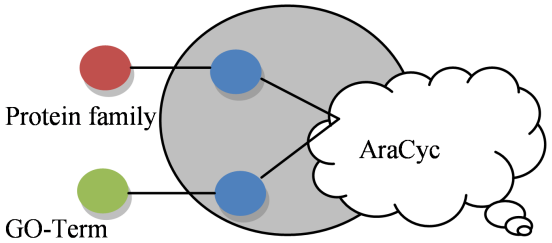

Figure 4: After the "collapse filter" was applied the proteins (blue) were merged in to one concept.

Step 6 to 9: Adding PFAM-family information to proteins. PFAM is a high quality, publicly available protein family database, which is based on Hidden Markov Model profiles

(HMM) and maintained by the Sanger Institute. It provides cross-references to structure information and GO annotations via GOA.

Firstly the PFAM database was parsed into an ONDEX graph representation. The sequence2pfam mapping was then used to map proteins to protein families based on sequence information. The Sequence2pfam method supports three ways of mapping a protein to a PFAM protein family:

a) Via a publicly available implementation of HMMER

b) Using the "TimeLogic" implementation of HMMER from Active Motif, Inc.

c) BLAST search with PFAM domain information derived from NCBI Conserved Domain Database (CDD) database.

For the workflow presented here, the "TimeLogic" (`http://www.timelogic.com`) implementation of HMMER was used because it has a higher throughput and sensitivity compared to the BLAST approach. Protein families with no associated proteins were removed using the unconnected filter. Afterwards, context information was copied to relations created via the sequence2pfam mapping method.

Step 10 to 12: Mapping structural information. After protein family classifications were added, PDB structures associated with each protein family were assigned to related proteins by traversing the new mappings. For reasons of space-efficiency, it was not practical to incorporate the entire set of crystallographic coordinate data for each protein into the graph representation directly. Instead, PDB coordinates are loaded on demand by the integrated Jmol PDB-viewer (http://www.jmol.org) whenever the structure view is requested by the user.

Step 13 to 15: Mapping GO terms to proteins. The final part of this workflow added GO annotations to the graph. In addition to GO terms extracted from the GOA database we have used a second way of deriving GO annotation using the PFAM family to GO mapping file. The reference file, which is provided by the GOA project, is processed by the external2Go mapping. Transitive mapping was used to infer protein to GO function based on protein family GO annotations (see).

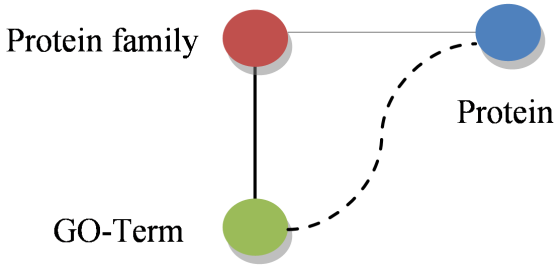

Figure 5: The dashed relation is created using the transitive mapping.

The final ONDEX integrated data graph was visualised using the OVTK. An example Pathway is displayed in Figure 3. As can be seen the Pathway (red stars) is enriched with GO-Terms (pink/orange/red circles) with PFAM families (green circles) and PDB structures (purple pentagons). Furthermore one PDB structure was selected and displayed.

3 Evaluation of Annotation Methods

The motivation for using a data integration approach to sequence annotation is to improve the accuracy of automatic annotation processes. This will enable tools to be developed that can assist users and database curators to assess the quality of a functional assignment. A quantitative measure of the accuracy of an assignment is required so that different annotation methods can be compared. This is particularly important for assessing annotations based on integrated data resources, because there will be additional uncertainties in the quality of the reference information being used and the success of the integration methods.

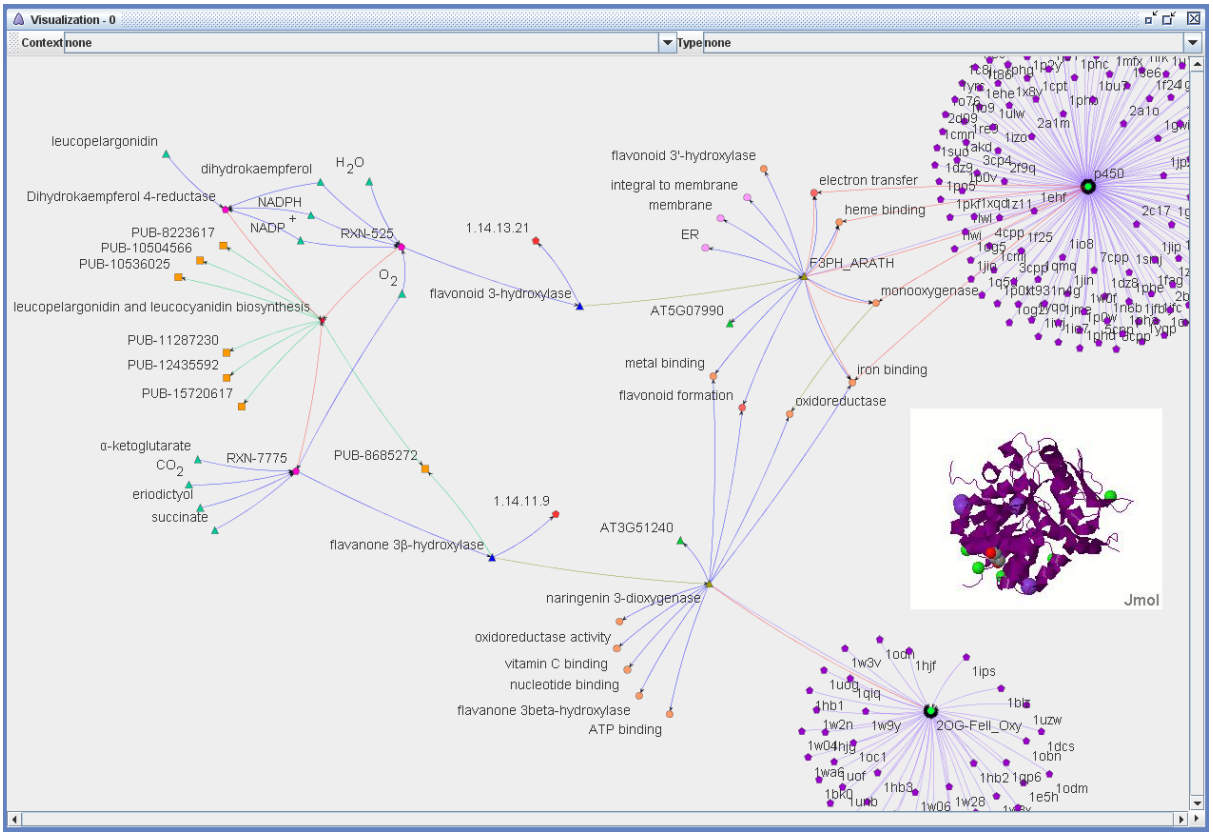

Figure 6: Resulting representation of an extract of the "leucopelargonidin and leucocyanidin biosynthesis". An AraCyc pathway in OVTK2 with GO, PFAM and Structure information.

The first step in such an assessment is to establish a reference set against which other annotations and annotation methods can be compared. To evaluate the quality of annotation derived from sequence homology and structural classification of proteins in this study we used the Gene Ontology Annotation (GOA) database as the reference set. Our reason for selecting GOA was because it provides high quality annotation of gene products from the UniProt Knowledgebase either by computationally deriving GO terms from SwissProt, InterProt, HAMAP and EC numbers or through manual curation of scientific publications. In our evaluation (Figure 4) we used the 20458 UniProt entries for Arabidopsis that are annotated in GOA (˜58% of entries). This relatively high coverage and the quality of annotation itself make it a very good choice for the reference dataset.

The test set was compared to the reference set using the weighted harmonic mean of precision and recall (F1), a recognised standard test for measuring performance of information retrieval methods (Goutte and Gaussier, 2005).

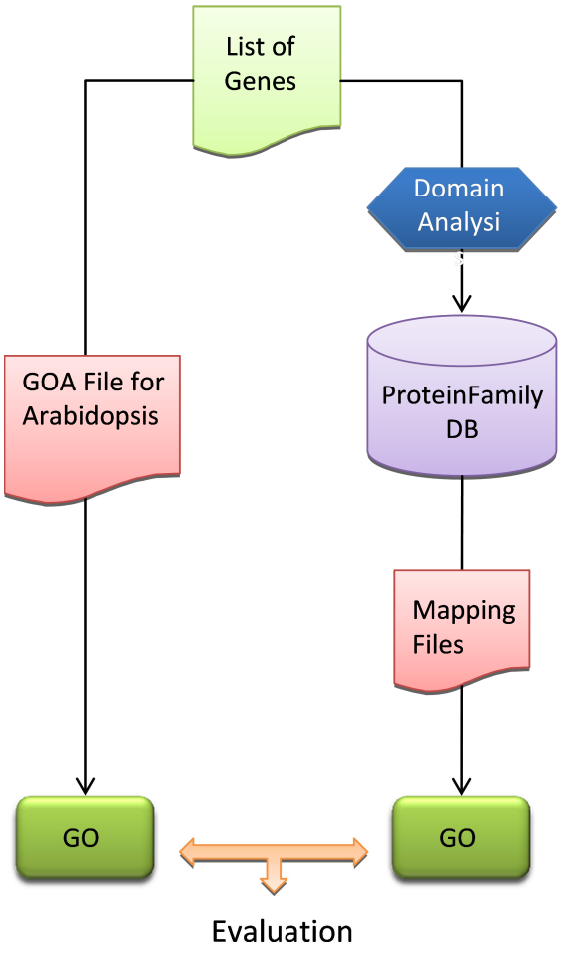

Figure 7: Evaluating our PFAM based GO mappings to an Arabidopsis reference set from publicly available GOA files.

	Molecular Function	Biological Process	Cellular Compo-nent	Overall
Precision	0.67	0.82	0.86	0.73
Recall	0.43	0.58	0.26	0.45
F-Score	0.52	0.68	0.40	0.56

Table 1: Evaluation of annotation quality for different Gene Ontology categories.

Table 1 summarises the results of the comparison between manual GOA annotation and our predicted annotation and shows precision, recall and F-Score for the three GO ontology categories. The best F-Score was achieved for identification of biological process. The second best performance was annotating by molecular function. These results are perhaps not surprising, because most PFAM entries are constructed from functionally equivalent protein families. The performance of the PFAM annotation pipeline for both biological process and molecular function undoubtedly reflects the importance of conserved sequence and structure features in determining function. The difference in granularity between the definitions of molecular function and biological processes may also contribute to the divergent performance of predictions in these categories. By contrast, the cellular component category annotation had the lowest recall and therefore the lowest F-Score. This is also reasonable, because the PFAM sequence profiles

are generally less able to predict cellular location (e.g. targeting signals) while the GO ontologies capture cellular location in some detail. Expert curators of protein databases will also have used other information such as the original publication to assign the cellular location for the protein. The result of comparing PFAM-based annotations with the reference GOA annotations for cellular component is perhaps notable for having a greater level of success than might have been expected.

4 Discussion

In this study we have presented some of our preliminary research into an integrative approach for the annotation and re-annotation of genome sequence data. Our first objective was to extend the range of databases used in the data integration framework ONDEX and develop a workflow that integrated the databases considered the most valuable for assigning protein function (i.e. curated protein sequences (UniprotKB), gene classification ontologies (GO), protein structure (PDB) and protein family assignment (PFAM)). The semantic integration of these data sources is the core of a developing platform that will be used to augment the annotation of emerging genome sequences being studied at Rothamsted Research and by our collaborators. The ONDEX visualisation toolkit (OVTK) has many features that will support more detailed scrutiny of the gene and protein function assignment by making explicit the links between the different data sources and when biochemical pathway information is also available, this can also be integrated and visualised to show the broader biological context of the genes of interest including displaying protein structure information where this is available.

During this research, it was clear that at key points in the integration process, alternative mapping methods could be used to provide the links between protein sequence and GO annotations and it was therefore important to quantify the success of these methods as part of the validation process for the integration workflow. Comparing the success of these mapping methods is similar to the bigger problem of assessing the success of assigning a biological function to each protein sequence derived from the coding sequences in a new genome. The comparison between different annotation approaches is however, not straightforward and a number of different statistical methods and criteria for successful classification are used, for example hierarchical evaluation measure (Kiritchenko, et al., 2005) and Fisher's false discovery rate (Bluthgen, et al., 2005). Although complex statistical measures have some advantages, such as dealing with partially correct annotations, the classic precision, recall and F-score still remain among the most widely recognised measures of information retrieval quality because they are much more straightforward to interpret and compare than more complex tests. For this reason we selected these common measures to explore the performance of the functional annotation methods presented in this paper. The results of the evaluation indicate that it is possible to use protein family categorization based on multiple sequence alignments to successfully infer biological and molecular function with a reasonable accuracy (82% and 86% respectively) by comparison with expert manual annotation. Recall depends on the actual classification made, i.e. molecular function, biological process or cellular component for the reasons detailed in Section 3.

The selection of a comprehensive reference set (Gold Standard) has proved to be a challenging task, as it is difficult to satisfy the needs of different methods with the same reference set. Constructing a reference set from publicly available GOA files has the advantage that the

information is readily available from GOA; no further annotation work is required and the resulting reference set can be used as a common benchmark since it is freely available. This comparison of GO mapping methods has highlighted the complexity of such evaluations and further research is needed to explore the issues identified in this preliminary analysis. Future developments of the ONDEX data integration system and user interfaces will extend the range of information types that can be incorporated into the annotation process and more extensive statistical methods will be developed to analyse the data and annotations. This research will also address the broader issue of assessing the contribution that data integration brings to the annotation process. This paper, as a report of work in progress, demonstrates that the ONDEX system can be adapted to this task with relative ease and provides a powerful platform for future genome annotation research.

Acknowledgements

The authors are pleased to acknowledge funding from the Biotechnology and Biological Sciences Research Council (BBSRC) through project grant BBS/B/13640. Rothamsted Research is in receipt of grant aided support from the BBSRC. The authors also wish to thank Paul Verrier for his helpful comments at various stages in the preparation of this manuscript.

References

[1] J. Taubert, R. Winnenburg, M. Hindle, J. Weile, J. Baumbach, S. Philippi, C. Rawlings, and J. Köhler. Data integration, information filtering and knowledge extraction with ONDEX. In preparation, 2008.

[2] L. J. Post, M. Roos, M. S. Marshall, R. van Driel, and T. M. Breit. A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data. *Bioinformatics*, 23(22):3080–3087, 2007.

[3] S. Salzberg. Genome re-annotation: a wiki solution? *Genome Biology*, 8:102, 2007.

[4] C. M. Smith, J. H. Finger, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald. The mouse gene expression database (GXD): 2007 update. *Nucleic Acids Research*, 35(Database issue):D618–D623, 2007.

[5] J. Taubert, K. P. Sieren, M. Hindle, B. Hoekman, R. Winnenburg, S. Philippi, C. Rawlings, and J. Köhler. The OXL format for the exchange of integrated datasets. *Journal of Integrative Bioinformatics*, 4(3), 2007.

[6] M. Baitaluk, X. Qian, S. Godbole, A. Raval, A. Ray, and A. Gupta. PathSys: integrating molecular interaction graphs for systems biology. *BMC bioinformatics*, 7:55, 2006.

[7] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rueegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ondex. *Bioinformatics*, 22(11):1383–1390, 2006.

[8] J. Küntzer, T. Blum, A. Gerasch, C. Backes, A. Hildebrandt, M. Kaufmann, O. Kohlbacher, and H.-P. Lenhof. Bn++ - a biological information system. *Journal of Integrative Bioinformatics*, 3(2), 2006.

[9] E. K. Neumann and D. Quan. Biodash: a semantic web dashboard for drug development. In *PSB'06: Pacific Symposium on Biocomputing 2006*, pages 176–187, Hawaii, USA, 1 2006. World Scientific Publishing.

[10] Oracle Corporation. Berkeley DB Java Edition Direct Persistence Layer Basics, 2006. `http://www.oracle.com/database/docs/BDB-JE-DPL-Basics-Whitepaper.pdf`.

[11] N. Bluthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel, and D. Beule. Biological profiling of gene groups utilizing gene ontology. *Genome Informatics*, 16:106–115, 2005.

[12] P. V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmen, and M. Oresic. Data integration and visualization system for enabling conceptual biology. *Bioinformatics*, 21(suppl_1):i177–i185, 2005.

[13] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In D.E. Losada and J.M. Fernandez-Luna, editors, *ECIR'05: European Colloquium on IR Research 2005*, pages 345–359, Santiago de Compostela, Spain, 2005. Springer.

[14] S. Kiritchenko, S. Matwin, and A. F. Famili. Functional annotation of genes using hierarchical text categorization. In *Proceedings of the ISMB BioLINK SIG on Text Data Mining, ISMB'05*, 2005.

[15] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(32):D115–119, 2004.

[16] J. Köhler. Integration of life science databases. *Drug Discovery Today: BIOSILICO*, 2(2):61–69, 2004.

[17] J. Köhler, C. Rawlings, P. Verrier, R. Mitchell, A. Skusa, A. Ruegg, and S. Philippi. Linking experimental results, biological networks and sequence analysis methods using ontologies and generalized data structures. *In Silico Biology*, 5:33–44, 2004.

[18] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research*, 13(4):662–672, 2003.

[19] L. A. Mueller, P. Zhang, and S. Y. Rhee. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology*, 132(2):453–460, 2003.

[20] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo. Pathway studio – the analysis and navigation of molecular networks. *Bioinformatics*, 19(16):2155–2157, 2003.

[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25:25–29, 2000.

[22] A. Bairoch. The ENZYME database in 2000. *Nucleic Acids Research*, 28(1):304–305, 2000.

[23] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, chapter The theory behind profile HMMs. Cambridge University Press, 1998.

[24] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D Biological Crystallography*, 54:1078–1084, 1998.

[25] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[26] E. L. Sonnhammer, S. R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics*, 28(3):405–420, 1997.

[27] W. R. Pearson. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, 183:63–98, 1990.