A significant proportion of knowledge about molecular biological processes is distributed over a large number of online databases. This knowledge has been obtained through experiments performed in laboratories all over the world. Overlaps often exist across the contents of these databases. The sub-discipline of integrative bioinformatics aims at collating this knowledge and making it accessible to both humans and computers. A popular integration paradigm is the construction of functional networks, which is more challenging, because each single type of association between entities to infer functional interactions. These methods work very well for functional networks. However, inferring confidence assessments for semantic networks, rather than functional networks, is more challenging, because each single type of association must be scored separately. Reliable gold standards are needed to assess the degree of confidence in each statement, that is, how likely the statement is to be correct. Several popular solutions to this problem exist for functional networks. These methods assess the quality of each input dataset against one or more additional datasets of higher quality, usually manually-curated collections. Based on the confidence measures gained from this comparison it is then possible to calculate a confidence measure for each functional interaction. The high-quality datasets used in these comparisons are often referred to as ‘gold standards’.

The method described by Lee et al. evaluates each evidential dataset against such a gold standard and obtains a log likelihood score (LLS). Subsequently, for each interaction in question, a weighted sum is formed over the LLS scores of those datasets that report the interaction. The weights are chosen in a manner that represents the degree of dependency between the datasets. Lycett describes a method that extends the original method of Lee et al. Not only one, but several different gold standards are used to generate LLS scores for the datasets. Furthermore, instead of creating a score for each interaction via the weighted sum described above, this method computes an existence probability from the original LLS scores and then averages over the different probabilities. These methods work very well for functional networks. However, inferring confidence assessments for semantic networks, rather than functional networks, is more challenging, because each single type of association must be scored separately. Reliable gold standards are needed to assess the degree of confidence in each statement, that is, how likely the statement is to be correct. Several popular solutions to this problem exist for functional networks. These methods assess the quality of each input dataset against one or more additional datasets of higher quality, usually manually-curated collections. Based on the confidence measures gained from this comparison it is then possible to calculate a confidence measure for each functional interaction. The high-quality datasets used in these comparisons are often referred to as ‘gold standards’.

The method described by Lee et al. evaluates each evidential dataset against such a gold standard and obtains a log likelihood score (LLS). Subsequently, for each interaction in question, a weighted sum is formed over the LLS scores of those datasets that report the interaction. The weights are chosen in a manner that represents the degree of dependency between the datasets. Lycett describes a method that extends the original method of Lee et al. Not only one, but several different gold standards are used to generate LLS scores for the datasets. Furthermore, instead of creating a score for each interaction via the weighted sum described above, this method computes an existence probability from the original LLS scores and then averages over the different probabilities. These methods work very well for functional networks. However, inferring confidence assessments for semantic networks, rather than functional networks, is more challenging, because each single type of association must be scored separately. Reliable gold standards are needed to assess the degree of confidence in each statement, that is, how likely the statement is to be correct. Several popular solutions to this problem exist for functional networks. These methods assess the quality of each input dataset against one or more additional datasets of higher quality, usually manually-curated collections. Based on the confidence measures gained from this comparison it is then possible to calculate a confidence measure for each functional interaction. The high-quality datasets used in these comparisons are often referred to as ‘gold standards’.
only exist for some of these types. The methods discussed above are thus only of limited use for assessing semantic networks. While solutions for specific types of data do exist, for example, PPIs (Bader et al., 2003; Artymiuk et al., 2005; Guzinski et al., 2006), these are again dependent on additional data. It would desirable to find a method that can infer confidence measures on biological networks in the absence of a gold standard, that is, based only on the existing experimental data.

To provide a more generic solution to this problem, we present a fully Bayesian approach which calculates, for each statement in a semantically-integrated dataset, the probability that it is true. We have evaluated the method’s effectiveness in comparison to related methods on simulated data with the same characteristics as real data. While desirable to find a method that can infer confidence measures on experimental measurements, is the product of the Bayes factors for these measurements with the prior odds of edge existence. The specification of prior odds $O(E^*)$ is described in the Supplementary Methods. Odds can obviously be converted into the corresponding probability using inversion.

As mentioned above, Equation 2 and thus also Equation 4 work under the assumption that experimental measurements are stochastically independent. This assumption is not valid for real data, and thus introduces a potential source of error into the methodology. Lee et al. address this problem by introducing dependency coefficients for their datasets (Lee et al., 2007), a solution which we argue is somewhat ad hoc. Instead we assume independence and focus on verifying that our method is robust to this assumption.

In order to calculate the Bayes factors above it is necessary to determine the rates of false positives and false negatives in each dataset. One approach is to compare each dataset to a gold standard and count the number of differences. However, due to the limited availability of gold standards as discussed above, this approach is not feasible here. Therefore, the only available option is to evaluate the datasets against a common consensus. A naive approach is to start with random values for the error rates, and to use these rates to create a candidate integrated network. Updated parameter values and the resulting integrated networks can then be computed iteratively.

However, a series of networks produced by this method does not typically converge to any sensible result (data not shown).

To overcome this problem, a fully Bayesian approach was employed to generate samples from the full joint posterior distribution of $\theta (\theta, G, X)$, where $\theta = (a_0, b_0, \ldots, a_n, b_n)$ is the vector of error rates associated with the members of the vector of experimental networks $X$, and where $a_i$ is the false positive rate of $X_i$ and $b_i$ is the false negative rate of $X_i$.

To determine the joint posterior distribution $\pi (\theta, G, X)$, one may exploit the following equation:

$$\pi (\theta, G, X) \propto \pi (G|X) \pi (\theta) \pi (X|\theta, G).$$

Thus, the posterior distribution $\pi (G|X)$ is proportional to the joint distribution $\pi (\theta, G, X)$. The joint distribution may in turn be factored as:

$$\pi (\theta, G, X) \propto \pi (G|X) \pi (\theta) \pi (X|\theta, G).$$

In summary, we conclude that

$$\pi (G|X) \propto \pi (\theta, G, X).$$

As a consequence, three values need to be determined: the prior distribution $\pi (G)$, the prior distribution $\pi (\theta)$ and the likelihood $\pi (X|\theta, G).$ We define the prior distribution of $\pi (G)$ as a random graph prior:

$$\pi (G) = \prod_{i=1}^{n} \pi (G_i),$$

which $q$ is the prior probability of an edge really existing.

To determine the prior distribution $\pi (\theta)$, we have to consider the nature of the error rates $a_i$ and $b_i$ as ‘success rates’ for misreading each potential edge. Modelling each observation event over a potential edge as a Bernoulli experiment with such a success rate, the number of false positives and false negatives in an experimental graph $X_i$ would follow a binomial distribution. The Beta distribution is conjugate to this binomial likelihood, and is dependent on two parameters, $\alpha$ and $\beta$. We make the assumptions:

$$a_i \sim \text{Beta}(\alpha, \beta), \quad b_i \sim \text{Beta}(\alpha, \beta).$$

For the sake of simplicity, we will later assume that all prior parameters are equal to 1, giving $\text{Beta}(1, 1)$ priors for all rates. Since sampling from $\pi (\theta, G, X)$ directly would be very difficult, we instead employ a Gibbs sampling approach (Gelfand and Smith, 1990) and alternately sample from $\pi (\theta|G, X)$ and $\pi (G|\theta, X).$
The sampling is accomplished by using the Bayesian method discussed above to infer posterior existence probabilities for each edge. These probabilities can then be used to sample a potential graph. The Gibbs sampling algorithm for sampling from their prior distribution.

The second step in each cycle is the sampling of a new error rate vector \( \theta \) based on \( G \). As explained above, Beta distributions can be used to describe uncertainty about \( \theta \). One can compare each \( X_i \) to the currently assumed true graph \( G \) and use it to count the number of supposed true positives (tp), false positives (fp), true negatives (tn) and false negatives (fn). Then, the full conditionals for \( a \) and \( b \) are as follows:

\[
\alpha_i \sim \text{Beta}(p_i + a, n_i + b) \\
\beta_i \sim \text{Beta}(n_i + a, p_i + b) 
\]

To initiate the algorithm as a whole, we need to generate an initial error rate vector \( \theta_0 \). It is sufficient to sample the initial values for each \( a_i \) and \( b_i \) from their prior distribution.

### 2.2 Evaluation method

In order to evaluate the method described above and to compare it against other probabilistic integration methods, a simulation and testing environment was created. The testing tool creates a random graph according to a specified model. In the simulation, this graph assumes the role of a true biological graph. The tool then derives a set of graphs from the true graph with predetermined error rates. In the simulation these graphs assume the role of experimental datasets. The simulated experimental datasets are subsequently passed to the integration method under investigation. The integrated graph resulting from the integration method is then compared with the original simulated true graph to evaluate the integration method’s performance.

Such a testing workflow can be programmed to be executed a large number of times in order to measure a method’s average behaviour. Furthermore, the testing tool allows for the automatic variation of different input parameters.

The simulated true graph is created as a scale-free graph using a preferential attachment algorithm, since many molecular-biological graphs have been shown to be approximately scale-free [Ravasz and Somfai 2002][1](#fn1). A description of the algorithm can be found in the Supplementary Methods. Not only does the choice of scale-free background graphs more closely match the topology of real biological graphs; it also poses an additional challenge for the new algorithm, since such graphs break with the assumption of a random graph prior distribution for \( G \).

The next problem is systematic error, also known as experimental bias. This phenomenon in particular leads to the violation of stochastic independence between datasets. We simulate this by sampling separate false negative probabilities for each interaction and false positive probabilities for each non-interacting node pair from Beta distributions. We then use these prepared probabilities to generate false positives and false negatives in the experiments, thus introducing the same bias/systematic error. More detail is provided in the Supplementary Methods.

After the simulated evidential graphs have been given to the integration method in question the resulting integrated network is evaluated against the original graph. There are a number of different quality measures that can be applied. One important aspect is to measure the accuracy of error rate estimates in the individual experiments. We can define the quadratic loss for the error estimates as follows:

\[
L_{\text{quad}} = \sum_{i=1}^{N} (\hat{\alpha}_i - \alpha_i)^2 + (\hat{\beta}_i - \beta_i)^2
\]

where \( \alpha_i \) and \( \beta_i \) are the estimates of the false positive and false negative rates for experiment \( i \) according to the integration method in question.

Also, to measure the accuracy of the final edge probabilities produced by the method, we can define further loss functions. Since we can expect a vast number of true negatives when working with sparse, scale-free graphs, it would be helpful to see the loss over interacting and non-interacting node pairs separately. These can be interpreted as analogue to the algorithm’s false negative rate and false positive rate:

\[
L_{\text{pn}} = \sum_{i \in G} (1 - \text{Pr}(e \in G_i))^2 \\
L_{\text{pp}} = \sum_{i \in \bar{G}} \text{Pr}(e \in G_i)^2
\]

We have evaluated the new algorithm in comparison with two gold standard-based methods presented by [Jeong et al. 2003][2](#fn2). One important aspect is to measure the accuracy of error rate estimates in the individual experiments.
experimental scale-free networks with 500 nodes, an average false negative rate 0.15 and average false discovery rate of 0.15 (corresponds to false positive rate of 0.0006). For the two gold standard-based methods (Lee and Lycett), one bias-free input experiment with the same FN and FDR rates was assigned as the gold standard. The additional input parameters for the gold standard-based methods (such as the dependency factor) have been set to optimal values to ensure peak performance. We have run 5000 replicates of the above test and averaged over the results. As a basic benchmark, we have also executed the same workflow for a naive integration method. The naive method simply assigns the observed proportion of experiments that support a given edge as its existence probability.

2.3 Application to biological networks

In addition to the evaluation on artificial data we applied the new method to semantically integrated biological data. We used the data integration system Ondex (Koehler et al. 2006) to gather as much data as possible on the Saccharomyces cerevisiae PPI and GI network. We imported all S. cerevisiae data from the BioGRID (Breitkreutz et al. 2006), MINT (Chatr-aryamontri et al. 2007), IntAct (Hermjakob et al. 2004) and MIPS-MPACT (Guldener et al. 2006) via the PSI-MI 2.5 XML (Kerrien et al. 2008) format. The resulting Ondex dataset represented proteins and genetic features, as well as their interactions, the experiments in which the interactions have been observed and the publications in which the experiments were described. To identify and interlink equivalent entries we used a semantic merger method. This method carefully identifies and resolves redundancies between data from the different databases, while preserving separate any separate lines of evidence. Finally, we summarized small low-throughput experiments (>20 nodes) into larger groups according to their experimental type, since low sample sizes can be expected to lead to very imprecise error rate estimates. To determine the above cut-off, we analyzed the variance of the sampling distribution with respect to the sample size. The chosen cut-off excludes datasets for which the statistical power substantially drops. Further details on the semantic integration procedure can be found in the Supplementary Material.

Having established a semantic knowledge network of physical and GIs within S. cerevisiae we executed the new algorithm over all 51,006 contained physical interactions (PSI-MI Ontology ID: Me0915) and all 15,006 contained synthetic GIs (PSI-MI Ontology ID: Me0794), as an exemplary type of GI. For comparison purposes, we also executed the naive integration method described above on the physical interaction data.

3 RESULTS AND DISCUSSION

3.1 Evaluation on simulated data

Figure 1 shows the performances of the four tested methods (Naive integration, Lee et al., Lycett and the method proposed here) over different numbers of input datasets. It is clearly visible that the new fully Bayesian method overall outperforms both gold standard-based methods. For low numbers of input datasets, the fully Bayesian method’s loss over existing interactions ($L^{\text{ex}}$) is substantially lower and is the only method to perform better than the naive approach for three input experiments. Given more than five experiments, all methods show comparably low loss of existing interactions. The variance in performance is at comparable levels in all methods.

Regarding non-existing interactions the fully Bayesian method constantly performs better than Lycett’s method. Lee’s method shows unexpected behaviour here as its performance on non-existing edges actually worsens with increasing numbers of datasets. Both the mean and the variance of its loss function increase steadily. This is most likely caused by the method not taking into account any available negative evidence. The Lee method’s central weighted sum only comprises the positive LLS scores for datasets that support an edge, but no negative components for datasets that do not support the same edge. Thus, with a rising number of datasets it becomes increasingly likely for non-existing edges to be misclassified. The proposed method, on the other hand, improves strongly with a growing number of input datasets. As expected, the SDs regarding the $L^{\text{ex}}$ metric are ~100 times smaller than those for the $L^{\text{ex}}$ metric, since the number of existing edges is very small compared with the number of potential edges.
Bayesian integration of networks without gold standards

The average loss regarding the \( \theta \) estimates (\( L_{\text{ER}} \)) shows the fully Bayesian method’s superior precision, which improves in both mean and variance with the number of input experiments, while both gold standard-based methods stagnate at constant large amounts of loss. This observation is to be expected, since these methods do not use the input datasets in order to predict error rates, but instead only rely on the gold standard.

3.2 Interaction probabilities for \( S.\) cerevisiae

As discussed in Section 2.3, we have applied the proposed method to physical and GI datasets for the yeast \( S.\) cerevisiae. Figure 2 shows histograms of the resulting existence probabilities. Regarding physical interactions, 20,180 (39.03%) of the experimentally reported interactions have been assigned probabilities <0.1. A total of 22,090 (43.3%) have been assigned probabilities >0.9. The remaining interactions are widely distributed around a small peak near probability 0.3.

When looking at the probability distribution over synthetic genetic interactions it is apparent that with 14,386 (95.8%) a vast majority of interactions have been assigned values <0.5. Of these, 8086 (53.9%) have a probability <0.1. Only 230 (1.5%) have a probability >0.9. This is most likely due to the much poorer coverage and data quality for synthetic GIs compared with physical interactions. An examination of the data shows that only 3.32% of the synthetic interactions are backed up by >2 experiments, whereas for physical interactions the same is true for 18.08%.

Even though it would be desirable to evaluate the correctness of these confidence assessments, without knowing the absolute truth regarding which interactions are real and which are not, a satisfying answer cannot be found. Apart from the discussed limitations of existing reference datasets often treated as gold standards, it would be problematic to evaluate the Lee and Lycett methods on the basis of the same gold standard data they received as input.

However, it is possible to illustrate the difference in results between the naive method and methods that take into account the reliability of each input dataset. Figure 3 shows a comparison between probabilities assigned to PPIs according to the naive method and according to the proposed method. It is clearly visible that only few agreements exist. A correlation of only 0.36 can be measured. This is to be expected. In contrast to the naive method, the proposed algorithm infers false positive and false negative error rates for each experiment and thus reaches more refined conclusions.

3.3 Summary and conclusion

We have presented a novel, fully Bayesian method for assessing the credibility of experimental network data in the absence of gold standards. We have evaluated the method’s performance in comparison with two existing gold standard-based methods using a rigorous testing environment. The new method has shown excellent performance despite the testing environment being designed to favour the competing gold standard-based methods: both competing reference methods have been given optimal input parameters as well as unbiased gold standards, which would not normally be available. Additionally, the environment has been set up to simulate scale-free graphs, which emulate the topology of real biological graphs and thus do not meet the proposed algorithm’s assumptions regarding the prior distribution of \( G \). Finally, the simulation of experimental data has been designed to introduce bias and thus violate the algorithm’s assumption of statistical independence of the datasets.

We have evaluated the performance of the new algorithm and the two reference methods using three different metrics. These metrics express the accuracy of error rate estimation as well as the accuracy of probability assignment to existing and non-existing
edges. Other performance metrics could have been used as well (e.g. ROC-AUC or weighted sums of loss functions). However, we argue that such metrics would have provided less detail, since they represent a method’s performance as a single number rather than yielding information on different categories of performance. Given biological networks or their simulated equivalents, it is crucial to highlight the difference between existing and non-existing edges, since their respective amounts are so vastly different. Thus errors on non-existing edges would always overshadow the errors on existing edges in summarizing metrics like ROC-AUC.

While we have shown that the new algorithm performs well on large input experiments, it remains yet to be shown how the method performs on smaller datasets originating from low-throughput experiments. A further limitation to the approach is that it does not take into account any potential pre-existing confidence assessments from the original experiments.

We conclude that the new fully Bayesian method is a valuable addition to the set of tools available for confidence assessment of experimental data. It is particularly useful for the application on semantically integrated knowledge networks that consist of heterogeneous data, since it allows for every sub-network to be addressed separately without need for reconfiguration or search for applicable gold standards. As shown in Section 2.3, we were able to easily apply the method to the PPI and synthetic GI sub-networks of the same dataset.

3.4 Outlook

Future work will include a further evaluation of the method on real biological data. It is possible to compare the algorithm’s performance to other methods when using the calculated confidence values for protein function prediction. Furthermore, we would intend to analyze the method’s performance on small low-throughput type input datasets. Finally, it would be interesting to explore more of the huge simulation parameter space. A great number of combinations of different numbers of small and large input experiments with varying false positive and false negative rates remain to be surveyed.

Another useful feature would be the incorporation of potential pre-existing prior information. This endeavour is largely complicated by the lack of a common standard. For example, PPI data as available in PSI-MI format from the databases mentioned above contains various different confidence values in different metrics, such as probabilities, letter grades or bit-scores. If these difficulties can be successfully tackled, the resulting improvements will make the presented method even more reliable and useful.

ACKNOWLEDGEMENTS

The authors would like to thank Matthew Pocock for his help with Java code optimization and the Newcastle Integrative Bioinformatics writing group for their suggestions towards the manuscript. D.W. and J.W. devised the proposed Bayesian method and the test harness. J.W. implemented the method, the test harness, performed the semantic integration of the \textit{S. cerevisiae} data and applied the Bayesian method on it. J.W., K.J. and J.H. re-implemented and analyzed the gold standard-based reference methods. J.W. and D.W. wrote the manuscript. J.H., S.C., P.L., A.W. and D.W. supervised the project.

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