Quality analysis and integration of large-scale molecular data sets

Lars J. Jensen and Peer Bork

One of the major challenges in bioinformatics today is to integrate and interpret the heterogeneous biological data that are being produced at an ever increasing pace. As this type of analysis is still in its infancy, all studies so far have relied on applying simple rule-based criteria on only a small subset of the available data. To enable comprehensive studies to be undertaken with a statistical framework, standardized repositories from which all datasets can be easily obtained and benchmarks that quantify the often high error rates of large-scale datasets are needed. Quality control, benchmark and integration efforts from protein interaction networks in the context of genome and transcriptome data are reviewed.

Lars J. Jensen
Peer Bork*
European Molecular Biology Laboratory
Meyerhofstrasse 1
D-69117 Heidelberg, Germany
Max-Delbrück-Centre for Molecular Medicine
Robert-Rössle-Strasse 10
D-13092 Berlin, Germany
*e-mail: bork@embl.de

▼ As molecular biology is changing from a descriptive to a quantitative discipline, data quality control is becoming increasingly important. Recent years have brought many new techniques for large-scale data acquisition, exemplified by genome sequencing methods, DNA microarrays technology and large-scale detection of protein–protein interactions.

After a short period of initial excitement (Table 1), data generation within each of these three fields was followed by discussions initially on data quality and subsequently on standards, annotation and interpretation. Quality assessments or benchmarks have revealed that the quality of the raw data was far worse than initially thought. Consequently, experimental protocols have been improved, for example, by performing replicate experiments. In addition, computational methods have been developed to identify and correct errors that exist within the data. However, as larger quantities of data accumulate standardization and annotation become increasingly important for interpreting and integrating the datasets.

In this review we briefly summarize recurring patterns for genome and transcriptome analysis, and discuss in more detail the powers and pitfalls of protein interaction data, one of the youngest of the large-scale data sources.

Genome sequencing

When tackling large complex systems it is essential to have at least a good overview of the parts list. In this respect, genome sequencing and annotation play a crucial role for biological systems, as an entirely sequenced and correctly annotated genome specifies the parts list for the entire organism.

Because sequencing technology was established long before the first large-scale sequencing efforts began in the early 1990s {exemplified by the first EST (expressed sequence tag) sets [1] and *Saccharomyces cerevisiae* chromosome III [2]}, standards and public repositories were already in place. Although both of these sequencing efforts [1,2] represented milestones in the era of genomics, the sequence quality was far from perfect. Learning from these first lessons, the standards were soon raised and the sequence quality improved considerably for shotgun and contig-based approaches (yeast chromosome III was fortunately completely re-sequenced at a later date to ensure proper quality).

Despite the yeast genome sequence being completed in 1996 [3], the parts list remains heavily debated with predicted gene numbers ranging from 4500 to 6500, although recent work has convincingly narrowed the value to around 5500 genes [4]. In metazoan genomes in general, and particularly in mammalian genomes, the current state of gene discovery is considerably worse as a result of complex gene structures; in a recent update of the *Drosophila* genome annotation, 85% of the annotated gene structures were changed [5]. This level of uncertainty is particularly worrying given that gene annotations often form the basis for interpreting other types of large-scale data.

Consequently, functional annotation remains the major challenge in genome sequencing,

1741-8372/04/$ – see front matter PII: S1741-8372(04)02408-9

Table 1. Timeline for transcriptome and interactome analysis[a]

	Year of advances in large-scale data types	Refs
First large-scale data set	1997, 2000	[9,13]
Concern over data quality	2001	[10,14]
Quantitative assessment of error rates	2001, 2002	[11,17]
Normalization and error detection	2001, 2003	[11,19]
Standards for reporting experiments	2001	[12]

[a]The general trends for new large-scale data types are indicated. Equivalent advances were made in the same order for both gene expression data and protein interaction data. In both cases, it took around three years from the first data sets were published until the first methods for correcting biases and errors had been developed. Furthermore, standards for reporting measurements and experimental conditions are typically developed at a very late stage. There is currently no standard for protein interaction data.

with the standards for annotation being far less than those for sequence quality. Immediately after the sequence of the first yeast chromosome was published, it was noted that function prediction could be improved considerably by using bioinformatics tools [6]. Even after multiple genomes have been sequenced, the annotation of genes differs considerably depending on the groups involved. For example, although the two consortia that produced the draft sequences of the human genome both made confident predictions of around 26 500 genes, their protein domain counts and consequent functional predictions using almost identical procedures differed by ~30% [7,8]. Expert-driven, manual and careful annotation, as it is being done for several genomes, is therefore crucial. However, comparative and integrative approaches remain hampered by the fuzziness of the term 'function'. When comparing orthologous genes between species it is usually the exception when respective annotation fields in the species-dependent databases are identical.

Transcriptome profiling

The first genome-wide gene expression experiment was published in 1997 [9]. In these early days of microarray analysis, statistical methods were rarely used for analyzing microarray data. Typically, all genes that were more than twofold up- or downregulated on an array were taken as being 'significantly' regulated. To our knowledge, Tsien *et al.* [10] were the first to point out the serious shortcomings of this early practice. Today, it is common knowledge that expression ratios are often intensity-dependent. Not only does noise-affect cause log-ratios for low-expressed genes to have higher variance than their more highly expressed counterparts, systematic biases in the log-ratios as function of intensity are also observed. These effects can be largely corrected experimentally by making die-swap replicates, and computationally using non-linear normalization techniques [11]. A vast variety of such methods are known today, the more successful of which rely on either loess regression or quantile normalization. This, together with a community acceptance of the need for doing replicate experiments, means that quality control of raw data is generally in place. Although numerous clustering and other analysis tools have also been developed, interpretation of microarray expression data is currently hampered by the lack of systematic annotation of most of the already published microarray data. It was only from the latter half of 2002 that the first journals required the use of a minimalist annotation standard (MIAME, minimum information about a microarray experiment) [12], which has been available for some time.

Protein–protein interaction screening

Large-scale interaction datasets are even younger than the other two types of data mentioned so far. The first genomics scale set was presented in 2000 by Uetz *et al.* [13], and was soon followed by several others [14–16]. Although methods for detecting protein interactions had been used at a smaller scale much earlier, standards and public repositories for such data had not been developed before the arrival of large-scale datasets and are currently still under development. Based on the surprisingly low overlap between the two first datasets, concern about the data quality had already been raised in 2001 by Ito *et al.* [14]. In parallel with the development of high-throughput assays, *in silico* interaction prediction methods progressed towards comparable quality, as was revealed in 2002 by a comprehensive quality assessment against known protein complexes from the MIPS (Munich information center for protein sequences) database [17,18]. We are currently still in the phase of establishing quality control methods to correct for shortcomings of high-throughput interactions screens, although significant progress has recently been made in distinguishing true interactions from false positives [19]. Analysis procedures for interpreting the raw data and identifying protein complexes are also currently emerging [20,21].

What is an interaction?

The interactome is less well defined than the genome and the transcriptome, as different communities use the term protein interaction to refer to anything from physical interactions to broadly defined functional interactions, such as neighbors in metabolic networks. Even if restricted to

physical interactions, it is important to discriminate between stable interactions and transient interactions: some proteins form stable complexes, whereas others only interact weakly or for short moments of time. An example of the latter is a protein kinase and its substrates. Although not intended, yeast two-hybrid (Y2H) assays have been shown to detect at least some transient interactions, whereas most of the interactions found by complex pull-down methods are, as the name suggests, stable complexes [22,23].

Comparison of different interaction sets is further complicated by the different nature of the datasets: Y2H experiments are inherently binary, whereas pull-down experiments and the MIPS database report larger complexes. To allow for comparisons, complexes are typically represented by several binary interactions, even when comparing different sets of complexes. However, it is important to realize that there is not a single, clear definition of a 'binary interaction'. In case of the MIPS protein complexes, the matrix representation, in which each complex is represented by the set of binary interactions corresponding to all pairs of proteins from the complex, is almost exclusively used. For complex pull-down experiments, two different representations have been proposed: the matrix representation and the spoke representation in which only bait–prey interactions are included (Figure 1) [17,24]. The binary interactions obtained using either of these representations are somewhat artificial as some interacting proteins might in reality never touch each other and others might have too low an affinity to interact except in the context of the entire complex bringing them together. Even in the case of Y2H assays, which inherently report binary interactions, not all interactions correspond to direct physical interactions. This is especially true for interactions reported between nuclear proteins in yeast, where indirect interactions involving other nuclear proteins can easily occur.

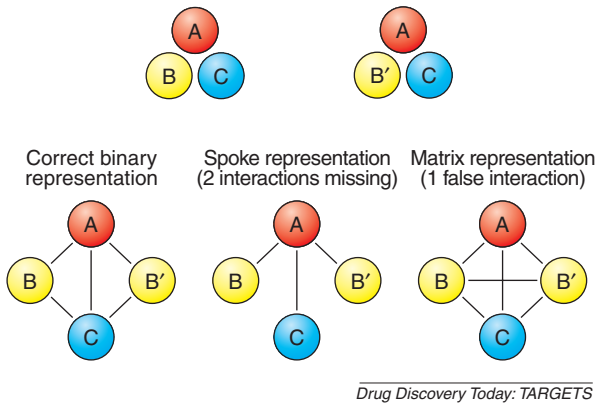

Figure 1. Spoke versus matrix representation. The two different binary representations of complex pull-down experiments each has its own pros and cons. The spoke representation will miss many true interactions unless every protein has been used as bait; in the example where A is used as bait, the interactions B–C and B'–C are missed. The matrix representation will capture more true interactions from sparse data sets, however, it is also prone to inferring some incorrect interactions such as B–B'.

Quality assessment of protein interaction networks

To make the best possible use of the various large-scale datasets, particularly the interaction datasets, it is important to quantify the accuracy of the different techniques. Ideally, such benchmarks should be performed by comparing the pairwise interactions or complexes that are suggested by each method with a reference dataset, which should be highly accurate and unbiased. Although accurate datasets do exist, such as the complexes in the MIPS [18] or PDB [25] databases, these are often redundant and strongly biased towards the types of proteins and protein complexes that are studied the most. In addition, internal comparisons of the different datasets with each other can also be quite informative. Several quality evaluations based on both types of comparisons have been published [17,22,24,26].

Poor agreement between the different protein interaction datasets is generally observed. However, given the lack of agreement on what constitutes an 'interaction', this is less of a surprise. Therefore, the small overlap of the datasets might be due to the different methods complementing each other by identifying different subsets of the interactome. In this context it is worth noting that the best agreement between the methods mentioned above compared with random expectation has been obtained for the two Y2H datasets [13,14], although this could alternatively be explained by common systematic errors of Y2H assays.

A more indirect approach for evaluating the quality of protein–protein interaction sets is to compare the suggested interactions with known subcellular localizations or protein functional classes, such as Gene Ontology [17,24,27]. The underlying assumption of such analysis is that interaction partners should belong to the same category, the validity of which depends strongly on the choice of classes. Co-expression of the corresponding genes has also been used as an evaluation criterion [17,23,28]. Although these methods can certainly give an idea of the relative accuracy of the different datasets, obtaining quantitative performance estimates this way is difficult to impossible, as they do not directly relate to physical protein interactions but rather to functional relationships in general.

There are therefore many reasons why slightly different conclusions are obtained regarding the relative accuracy of

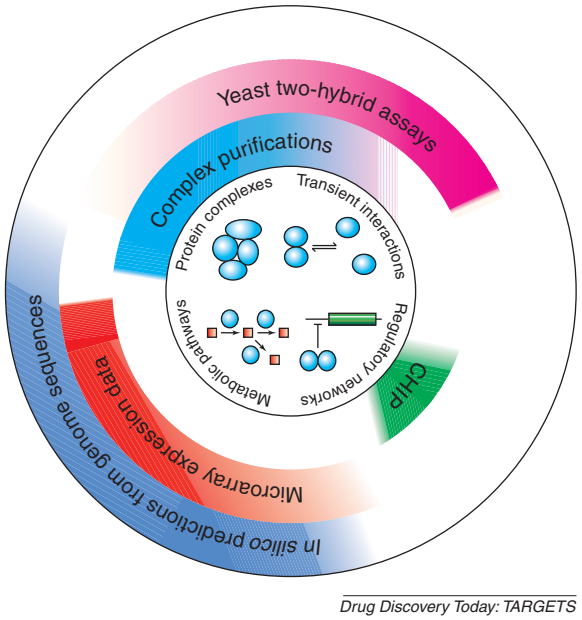

Figure 2. Overview of the interaction classes covered by different types of large-scale data sets. The colored regions show the classes of interactions on which information can be obtained from each data type. As a consequence of the partial overlap of the data types, both agreement and disagreement can provide information of the class of interaction.

the different assays depending on the type of benchmark used. For instance, using a reference set of stable complexes will tend to favor complex pull-down experiments, for which the estimated accuracy and coverage further depends on the choice of binary representation, as the matrix representation gives better coverage but lower accuracy than the spoke representation [17,24]. Similarly, one would expect *in silico* predictions to be favored by benchmarks that compare functional classes [17]. Despite these many complications, the general conclusion remains the same: the error rate on large-scale experiments is high (in the order of 50%) [17,27,28], which is also a major reason for the poor agreement between individual datasets.

Integration of genome, transcriptome and interactome data

Integration of the various types of large-scale data is currently receiving much attention. There appears, however, to be little agreement on what exactly is meant by 'integration', not to mention how to achieve it. The word 'integration' is being attached to almost any analysis that involves the combined use of two or more large datasets.

A very down-to-earth problem when combining datasets is the lack of standardization of gene identifiers. The different types of data (or even different datasets of the same type) do not usually refer to the same gene using the same identifier. Furthermore, the various gene identifiers often originate from different databases that do not refer to each other. As a result, even apparently simple operations, such as pooling or comparing datasets, often involve a lot of work. This might partly explain why most of the data integration papers published to date make use of only a small subset of the datasets available. In an attempt to alleviate these problems, we have recently compiled lists of synonymous gene identifiers for the most prominent eukaryotic model organisms; these can be downloaded from http://www.bork.embl.de/synonyms.

Once standardization-related problems have been overcome, one of the major challenges in combining datasets is to develop analysis methods that are sufficiently robust to handle the typically high error rates, for example, the large numbers of falsely reported protein–protein interactions. Fortunately, as many people have pointed out, the confidence of binary interactions can be much improved by only considering those interactions that are supported by two or more data sources (i.e. study the overlap between networks) [17,23,26,27,29–32]. There are, however, several problems with such simplistic approaches; most importantly, a very considerable fraction of true interactions will probably be lost as very little overlap is generally observed between networks. Also, as the number of datasets increases and those that are larger in size are considered, the number of false-positive interactions that are supported by multiple datasets will increase dramatically. This has already become an issue for certain types of *in silico* interactions [33].

In addition to the integration of different types of large-scale datasets within species, integration across species through genome-wide assignment of orthologous genes can potentially increase the signal-to-noise ratio. Recent studies comparing microarray expression data for *S. cerevisiae* and *Caenorhabditis elegans* have revealed that evolutionary conservation of co-expression provides much stronger evidence for a functional relationship than co-expression in just one species [34,35]. Also, computationally predicted interactions based on conserved gene neighborhood, co-occurrence and protein fusion are by nature cross-species comparisons [33].

One problem that arises when using different data sources as independent lines of evidence is that different experiments often do not detect the same kind of interactions (Figure 2). There might be perfectly good reasons why networks derived from Y2H interaction screens show poor overlap with networks based on co-expression; one could easily imagine both clusters of co-expressed genes that do not encode physically interacting proteins or proteins that

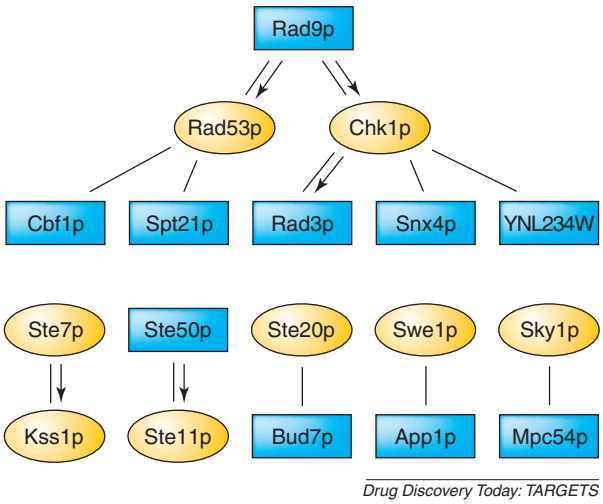

Figure 3. Combining data sources to predict transient interactions. All pairs of proteins interacting according to either the Uetz *et al.* [13] or Ito *et al.* [14] core data sets were extracted, and only those additionally supported by microarray co-expression data were kept. From this set, all interactions reported by either of the complex purification data sets [15,16] were eliminated, leaving a set of 12 binary interactions that are expected to be enriched for transient interactions. These interactions are all shown as lines. All interactions involve at least one kinase (yellow ellipses) and, with the exception of Ste7p–Kss1p, one other protein (blue boxes). Arrows indicate interactions known from the literature: Rad9p is required for activation of the Rad53p and Chk1p kinases; Chk1p phosphorylates Rad3p; Ste7p phosphorylates Kss1p; and Ste50p is a kinase activator of Ste11p. Of the remaining interactions, Ste20p–Bud7p is of particular interest as both proteins are known to be involved in bud-site selection. We therefore suggest that Bud7p either regulates Ste20p, or is itself phosphorylated by it.

interact if both are present but not co-expressed. As mentioned above, there are also differences in the types of physical interactions that can be captured by different assays.

It might therefore be fruitful to consider support, or lack of it, from the various datasets as informative when interpreting interactions. Figure 3 shows the complete set of 12 binary interactions supported by both Y2H [13,14] and microarray expression data, but which are not found by complex purification methods [15,16]. Interestingly, all interactions involve at least one kinase, and current literature indicates that five of the interactions also involve either a substrate or an activator of the kinase in question, which demonstrates that Y2H data are capable of identifying transient interactions not found by complex purification. Of the remaining seven interactions, that occurring between Ste20p and Bud7p is probably also a true transient interaction, as both proteins have been shown to be involved in bud formation.

In addition to using combined evidence to resolve the type of individual binary interactions, integration can also provide insight into complex networks where all interactions cannot be discovered by any single experimental method. An early example of this is the reverse engineering of galactose metabolism in *S. cerevisiae* from Y2H, chromatin-IP and microarray expression data [36].

Concluding remarks

The availability of a variety of genomics-scale data raises hope that a combination approach might lead to many biological discoveries. However, to facilitate integration of data it is important that all data of a particular type can be obtained from a single curated repository. Standards for how to report experimental conditions and measurements are a prerequisite for the success of such databases. Preferably, standards should be agreed upon at an early stage, before large amounts of data have accumulated. To successfully integrate the often highly error-prone large-scale datasets, comparable quality assessments (benchmarks) are equally important, as knowing the error rates permits proper statistical frameworks to be applied. So far, however, most approaches to data integration have merely looked for agreement between datasets to lower the error rate. However, it is important to keep in mind that the different methods do not all measure the same type of interactions, and that agreement cannot always be expected. There is consequently a need for more intelligent integration methods to be developed.

Acknowledgements

Lars Juhl Jensen is funded by the Bundesministerium für Forschung und Bildung, BMBF-01-GG-9817. We wish to thank Christian von Mering for help in orthology assignment and data supply and other members of the group for helpful discussions.

References

1 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656
2 Oliver, S.G. *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46
3 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 546, 563–567
4 Kellis, M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254
5 Misra, S. *et al.* (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* 3, Epub
6 Bork, P. *et al.* (1992) What's in a genome? *Nature* 358, 287
7 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
8 Venter, J. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
9 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
10 Tsien, C. *et al.* (2001) On reporting fold differences. *Pac. Symp. Biocomput.* 496–507
11 Tseng, G.C. *et al.* (2001) Issues in cDNA microarray analysis: quality

filtering, channel normalization, models of variation and assessment of gene effects. *Nucleic Acids Res.* 29, 2549–2557
12 Brazma, A. *et al.* (2001) Mininum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365–371
13 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
14 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
15 Gavin, A-C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
16 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
17 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403
18 Mewes, H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34
19 Saito, R. *et al.* (2003) Construction of reliable protein–protein interaction networks with a new interaction generality measure. *Bioinformatics* 19, 756–763
20 Bader, G.D. and Hogue, C.W.V. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4, 2
21 Krause, R. *et al.* A comprehensive set of protein complexes in yeast: mining large scale protein-protein interaction screens. *Bioinformatics* (in press)
22 Aloy, P. and Russell, R.B. (2002) The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* 27, 633–638
23 Kemmeren, P. *et al.* (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* 9, 1133–1143
24 Bader, G.D. and Hogue, C.W.V. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* 20, 991–997
25 Westbrook, J. *et al.* (2003) The protein data bank and structural genomics. *Nucleic Acids Res.* 31, 489–491
26 Edwards, A.M. *et al.* (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* 18, 529–536
27 Sprinzak, E. *et al.* (2003) How reliable are experimental protein–protein interaction data. *J. Mol. Biol.* 327, 919–923
28 Deane, C.M. *et al.* (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* 1, 349–356
29 Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86
30 Ge, H. *et al.* (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* 29, 482–486
31 Gerstein, M. *et al.* (2002) Integrating interactomes. *Science* 295, 284–287
32 Troyanskaya, O.G. *et al.* (2003) A bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. U. S. A.* 100, 8348–8353
33 von Mering, C. *et al.* (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261
34 Teichmann, S.A. and Babu, M.M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* 20, 407–410
35 van Noort, V. *et al.* (2003) Predicting gene function by conserved co-expression. *Trends Genet.* 19, 238–242
36 Ideker, T. *et al.* (2001) Integrated genomics and proteomic analysis of a systematically perturbed metabolic network. *Science* 292, 929–934