BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 9 2011, pages 1299–1306
doi:10.1093/bioinformatics/btr134

Systems biology

Advance Access publication March 16, 2011

Customizable views on semantically integrated networks for systems biology

Jochen Weile[1], Matthew Pocock[1], Simon J. Cockell[2], Phillip Lord[1], James M. Dewar[3], Eva-Maria Holstein[3], Darren Wilkinson[3,4], David Lydall[3,5], Jennifer Hallinan[1] and Anil Wipat[1,3,*]

[1]School of Computing Science, Faculty of Science Agriculture and Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, [2]Bioinformatics Support Unit, Institute for Cell and Molecular Biosciences, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE2 4HH, [3]Centre for Integrative Systems Biology of Ageing and Nutrition, Institute for Ageing and Health, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE4 5PL, [4]School of Mathematics and Statistics, Faculty of Science Agriculture and Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU and [5]Institute for Cell and Molecular Biosciences, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: The rise of high-throughput technologies in the post-genomic era has led to the production of large amounts of biological data. Many of these datasets are freely available on the Internet. Making optimal use of these data is a significant challenge for bioinformaticians. Various strategies for integrating data have been proposed to address this challenge. One of the most promising approaches is the development of semantically rich integrated datasets. Although well suited to computational manipulation, such integrated datasets are typically too large and complex for easy visualization and interactive exploration.

Results: We have created an integrated dataset for *Saccharomyces cerevisiae* using the semantic data integration tool Ondex, and have developed a view-based visualization technique that allows for concise graphical representations of the integrated data. The technique was implemented in a plug-in for Cytoscape, called OndexView. We used OndexView to investigate telomere maintenance in *S. cerevisiae*.

Availability: The Ondex yeast dataset and the OndexView plug-in for Cytoscape are accessible at http://bsu.ncl.ac.uk/ondexview.

Contact: anil.wipat@ncl.ac.uk

Supplementary information: Supplementary data is available at *Bioinformatics* online.

Received on August 27, 2010; revised on February 10, 2011; accepted on March 9, 2011

1 INTRODUCTION

1.1 Systems biology and integrative bioinformatics

Systems biology considers multiple aspects of an organism's structure and function at the same time, using the plethora of data that is publicly available online. Biologists have access to heterogeneous data covering many different aspects of biology; 1230 entries are listed in the 2010 database issue of *Nucleic Acids Research* (http://nar.oxfordjournals.org/content/38/suppl_1). For model organisms such as *Saccharomyces cerevisiae*, the abundance of freely available data has long since exceeded the point at which it can be analysed manually.

Most data sources still exist in isolation, each with its own specialization and focus (Stein, 2002). In many cases, databases lack semantic links to each other, even when they are providing data about the same entities. This partitioning of knowledge is a particular problem for systems biology, hampering the examination of the emergent behaviours inherent in complex biological systems. The problem of using distributed, heterogeneous datasets is addressed by the subdiscipline of integrative bioinformatics. There are many different approaches to the integration and querying of large heterogeneous datasets. Early XML-based approaches (Achard *et al.*, 2001) were soon replaced by more sophisticated methods. Federated approaches, such as the distributed annotation system (DAS) (Prlić *et al.*, 2007) establish a central view on the data by translating each internal query into a set of external queries in order to retrieve the required data from the outside sources on the fly (Heimbigner and McLeod, 1985). Semantic web approaches, as endorsed by the Open Biological and Biomedical Ontologies foundry (OBO) (Smith *et al.*, 2007) or as implemented in YeastHub (Cheung *et al.*, 2005), establish a network of interlinked ontologies using a set of controlled vocabularies (Brinkley *et al.*, 2006). Data warehousing approaches such as BioMART (Haider *et al.*, 2009) EcoCyc (Keseler *et al.*, 2010) and SAMBA (Tanay *et al.*, 2004), convert and incorporate data from their external sources into their own database schema and provide custom queries.

1.2 Ondex

Ondex is a graph-based data integration framework (Kohler *et al.*, 2006), which takes a semantic warehousing approach. Links between entities in different datasets may be directly extracted from the available data sources, inferred as part of the integration, or generated by external tools, such as BLAST (Altschul *et al.*, 1990). These data can subsequently be matched and interlinked with each other semantically. The result is a semantically enriched dataset, with which users can interact and also visualize.

*To whom correspondence should be addressed.

Ondex incorporates data into a network of entities termed 'concepts', connected by 'relations', all of which can carry 'attributes'. All concepts, relations and attributes have 'types', which are organized in a hierarchical fashion. For example, the concept type `Protein` is a subtype of `Molecule`, which is itself a subtype of `Thing`. This hierarchy means that every `Protein` concept is also a `Molecule` and a `Thing`. Similarly, the relation type *catalyzes* is a subtype of *actively_participates_in*. Therefore, every statement that *p*::`Protein` *catalyzes r*::`Reaction` means that *p actively_participates_in r*. Adding this type of information means that the computer stores not only data, but also its meaning, providing a 'semantic representation' of the data.

A special kind of attribute that all concepts in the an Ondex graph have is the cross-reference (called 'concept accession' in Ondex). Concept accessions make it easy to connect concepts originating from different sources. For example, if an Ondex dataset contains two `Gene` concepts which have been imported from two different data sources, but which share the same concept accession, Ondex can connect these concepts with a new relation of type *same_as*, and can at a later time merge these concepts into a single concept. All concepts and relations also have attached provenance information stating their origin and any associated evidence codes.

The process of integrating data into this data structure is performed by the Ondex workflow engine, which employs a plugin architecture. Parsers, mapping methods and other plugins can be developed using an open API (Taubert *et al.*, 2007) and can subsequently be used in workflows.

In this work, we describe a mechanism to collapse groups of concepts into a simpler, more easily visualized and conceptualized representation. We have implemented this mechanism as a plugin for Cytoscape, called OndexView, which facilitates the focused analysis of parts of a large, complex network. We demonstrate the value of this approach by applying it to an investigation into the systems biology of telomere maintenance in the baker's yeast *S. cerevisiae*.

1.3 Telomere maintenance and *BMH1/2*

Telomeres are structures composed of the ends of eukaryotic chromosomes together with their capping nucleoprotein complexes. These structures appear to play a major role in the ageing process (Blasco, 2007; Cheung and Deng, 2008). Without the capping proteins, chromosome ends appear as double-stranded breaks to the cell's DNA damage detection mechanism, triggering a checkpoint response and ultimately cell cycle arrest (Longhese, 2008; Sandell and Zakian, 1993). With each cell division, telomeres shorten (Longhese, 2008). When telomere length falls under a certain threshold, the checkpoint response is triggered. This mechanism appears to contribute to the establishment of a cell's finite lifespan.

Saccharomyces cerevisiae is an excellent subject for studying telomere biology, because it is one of the simplest and most well-studied eukaryotic model organisms, and telomere biology is highly conserved across eukaryotes. In *S. cerevisiae*, an important component of the telomere maintenance mechanism is the protein Cdc13, which binds to the single-stranded DNA overhangs of telomeres. Cdc13 has two major functions: capping the telomeric DNA and recruiting telomerase (Garvik *et al.*, 1995; Nugent *et al.*, 1996), which is part of the telomere repair mechanism.

CDC13 is an essential gene, so it is difficult to characterize using knock-out mutants. However, there is a temperature-sensitive mutant called *cdc13-1*. This mutant has a wild-type phenotype below 26°C, but above this temperature telomeres become uncapped, the checkpoint response is induced and the cells stop dividing (Weinert and Hartwell, 1993).

Epistatic interactions between *cdc13-1* and every non-essential gene of the *S. cerevisiae* genome have been studied (Addinall *et al.*, 2008) using a high-throughput synthetic genetic array (SGA) assay (Tong *et al.*, 2001). Addinall and co-workers identified a large number of genes which appear to genetically interact with *cdc13-1*. The biological function of many of these genes is already understood, but the role of others requires further investigation.

A particularly challenging problem is the phenotype of the two paralogues *BMH1* and *BMH2*. The deletion of the gene *BMH1* strongly suppresses the *cdc13-1* phenotype, while the deletion of its paralogue *BMH2* suppresses *cdc13-1* rather weakly. *BMH1* and *BMH2* share 91.6% sequence identity. They both encode 14-3-3 proteins, a class of proteins which usually occur as dimers and which bind to phosphoproteins (Chaudhri, 2003). Members of the 14-3-3 family have a variety of different functions in eukaryotes, including directly modifying the functionality of their target proteins, mediating and controlling transport processes between the cytoplasm and different organelles and serving as scaffolds for interactions between different proteins (Tzivion *et al.*, 2001). An important role for *BMH1* and *BMH2* in the regulation of carbohydrate metabolism has also been suggested (Bruckmann *et al.*, 2007).

It is not immediately obvious why *BMH1* and *BMH2* behave differently despite their close homology. Clearly, the difference in phenotype between two such closely homologous genes cannot be understood by studying the genes in isolation; a systems biology approach is essential.

In order to investigate this problem, we used Ondex to integrate five publicly available data sources as well as a homology dataset generated from BLAST results. We then enriched this data with semantic links between the concepts from the different data sources. In addition, we developed a novel, view-based visualization approach which we used to analyse this large, complex dataset. This analysis produced several testable hypotheses.

2 METHODS

2.1 Data sources

We integrated six different data sources using Ondex. Genomics data, as well as the latest Gene Ontology (Ashburner *et al.*, 2000) annotations, were obtained from the *Saccharomyces* Genome Database (SGD) (Cherry *et al.*, 1997). A yeast regulatory network was acquired from Balaji *et al.* (2006). A curated model of the yeast metabolic network was sourced from Herrgard *et al.* (2008). A yeast protein–protein interaction (PPI) network and a network of known genetic interactions (GI) were taken from the BioGRID database (Stark *et al.*, 2006). The BioGRID data includes the GIs reported by Addinall *et al.* (2008). Homology links between genes were created using BLAST (Altschul *et al.*, 1990) with an 85% sequence identity threshold (Table 1). Further information regarding the data sources can be found in the Supplementary Material.

2.2 Integration

The source datasets were integrated to form a combined Ondex *Saccharomyces* knowledge network. A metadata model was designed for the knowledge network to capture the semantics of concept and relationship

Table 1. Data sources used in this work

Data	Source	Version/date
Genome	SGD	11/02/2010
GO annotations	SGD	11/02/2010
Interactome	BioGRID	v2.0.61
Regulatory network	Balaji *et al.* (2006)	NA
Metabolic network	Herrgard *et al.* (2008)	v1.0
Homology	BLAST (id.> 85%)	NA

NA = not applicable.

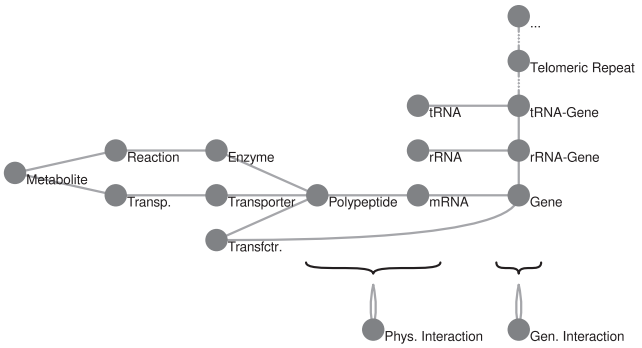

Fig. 1. Simplified schematic of the data structure underlying the Ondex *Saccharomyces* knowledge network. Circles represent concept types, and lines represent relations between them.

types found within the different datasets (Fig. 1). This model provides information about how certain concept types inherit features from one another (for example, an `Enzyme` is always a `Protein`), and allows such concepts to be treated as both concept types in the downstream workflow.

Ondex parsers for each data source produce an Ondex-compatible representation of the data. For example, the BioGRID parser creates an appropriately typed concept for every reported gene, protein and RNA in the database, and an interaction concept of a corresponding type for every interaction. All genes, proteins and RNAs that take part in an interaction are linked to that interaction concept with a *participates_in* relation.

A mapping algorithm based on matching cross-references was used to find identical concepts in the Ondex graph. For example, the protein encoded by the ORF *YDL220C* from BioGRID matches a concept with the same accession number in the metabolic network. When the procedure identifies a group of matching concepts, it merges them into one single concept that captures all of the information that was represented by the group members. The procedure checks whether all concepts in a group have compatible types. If so, it automatically uses the most specific type present in the group for the merged concept. Otherwise it reports the inconsistency. For example, a concept cannot be both a `Gene` and a `Pseudogene`, but a concept can be both a `Protein` and an `Enzyme`, in which case `Enzyme` is the more specific type.

The semantic model reflects current understanding of molecular cell biology: concepts of type `Gene` have *transcription* relations linking them to `mRNA` concepts, which in turn connect to `Polypeptide` concepts via *translation* relations. `Polypeptides` can be *part_of* `Proteins`. `Enzymes`, a subtype of `Proteins`, may *catalyze* `Reactions`, which *consume* and *produce* `Molecules` such as `Metabolites` and `Proteins`.

`Transcription_factors`, a subtype of `Protein`, connect to `Genes` via *regulation* relations. `Genes` in turn are `Nucleotide_features`, which connect via *adjacency* or *overlap* relations, establishing genomic context. All `Nucleotide_features` can *participate_in* various `Genetic_interactions`. Similarly, `Molecules`, such as `RNAs` or `Polypeptides` can *participate_in* `Physical_interactions`.

Non-physical data can also be associated with existing concepts. `Nucleotide_features` can contribute to `Biological_processes`. `Proteins` can have a `Molecular_function`. Various parts of the knowledge network can also be associated with `Publication` concepts.

All of the parsers and mapping methods, as well as the metadata that were created for this work, are available as part of the Ondex suite, which is licensed under the GNU GPL v3, and is downloadable from http://www.ondex.org/. Further details regarding the integration process can be found in the Supplementary Material.

2.3 A novel visualization strategy

The graph produced by the integration is stored in the Ondex XML format OXL (Taubert *et al.*, 2007) and can be browsed in Ondex. The resulting network is large and complex, and contains many types of concepts and relations. In order to present a more succinct and biologically focused view of the data, we developed a plugin for the network visualization tool Cytoscape (Shannon *et al.*, 2003) called OndexView.

This new visualisation allows the user to define 'views' on the data. A view focusses on one type of concept. All concepts of that type found in the knowledge network are visualized as network nodes. The user can query the knowledge network for associations between the concepts. These associations will be visualized as edges between the corresponding nodes. Querying for associations is accomplished by specifying a set of metadata motifs. The instances of these motifs found in the underlying knowledge network are then used to create corresponding associations in the view. Metadata motifs are alternating sequences of concept types and relation types that begin and end on the same concept type (Fig. 2E).

A modified depth-first search algorithm is used to extract motif instances from the underlying Ondex knowledge network. At each depth level, the algorithm checks whether the currently explored path matches the target motif. For example, a user interested in proteins and metabolic pathway relationships between them could specify a motif representing this type of interaction. Applying the algorithm using this motif produces a view on the graph, containing only those elements in which the user is interested (Fig. 2). It is possible to combine multiple motifs to form a view as long as the motifs share the same start point concept type.

2.4 Using OndexView with the *Saccharomyces* knowledge network for generating hypotheses

A semantically collapsed view of a complex network facilitates hypothesis generation by providing a simple representation of genes of interest and their interactions. A subnetwork consisting only of these genes and their neighbours can be generated, and inspected to identify edges of potential interest. Examination of the annotations attached to these edges and their adjacent nodes in the Ondex network provides links to available knowledge about the underlying biology. The semantically collapsed view also provides a way of perusing relevant literature in a focused manner, making hypothesis generation considerably more efficient than the alternative of trawling through large databases of publications.

Using OndexView, we defined five motifs over `Gene` concepts covering genetic and physical interactions, homology, regulation and metabolic precedence (Table 2).

We created a view using all five motifs, querying for the immediate neighbourhood of *BMH1*, *BMH2* and *CDC13*. We then laid out the view, separating all neighbouring genes into groups: Exclusive neighbours of *BMH1*, exclusive neighbours of *BMH2*, joint neighbours of *BMH1* and *BMH2* and others.

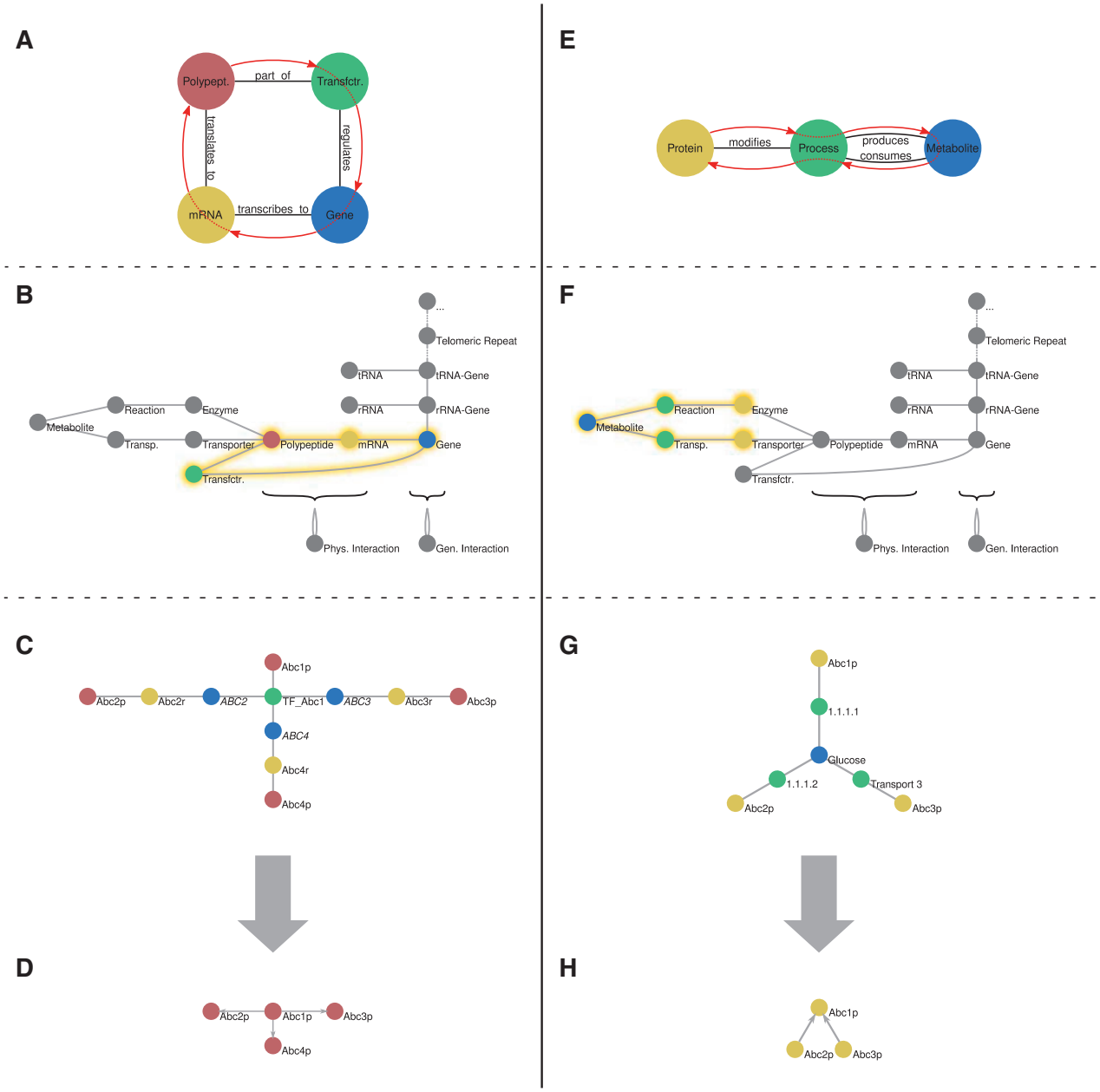

Fig. 2. Schema illustrating two examples for constructing associations based on metadata motifs. (**A**) The motif shown selects all paths from a polypeptide that is part of a transcription factor which regulates a gene that is transcribed to an mRNA which is translated into another polypeptide. (**B**) The motif from (A) is shown in context of the complete metadata structure. (**C**) A small example subnetwork to which the motif can be applied. Finding the motif (A) in the subnetwork, we identify three matching paths. Each element in these paths matches its corresponding element in the motif. (**D**) The view resulting from collapsing the paths identified in (C) according to motifs from (A) contains three edges; one for each matching path. (**E**) The motif selects all paths from a protein that modifies a process that produces a metabolite which is then consumed by another process back to another protein that modifies that process. Such a motif could be described as 'precedence in a metabolic pathway'. (**F**) The motif from (E) is shown in context of the complete metadata structure. (**G**) A small example subnetwork to which the motif can be applied. Finding the motif (E) in the subnetwork, we identify two matching paths. Each element in these paths either matches or is a subtype of the corresponding element in the motif. (**H**) The views resulting from collapsing the paths identified in (G) according to motifs from (E) contains two edges; one for each matching path.

Table 2. Motif definitions used in this work

Association name	Motif
homologue	`Gene` *is_homologue* `Gene`
metabolic path	`Gene` *encodes* `Polypeptide` *is_part_of* `Protein` *participates_actively_in* `Process` *gives* `Metabolite` *taken_by* `Process` *has_active_participant* `Protein` *has_part* `Polypeptide` *encoded_by* `Gene`
ph. interaction	`Gene` *encodes* `Polypeptide` *participates_actively_in* `Physical_Interaction` *has_passive_participant* `Polypeptide` *encoded_by* `Gene`
regulation	`Gene` *encodes* `Polypeptide` *is_part_of* `Transcription_Factor` *regulates* `Gene`
gen. interaction	`Gene` *participates_actively_in* `Genetic_Interaction` *has_passive_participant* `Gene`

3 RESULTS AND DISCUSSION

The work described here has four major outcomes: a knowledge network for the yeast *S. cerevisiae*; an algorithm for semantic collapsing of a network; a Cytoscape plugin for visualization of the knowledge network implementing this algorithm; and a set of hypotheses relating to telomere maintenance in *S. cerevisiae*, generated using this approach.

3.1 The Ondex *Saccharomyces* knowledge network

Using the Ondex data integration platform, we integrated five publicly available data sources covering various aspects of *S.cerevisiae* biology into a semantically enriched and interlinked knowledge network. The resulting knowledge network consists of a total of 240964 concepts of 59 different types, connected by 754391 relations of 28 types.

3.2 OndexView: a Cytoscape plugin

We implemented a new method for semantic simplification as the Cytoscape plugin OndexView, that uses views on the underlying data. Users can load integrated Ondex networks into OndexView and open views on them.

A user can invoke predefined views over the knowledge network or define custom views. A built-in motif editor is included in OndexView for this purpose. To apply a view, a concept type is selected from the knowledge network. The user then chooses one or more motifs for the selected concept type. OndexView extracts all the required data from the underlying Ondex graph and constructs associations according to the algorithm outlined in Section 2.3. The program then offers a query interface to the user, which can be used to focus on specific nodes (such as genes) or collections of nodes (such as pathways and complexes) and their neighbourhoods. Once the query has been processed, OndexView displays the selected view in the main Cytoscape window.

3.3 Telomere maintenance in *S. cerevisiae*

BMH1 and *BMH2*, despite their 91.6% sequence identity, have very different interactions with the temperature-sensitive telomere uncapping mutant *cdc13-1*. Our analysis of the Ondex *Saccharomyces* knowledge network, conducted using OndexView, led to the formulation of several hypotheses to explain the different interaction profiles of *BMH1* and *BMH2*.

Table 3. Gene groups in the neighbourhood of *BMH1* and *BMH2* in the created view and their cardinalities

	Total	Cell cycle related	Glucose metabolism related	Histone related
		No. (%)	No. (%)	No. (%)
Neighb. of *BMH1*	68	15 (22.1)	7 (10.3)	4 (5.9)
Neighb. of *BMH2*	79	15 (18.9)	8 (10.1)	4 (5.1)
Intersection of *BMH1/2* neighbourhoods	28	7 (25.0)	3 (10.7)	4 (14.3)
Union of *BMH1/2* neighbourhoods	119	23 (19.3)	12 (10.1)	4 (3.3)

From the semantically collapsed graph, it is apparent that the neighbourhoods of *BMH1* and *BMH2* have only 23.5% of genes in common; *BMH1* and *BMH2* have completely separate sets of transcriptional regulators. This information is not readily apparent in the uncollapsed graph. Analysis of the node description fields reveals that 19.3% of the two genes' combined neighbourhood are cell cycle-related genes, while 10.1% are glucose metabolism related genes. Of the genes that physically interact with both *BMH1* and *BMH2*, 14.3% are histone related (Table 3). An annotated screenshot of this view is available as Supplementary Figure S2.

The over-representation of genes involved in regulation of the cell cycle[1] in the joint neighbourhood of *BMH1* and *BMH2* suggests that *BMH1* and *BMH2* may function as cell cycle regulators. Notably, Rad53, a key element of the cell's checkpoint signalling pathway, interacts physically with Bmh2. Further, two genes that have previously been identified as suppressors of *cdc13-1*: *BNR1* and *CYK3*, are also present in the neighbourhood of *BMH1* and *BMH2*. The joint neighbourhood of *BMH1* and *BMH2* also contains a relatively large number of genes related to regulation of glucose metabolism,[2] indicating that the pair of genes may also play a major role in the regulation of glucose metabolism (Bruckmann *et al.*, 2007).

3.4 Differential regulation of *BMH1* and *BMH2*

The semantically collapsed view of the neighbourhood of *BMH1* and *BMH2* (Section 2.4) shows that Bmh2 physically interacts with the protein Rad53. After selecting Rad53 in the view, we can learn from its description field that Rad53 mediates the activation of the cell-cycle checkpoint (Schwartz *et al.*, 2002), thus interacting with the *cdc13-1* phenotype. Examination of the interaction between Bmh2 and Rad53 reveals an annotation indicating that this interaction was originally reported by Usui and Petrini (2007). These authors showed that both Bmh1 and Bmh2 directly bind to the active (phosphorylated) Rad53 protein, thus enhancing its signalling effect. An edge between *BMH1* and *RAD53* is absent in the Ondex network, since Usui and Petrini experimentally verified only the physical

[1]Over-represented with respect to the distribution of GO terms in the *S.cerevisiae* genome; hypergeometric test, $P = 0.00046$.
[2]Over-represented with respect to the distribution of GO terms in the *S.cerevisiae* genome; hypergeometric test, $P = 0.0079$.

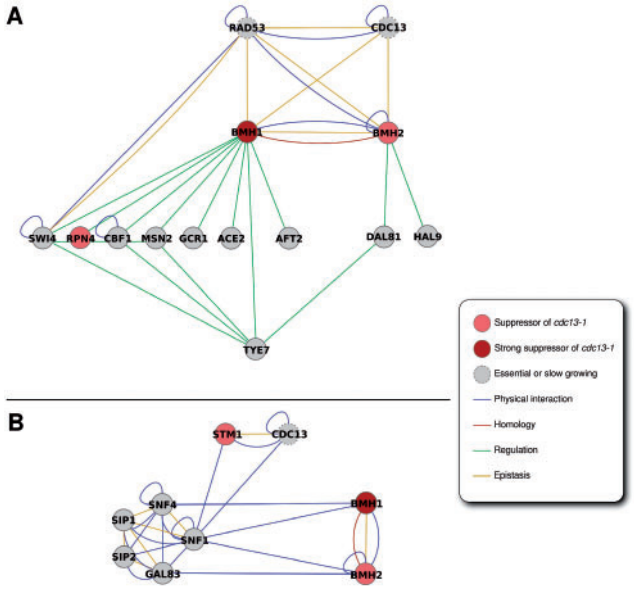

Fig. 3. Screenshots from OndexView, summarizing the hypotheses presented. Edges represent physical interactions (blue), homology (red), regulation (green) and epistasis (yellow). Nodes represent genes, some of which are weak (light red) and some strong (dark red) suppressors of *cdc13-1*. Genes which show lethal or slow growing phenotypes upon deletion are marked with dashed outlines, as knowledge of their epistatic behaviour can be expected to be incomplete.

interaction between Bmh2 and Rad53. However, they hypothesise that Bmh1 interacts with Rad53 in the same way that Bmh2 does.

Accepting Usui and Petrini's assumption, a significant difference in protein abundance during the time of checkpoint induction could explain the different effects of *BMH1* and *BMH2* deletions. The collapsed network neighbourhood of *BMH1* and *BMH2* shows clearly that the two genes are regulated completely independently from one another. Furthermore, according to their description fields in the network, two of *BMH1*'s transcriptional regulators, *ACE2* and *SWI4*, are known to be active during G_1 phase (Andrews and Moore, 1992; McBride *et al.*, 1999). This observation leads us to conjecture that *BMH1* could be more strongly expressed than *BMH2* during the G_1 phase, resulting in a higher abundance of Bmh1 protein during and after G_1 phase.

We hypothesize that due to its independent transcriptional regulation, *BMH1*'s products are present at higher levels than those of *BMH2* at checkpoint time. Thus, deleting *BMH1* would remove the majority of the Bmh1 and Bmh2 protein available as binding partners for Rad53, severely impairing the cell's checkpoint response in the face of uncapped telomeres. Removing *BMH2*, in contrast, removes a smaller proportion of binding partners for Rad53, thus resulting in a milder suppression effect on *cdc13-1*.

The differential regulation hypothesis has been summarized in Figure 3A, by generating a neighbourhood graph of *BMH1* and *BMH2* and removing all nodes except for those mentioned in the above discussion.

Experimental validation of the differential regulation hypothesis could be performed in several ways. Expression profiling of *BMH1* and *BMH2* over the course of the cell cycle could be performed. Consultation of pre-existing data has so far been inconclusive: a time course microarray performed by Spellman and colleagues showed *BMH1* expression levels to be slightly higher than those of *BMH2* throughout the the cell cycle and spiking 4-fold during G_1 phase (Spellman *et al.*, 1998) (Supplementary Fig. S2). However, due to the low resolution of the assay, the observed peak is represented by only one datapoint, rendering its significance rather doubtful. Another possibility would be the examination of an *ace2*Δ*swi4*Δ double mutant, which under the above hypothesis should replicate the *bmh1*Δ phenotype regarding suppression of *cdc13-1*.

3.5 Phosphorylation of Cdc13

A different hypothesis was formed after further examining the OndexView network neighbourhood of *BMH1* and *BMH2*. The network shows a set of edges, indicating that Bmh1 and Bmh2 physically interact with Snf1, Snf4 and Gal83, which according to their node description fields, are members of the AMP-dependent kinase (AMPK) complex. Snf1 also interacts with Cdc13 and Stm1, another telomere-capping protein. The publication linked in the Ondex knowledge network reveals that this interaction is a phosphorylation, (Ptacek *et al.*, 2005).

Publications linked from the nodes reveal that the AMPK complex is a heterotrimer composed of the α-subunit Snf1 (the actual kinase), its activating γ-subunit Snf4 and a third component β-subunit that tethers Snf1 and Snf4 together. Sip1, Sip2 and Gal83 compete for the place of this third component (Jiang and Carlson, 1997). These three competing proteins have been shown to determine the AMPK's substrate specificity and its cellular localization (Lin *et al.*, 2003; Schmidt and McCartney, 2000). For example, the Gal83 variant of the AMPK complex has been shown to be able to enter the nucleus (Vincent *et al.*, 2001).

In the network, two edges exist indicating that Bmh1 and Bmh2 both physically interact with Snf1 (Elbing *et al.*, 2006). However, Bmh1 alone interacts with the γ-subunit Snf4 (Gavin *et al.*, 2002), while Bmh2 alone interacts with the β-subunit Gal83 (Krogan *et al.*, 2006).

The nature of these interactions is currently unknown, but 14-3-3 proteins have been reported to affect kinases in several different ways, including scaffolding and direct alteration of the target's function (Tzivion *et al.*, 2001). If Bmh1/2 are involved in scaffolding for the formation of different AMPK variants, then the deletion of *BMH1* and the subsequent over-representation of Bmh2 dimers could favour the formation of Gal83-AMPK variants which can enter the nucleus to phosphorylate Cdc13. We, therefore, hypothesize that the phosphorylation of Cdc13 could potentially affect its temperature sensitivity and thus the *cdc13-1* phenotype.

The phosphorylation hypothesis has been summarized in Figure 3B, by generating a neighbourhood graph of *BMH1* and *BMH2* and removing all nodes except for those mentioned in the above discussion.

Experimental validation of the phosphorylation hypothesis is more difficult. An examination of the phenotype of *gal83*Δ mutants could offer further insight. If such a test would corroborate the importance of the AMPK complex for the observed phenotype, one could examine whether *in vitro* phosphorylation of Cdc13 by the AMPK is possible. However, without further evidence we have to consider the effect of a potential phosphorylation of Cdc13 on its temperature sensitivity in particular to be mere speculation.

3.6 Impact of noisy data

Like any other data integration approach, the Ondex *Saccharomyces* knowledge network is limited by the state of knowledge contained in its data sources. There are two main types of errors that affect the system. (i) Incorrect information from the databases, such as curator errors, will also be present in the integrated network, unless detectable by Ondex's inconsistency checks as discussed in Section 2.2. For example, contradictions between data sources can be detected and rectified. (ii) Information that is missing from the source databases will also be missing in the integrated dataset. A particular problem in this respect is the lack of negative knowledge recording throughout the systems biology community. In many sources, it is not possible to decide if non-existence of a database entry is indicative of the subject being known not to exist or not being known to exist.

It is obvious that such problems also impact downstream analyses with OndexView. However, by enabling the user to review the evidence underlying the integrated data, she/he can be pointed at the original publications that can be consulted. For example, as described in Section 3.4, the publication linked from the physical interaction edge between Bmh2 and Rad53 clarified the circumstances around the non-existence of an edge between Bmh1 and Rad53. So it can be argued that performing visual analyses with OndexView on the Ondex *Saccharomyces* knowledge network also impacts on missing data, as it allows for clarifications and corrections.

3.7 Comparison to related works

While the Ondex *Saccharomyces* knowledge network in conjunction with OndexView's semantic simplification method shows parallels to previous data integration approaches, there are a number of important differences. Like the work presented in this article, semantic web approaches, such as YeastHub (Cheung *et al.*, 2005) store their integrated data in a machine-interpretable fashion, thus providing flexible platforms that can be adapted for various purposes. However, unlike OndexView, such systems require complex querying language constructs to access them. Data warehousing approaches such as SAMBA (Tanay *et al.*, 2004) and EcoCyc (Keseler *et al.*, 2010), on the other hand, are related to the presented work in that they collect the contents of heterogeneous data sources in one centralized repository. Unlike the Ondex *Saccharomyces* knowledge network, they offer web interfaces, which makes them very easy to query for standard use. However, they are based on less semantically rigorous data structures, rendering them less flexible. In summary, the Ondex *Saccharomyces* knowledge network combines aspects from both semantic web approaches and data warehousing approaches; featuring both their strengths, but also some of their weaknesses.

4 CONCLUSIONS

The data produced by high-throughput approaches has the potential to revolutionize our understanding of biology, but can do so only if the computational techniques necessary to visualize and analyse the data can scale with the amount of data generated. Data integration methods have proven to be a successful way to face this challenge. We have shown that a view-based approach can be a powerful tool to simplify the visualization of the complex knowledge networks generated by semantic data integration, without loss of the underlying information. The view-based approach provides concise visualizations tailored to providing only the information relevant to a particular investigation.

Biological phenomena such as the different phenotypes arising from deletion of *BMH1* and *BMH2* emerge from the interplay of several independent molecular biological networks. OndexView enables users to visualize not only these networks but also the ways in which they dovetail. This visualization serves as a starting point for the user, who can now easily explore genes, their various relationships and the underlying evidence, thus gathering inspirations facilitating the generation of testable hypotheses regarding the functions of these genes.

4.1 Outlook

There are a number of improvements that could be applied to OndexView and the presented knowledge network, as well as a number of ideas that build upon this work. The exploration of evidence trails behind the simplified edges in OndexView could be made more easily accessible. Rather than showing evidence in a textual representation in graph attributes, an option to visualize it graphically on demand would further increase the tool's usefulness. However, due to the limitations of the Cytoscape API, which renders graph views immutable, such new features will more likely be included in a re-implementation of OndexView for Ondex's own graph visualization frontend.

Furthermore, the Ondex *Saccharomyces* knowledge network could be enriched with probabilities on the connections between concepts, which could be inferred from the original data sources as well as the evidence coverage. Then, if more experience can be gathered on molecular pathways that potentially qualify for explaining observed epistasis effects, they could be generalized into a collection of semantic motifs. Instances of these motifs in the graph could be ranked according to their overall probability. The generation of such ranked lists could further assist biologists with the formation of hypotheses.

ACKNOWLEDGEMENTS

J.W. and M.P. created the Ondex yeast dataset and all required Ondex workflow modules not already present in the Ondex distribution; J.W. and A.W. designed the view-based visualization approach; J.W. implemented the OndexView plug-in for Cytoscape; J.W., E.H. and J.D. investigated BMH1/2 epistasis with OndexView and generated the presented hypotheses; J.W., S.J.C., P.L., J.H. and A.W. wrote the paper; S.J.C., P.L., D.W., D.L., J.H. and A.W. supervised the project. Furthermore, the authors would like to thank the Ondex development team and the Newcastle Integrative Bioinformatics writing group for their help.

Funding: The authors are pleased to acknowledge funding from the Biotechnology and Biological Sciences Research Council (BBSRC) Systems Approaches to Biological Research (SABR) initiative [Grant number BB/F006039/1].

Conflict of Interest: none declared.

REFERENCES

Achard,F. *et al.* (2001) XML, bioinformatics and data integration. *Bioinformatics*, **17**, 115–125.

Addinall,S.G. *et al.* (2008) A genomewide suppressor and enhancer analysis of *cdc13-1* reveals varied cellular processes influencing telomere capping in *Saccharomyces cerevisiae*. *Genetics*, **180**, 2251–2266.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Andrews,B.J. and Moore,L.A. (1992) Interaction of the yeast Swi4 and Swi6 cell cycle regulatory proteins in vitro. *Proc. Natl Acad. Sci. USA*, **89**, 11852–11856.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Balaji,S. *et al.* (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.

Blasco,M.A. (2007) Telomere length, stem cells and aging. *Nat. Chem. Biol.*, **3**, 640–649.

Brinkley,J.F. *et al.* (2006) A framework for using reference ontologies as a foundation for the semantic web. *AMIA Annu. Sympos. Proc.*, 96–100.

Bruckmann,A. *et al.* (2007) Post-transcriptional control of the *Saccharomyces cerevisiae* proteome by 14-3-3 proteins. *J. Proteome Res.*, **6**, 1689–1699.

Chaudhri,M. (2003) Mammalian and yeast 14-3-3 isoforms form distinct patterns of dimers *in vivo*. *Biochem. Biophys. Res. Commun.*, **300**, 679–685.

Cherry,J.M. *et al.* (1997) Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, **387** (Suppl. 6632), 67–73.

Cheung,A.L.M. and Deng,W. (2008) Telomere dysfunction, genome instability and cancer. *Front. Biosci. J. Virt. Lib.*, **13**, 2075–2090.

Cheung,K.-H. *et al.* (2005) Yeasthub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics*, **21** (Suppl. 1), i85–i96.

Elbing,K. *et al.* (2006) Purification and characterization of the three Snf1-activating kinases of *Saccharomyces cerevisiae*. *Biochem. J.*, **393**(**Pt 3**), 797–805.

Garvik,B. *et al.* (1995) Single-stranded DNA arising at telomeres in *cdc13* mutants may constitute a specific signal for the RAD9 checkpoint. *Mol. Cell. Biol.*, **15**, 6128–6138.

Gavin,A. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Haider,S. *et al.* (2009) BioMart central portal–unified access to biological data. *Nucleic Acids Res.*, **37** (Suppl. 2), W23–W27.

Heimbigner,D. and McLeod,D. (1985) A federated architecture for information management. *ACM Trans. Inf. Syst.*, **3**, 253–278.

Herrgard,M.J. *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, **26**, 1155–1160.

Jiang,R. and Carlson,M. (1997) The Snf1 protein kinase and its activating subunit, Snf4, interact with distinct domains of the Sip1/Sip2/Gal83 component in the kinase complex. *Mol. Cell. Biol.*, **17**, 2099–2106.

Keseler,I.M. *et al.* (2010) Ecocyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, 334–337.

Kohler,J. *et al.* (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**, 1383–1390.

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

Lin,S.S. *et al.* (2003) Sip2, an N-myristoylated β-subunit of Snf1 kinase, regulates aging in *Saccharomyces cerevisiae* by affecting cellular histone kinase activity, recombination at rDNA loci, and silencing. *J. Biol. Chem.*, **278**, 13390–13397.

Longhese,M.P. (2008) DNA damage response at functional and dysfunctional telomeres. *Genes Dev.*, **22**, 125–140.

McBride,H.J. *et al.* (1999) Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation. *J. Biol. Chem.*, **274**, 21029–21036.

Nugent,C.I. *et al.* (1996) Cdc13p: a single-strand telomeric DNA-binding protein with a dual role in yeast telomere maintenance. *Science*, **274**, 249–252.

Prlić,A. *et al.* (2007) Integrating sequence and structural biology with DAS. *BMC Bioinformatics*, **8**, 333.

Ptacek,J. *et al.* (2005) Global analysis of protein phosphorylation in yeast. *Nature*, **438**, 679–684.

Sandell,L.L. and Zakian,V.A. (1993) Loss of a yeast telomere: arrest, recovery, and chromosome loss. *Cell*, **75**, 729–739.

Schmidt,M.C. and McCartney,R.R. (2000) β-subunits of Snf1 kinase are required for kinase function and substrate definition. *EMBO J*., **19**, 4936–4943.

Schwartz,M.F. *et al.* (2002) Rad phosphorylation sites couple Rad53 to the *Saccharomyces cerevisiae* DNA damage checkpoint. *Mol. Cell*, **9**, 1055–1065.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Smith,B. *et al.* (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.

Spellman,P. T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.

Tanay,A. *et al.* (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl Acad. Sci. USA*, **101**, 2981–2986.

Taubert,J. *et al.* (2007) The OXL format for the exchange of integrated datasets. *J. Integr. Bioinformatics*, **1**, 62.

Tong,A.H.Y. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.

Tzivion,G. *et al.* (2001) 14-3-3 proteins; bringing new definitions to scaffolding. *Oncogene*, **20**, 6331–6338.

Usui,T. and Petrini,J.H.J. (2007) The *Saccharomyces cerevisiae* 14-3-3 proteins Bmh1 and Bmh2 directly influence the DNA damage-dependent functions of Rad53. *Proc. Natl Acad. Sci. USA*, **104**, 2797–2802.

Vincent,O. *et al.* (2001) Subcellular localization of the Snf1 kinase is regulated by specific β subunits and a novel glucose signaling mechanism. *Genes Dev.*, **15**, 1104–1114.

Weinert,T.A. and Hartwell,L.H. (1993) Cell cycle arrest of *cdc* mutants and specificity of the *RAD9* checkpoint. *Genetics*, **134**, 63–80.