Robust prediction of hubs in the yeast synthetic lethal network

*Daniel Schöner[1,3,4], Corinne Dahinden[2,3] , Wilhelm Gruissem[1,3], Peter Bühlmann[2,3]

[1]Department of Biology, ETH Zurich, Universitaetsstr. 2, 8092 Zurich, Switzerland
[2]Seminar for Statistics, ETH Zurich, Leonhardstr. 27, 8092 Zurich, Switzerland
[3]Competence Center for Systems Physiology and Metabolic Diseases (CC-SPMD)
[4] Present adress: Roche Diagnostics, Forrenstr, 6343 Rotkreuz, Switzerland

Email: Daniel Schöner - dhs@ethz.ch; Corinne Dahinden - dahinden@stat.math.ethz.ch; Wilhelm Gruissem - wgruissem@ethz.ch; Peter Bühlmann - buhlmann@stat.math.ethz.ch;

*Corresponding author

Abstract

In the genetic model organism Saccharomyces cerevisiae, large scale screening for synthetic lethality has greatly increased our knowledge of genetic network organization and the functional relationships between genes. Only 5% of all possible query genes in the yeast genome have been analyzed in published synthetic lethal screens, indicating that a large part of the synthetic lethal network still remains to be uncovered. We present the first reliable method to predict highly connected query genes for the synthetic lethal network from a list of non-essential genes not yet considered for screening. Since these genetic network hubs exhibit far more interactions than an average query gene, prioritizing them for screening will help to elucidate the synthetic lethal network more efficiently by minimizing the necessary experimental efforts. We apply a twofold strategy assessing pairwise relationships between non-essential genes as well as their individual properties as single genes to reliably predict genetic network hubs. Integration of the results into a unified list leads to a robust estimation of promising query genes for future experiments. Application of the method provides a roadmap for fast, comprehensive assessment of synthetic lethality in yeast. It can also be used for the design of future experiments for other biological networks.

1 Background

Technological advances in biological experimentation have enabled researchers to investigate living systems on an unprecedented scale, studying genomes, proteomes or molecular networks in their entirety. Whether gene expression analysis using microarrays, proteome analysis using mass-spectrometry or, as in the presented case, large-scale screens for genetic interactions, high-throughput technologies provide a rich source of biological information. This comprehensiveness comes at a cost, however, because experimental results from large-scale screens can be rather unspecific, biased or prone to experimental imprecision. Combining high-throughput methods with computational analysis, either prior to or after experimentation, can significantly improve the prospects of success, since it can assign probabilities for likely molecular targets and narrow down the search to refined subsets of interesting genes or proteins. In this work, we demonstrate how to combine two computational methods in order to predict yeast genes that are likely to show a large number of genetic interactions with other genes. Our methodology provides a basis for improved experimental design for the study of the yeast genetic network. Genetic interactions can be inferred on a systematic scale using synthetic lethality screening. Crossing mutants depleted of a query gene of interest into the complete selection of single-deletion mutants and scoring for lethality can uncover genetic buffering on an unprecedented scale (1; 2) and reveals pathway architecture (3) as well as the role of central proteins in the networks. Despite the wide use of this genome-scale technology the complete structure of the cellular network remains far from being understood in detail. In the genetic network of SL relationships, only a small fraction of the genome has been mapped as a query gene and approaches to reveal the complete genetic network are currently ongoing. The success and speed of this approach will largely depend on a careful selection of new query genes for the next set of experiments. While the average number of hits in a screen is 34 out of 4800 possible non-essential target genes (2), some important genes for proteins with many cellular functions exhibit significantly higher numbers of interaction partners. These hubs thus have a high degree or connectivity in the genetic network. Figure 1 shows the degree distribution of genes used as query genes in systematic SL experiments and illustrates that many query genes only have a few connections while a few genes have high degrees (> 100). Mapping genetic interactions for those hubs first will help to significantly increase the information gain for uncovering the SL network, because less experiments have to be performed for the same number of interactions to be revealed.

The properties of molecular hubs have previously been studied in other networks. Based on a large set of protein-protein interactions, Jeong et al. (4) showed that some nodes in the network show a high number of connections to the other nodes: These hubs represent proteins for essential cell functions. Ozier et al. (5) demonstrated that genetic interactions are far more likely to occur between these protein hubs and that most genetic interactions involve at least one hub protein. This suggests that hubs in the physical and the genetic network may coincide.

Because the cellular network is complex, the behavior of a genetic network is not an isolated phenomenon but strongly correlated with other measurements and networks. Several studies have therefore investigated cellular networks with respect to their topological features and their relationships to other functional genomic data. Kelley et al. (6) showed that overlaying the protein network and the genetic network allows to classify SL interactions into distinct classes with respect to pathway architecture. Wong et al. (7) exploited dependencies of SL interactions with other data types to predict synthetic sick and lethal gene pairs.

Here we made use of the available information about the SL genetic network and other genomic and proteomic data to predict genetic network hubs from a list of non-essential query genes. All of the genes in this list have not been used for systematic SL screening so far and thus represent the query gene pool from which new candidates for screening can be selected. Our approach differs significantly from the approach used by Wong and colleagues (7), who focussed on predicting genetic interactions. Not only do we predict interacting pairs but we are the first to predict genetic network hubs and propose them as the most suitable query gene candidates for further SL screens. The results are therefore of considerable practical value for designing future experiments. Our resulting list of genes serves as a guide for the elucidation of the genetic network using SL screening and represents a ranking list of putative network hubs sorted by their network connectivity.

1.1 Modeling approach

We applied a twofold strategy in order arrive at a robust estimate of those non-essential genes that will show a high number of interactions when used for systematic SL screening. We predicted (A) the connectivity of a query gene (hub prediction) and (B) the probability of SL interactions of a query gene with possible target genes (SL prediction) using Random Forests of decision trees (8; 9). Random Forest regression allowed us to predict the number of target gene hits that can be expected from SL screening of new query gene (A). In a parallel approach, we applied Random Forest classification to predict the interactions partners of a query gene and subsequently inferred the number of SL hits from these predictions (B). Random Forests have the advantage to make accurate predictions using mixed inputs ranging from real-valued to categorical data. Furthermore, the method effectively incorporates high-dimensional and correlated features. This was important for our approach, because we included a number of different highly redundant scores derived from the same data sources (see 1.2).

During model training, we applied a special leave-one-query-out cross-validation procedure that realistically mimics the situation of a new genetic interaction screen (loqo-CV, Section 5.1). For each query gene in our data set used for parameter estimation we omitted all known high-troughput SL interactions, trained the model on the remaining data and modeled the respective interactions using the selected input features. Deviations of the estimated result to the true outcome yielded the prediction errors. Since some of the input features were calculated from the synthetic lethal interaction network and were thus based exactly on the information we aimed to predict, this required some special adjustments and corrections, which are described in detail in Section 5.1.

1.2 Data-sources

In both approaches, we considered different types of genome and proteome data sets to derive some scores that can be used as input features in the applied Random Forest regression and classification models. We provide detailed descriptions of the derivation of scores in Sections 2.1.1 and 2.2.1.

Genetic Networks

We assembled the genetic networks derived from genetic information available in the GRID-database as of May 2008 (10). We included information from all genetic interaction experiments. In contrast to the response variable or the model output, which contains the number of high-throughput (htp) SL hits only, we also included small scale experiments and synthetic sick phenotypes in the network and to derive a set of scores and coefficients as input features for both approaches.

Protein Networks

We assembled the protein networks constructed from protein interaction information available in the GRID-database as of May 2008 (10). This included small and large-scale yeast two-hybrid as well as mass-spectrometic measurements of proteins from purified complexes. To derive input features we used the same measures as for the genetic networks.

Mixed Networks

In order to increase network coverage, we also considered a unified interaction network regardless of the interaction type, including protein as well as genetic interactions. From this mixed interaction network we derived the same scores as for the other networks.

Gene Expression

We considered two large compendia of gene expression profiles. Hughes et al. (11), reported gene expression results of the effects of a large set of chemical conditions and single-gene deletions. Mnaimneh et al (12) reported gene expression data for promoter silencing alleles of a large set of essential genes. We used these data to calculate input features for the classification approach.

Gene Ontology

We also included information Gene Ontology (13). For the classification approach, we obtained scores from the three categories “biological process“, “cellular component“ and “molecular function“ .

Sequence information

Basic protein or gene-features that can be derived from their DNA sequences also provided important information in our analysis. For example, we used the amino acid content or the codon bias of a protein. Other types of information included molecular weight, protein length, and GRAVY or AROMA scores that refer to hydrophobic or aromatic residue content of the protein.

Chromosomal location

We also assessed the relative positioning of the gene loci on their respective chromosomes based on their name identifiers to derive input features.

2 Results

2.1 Regression for network degree

We used a single gene approach assessing a data set comprising 256 query genes already used for htp SL screening. For each query gene in the data set we calculated the genetic network degree, i.e. the number of SL interaction partners found by htp SL screening. This value was used as the response variable or output of the model. As input features, we assembled a list of 38 decriptors from other genomic data that describe the properties of each single query gene as explained in the following.

2.1.1 Input features

Genetic networks

We assessed the network degree, by counting the number of genetic interactions per query gene. When calculating the genetic network input features for a given query gene, we excluded the SL information from the corresponding screen where the gene acts as a query gene in order to avoid redundancies between input and output. That way, the part of the genetic network revealed by mapping a given query gene was unknown when making predictions for it. In contrast to the response variable, where only htp SL information was considered, here, we also counted genetic interactions found by ltp experiments and combine weaker synthetic sick phenotypes with synthetic lethals (SSL) from other screens. The aim was to include as much knowledge about the gene as possible and use features that are was available for novel query genes, as well. We counted the interactions with a given query gene, to obtain the network degree. Even though we omitted information in order to mimic the situation of a new query gene, these features were still partially correlated with the response variable in the model, depending on how often the considered gene had been found as a target in other screens. We used these features in our input because they represented very useful information that would also be available for new genes that have not been used in systematic SL screens, so far. In fact, they were found to be a strong predictor (see Figure 3). Considering htp and ltp data seperately and also together, this resulted in the integer-valued degree scores *SSL.htp.deg*, *SSL.ltp.deg*, *SSL.deg*.

Protein networks

We used the number of links in the protein interaction network, found by yeast two-hybrid or mass-spectrometric experiments using protein purification, as the protein network degree of the query genes. Beyer et al. (14) showed that genetic interactions bridge highly connected genes in the protein interaction network. With these inputs we assessed whether such protein network hubs are also genetic network hubs. As above, we distinguished between htp and ltp data and we also used a combination of both to derive the integer-values inputs *PI.htp.deg*, *PI.ltp.deg*, *PI.deg.*

Mixed networks

We constructed a mixed biological network featuring both protein interactions and genetic interactions as in both paragraphs above. We calculated the degree in a network containing only htp or ltp data, and a combination of both. The features *Mixed.htp.deg*, *Mixed.ltp.deg*, *Mixed.deg* have integer scale.

Amino acids

Each gene or protein sequence consists of a given number of amino acids (AA) related to its functional role in the cell. We used the count of each amino acid present in a protein as separate features to characterize each query gene. Thus, these inputs would reveal whether genetic network hubs are enriched in specific residues in their amino acid profile. For the resulting integer input variables we used the three letter code of each AA: *ALA*, *ARG*, *ASP*, *CYS*, *GLN*, *GLU*, *GLY*, *HIS*, *ILE*, *LEU*, *LYS*, *MET*, *PHE*, *PRO*, *SER*, *THR*, *TRP*, *TYR*, *VAL.*

Protein sequences

Some additional, informative features can be derived from the protein sequences. We included the molecular weight and the length of the protein to see whether genetic network hubs tend to have a certain size. To capture biochemical properties of the gene products, we included the hydrophathy-score (GRAVY), the occurence of aromatic residues (AROMA) as well as the pI of the protein. We further included codon bias of the gene sequences in the form of the CAI (codon adaptation index) and FOP (frequency of optimal codons) score (15). The resulting list of inputs comprises the integer and continuous variables *Mol.weight*, *Protein.length*, *GRAVY*, *AROMA*, *Codon.bias*, *CAI*, *FOP.*

Chromosomal location

Several lines of evidence suggest that essential genes in yeast cluster at genomic locations with low recombination frequency (16). Potentially, genetic network hubs also cluster to specific genomic locations to be protected from random mutation. To investigate chromosomal locations for the genes, we calculated their ranks on each chromosome arm and assessed their relative distance to the centromer as well as the relative distance to the centromer and the telomeres. These numbers were obtained from the spatial order of the genes as given in the name identifiers. All the genes on a given chromosome were sorted for each chromosome arm (R and L: right and left of centromer) and their relative position was estimated by taking the fraction of the number in the gene identifier (eg 109 for YOL109W) and the total amount of loci on the respective chromosome arm (e.g. 166 for the left arm of chromosome "O"). This number contains crude spatial information and serves as a measure for the distance between a given locus and the centromer. We obtained the relative distance to the center of the chromosome arm, a region expected to have high random mutation rate, by substracting the number 0.5 of this value and taking the absolute value of the result. We named the resulting real-valued inputs accordingly as *gene.rel.cent* and *gene.rel.mid*. A low value for *gene.rel.cent* means that the gene is located close to the centromer and far from the telomere, while a low value for *gene.rel.mid* means the gene is located close to the center of the chromosome and thus far from both centromer and telomere.

This list of features contains correlated input information that could affect a regression model. As briefly mentioned in the introduction, the Random Forest regression model applied here is not harmed by inputs correlated to each other so that the entire list of features could be included in the analysis.

2.1.2 Cross-validation and feature selection

We trained a Random Forest regression model on the data set with all genomic inputs as explanatory variables and the genetic network degree derived from the htp SL network as the response variable. The random Forest algorithm yields the weights of each explanatory variable in the form of a variable importance score, as shown in Figure 3. We succesively deleted the least important variable from the current submodel, until the mean squared prediction error (MSE) became minimal. For a good model the MSE should be lower than the variance of the response variable. The MSE itself is estimated using leave-one-query-out cross-validation (see Section 5.1), omitting the observation corresponding to a given query gene at each run, training the model on the rest of the data and predicting the network degree of the omitted query gene. We found that a model containing the genetic and protein network inputs *SSL.htp.deg*, *SSL.deg*, *PI.ltp.deg*, *PI.deg*, *Mixed.ltp.deg*, *Mixed.deg*, as well as the amino acid counts *PRO*, *VAL* has a minimal MSE of 793.5 (with the variance of the response variable being 999.3). Additional File 1 shows the results of the loqo-CV procedure for the regression model. Note that in Figure 3 some other inputs rank higher or almost as high as these selected variables, but are not included in the finally selected model. This is because the picture shows their weights in the complete model. We assessed all reduced submodels, omitting one feature after the other. In those submodels the ranking of the remaining features can vary due to partial redundancies between the features. The final model that we selected results in a minimal MSE and thus has the strongest capacity to predict genetic network hubs. It incorporates information from genetic networks, protein networks and from the amino acid counts of proline and valine. It is interesting that htp information was more informative than ltp for the genetic network, while for the protein networks and the mixed network it is the other way round. Apparently, for synthetic lethality prediction htp protein interactions are less informative than ltp, and ltp SL interactions in the input are less informative than interactions from high-throughput experiments. A possible explanation is that for protein interactions htp experiments are not a very reliable source of information compared to ltp and for synthetic lethals htp screening provides much more comprehensive information than ltp.

2.1.3 Prediction of new network hubs

We trained a model on the input features in our final model, as described in Section 2.1.2. For the unknown part of the SL network, we assembled a new data set for all 1350 possible non-essential query genes that have not been tested yet, but were found as a target in a screen at least once. We restricted ourselves to this number instead of all 4544 remaining possible queries (4800 - 256 already tested), in order to include only those genes in our pool, that are known to be amenable for SL screening. We then used the fitted model for prediction of novel network hubs based on the selected input variables. The resulting list is sorted by the expected degree, representing the likely number of interactions it displays with other genes. It is shown in Additional File 2.

2.2 Classification for SL interaction

Here, we used an approach based on pairs of genes. We included all SL interactions from large scale-screens as positive instances. This resulted in 7626 non-redundant SL pairs. As negative samples, we produced an equal number of non-SL gene pairs by randomly shuffling the query and target genes in the list of positive SL interactions. From this list of random gene pairs we excluded any pair of genes for which a genetic interaction had been reported, also considering small-scale experiments and synthetic sick phenotypes. Thus, we only included negative pairs that have actually been tested for genetic interactions and were found to be negative. We excluded ambiguous pairs, for which a genetic interaction had been found by other experimental means. From current htp SL data, it is known, that an average query gene only interacts with a small fraction of all genes in the target gene pool (i. e. on average 34 out of 4800 (1)).

Consequently, the negative set should be much larger than the positive set. In our scenario, though, we use a balanced data set. First, as in the previous approach, we do not consider all 4800 possible targets, but a set of non-essential genes that exhibit interactions with at least one query gene. Thus, the gene pairs in the positive and the negative class only contained genes that are reportedly amenable for SL screening, and genes that are possibly not sensitive to the method are excluded. Secondly, the employed Random Forest classifier is known to be sensitive to very imbalanced input data. We thus used a reduced balanced data set, because we believe that the essential features discriminating positive from negative instances can be readily retrieved from less data. The response variable is binary and equals 1 if the considered gene pair exhibits SL and 0 otherwise. The explanatory variables consist of 45 continuous and integer scores derived from genomic and proteomic data.

2.2.1 Input features

As for the previous regression approach, we calculated different input scores that characterize each query-target gene pair based on the data described in the introduction (1.2). These scores assess pairwise relationships between gene pairs, for instance shared interaction patterns in different networks. We used the network clustering coefficients as described by Goldberg et al. (17) including the methods meet/min (mm), the Jaccard index (j) or the geometric (g) and the hypergeometric clustering (hg) coefficients to obtain real-valued inputs from these networks.

Genetic network

We assessed the network of genetic interactions considering the amount of shared interaction partners of two genes, taking into account their connectivity in the network. Note that the hypergeometric clustering coefficient used here is the same as the term "genetic congruence" used by Ye et al. (18). The resulting input features are the j, mm, g and hg clustering coefficients *j.gene.coeff*, *mm.gene.coeff*, *g.gene.coeff*, *hg.gene.coeff*.

Protein Network

We considered the network of protein interactions and calculated the same coefficients as above: *j.protein.coeff*, *mm.protein.coeff*, *g.protein.coeff*, *hg.protein.coeff*.

Ozier et al. (5) showed that protein hubs tend to show many SL relationships. They refer to these relationships as the combined physical connectivitiy, a score based on the product of the network degrees of both proteins in the protein interaction network. Here, we calculated these scores and distinguished between low- and high-throughput (htp/ltp) and the union of both (all): *C.htp*, *C.ltp*, *C.all*.

Mixed Network

For the mixed network, we applied different schemes to obtain suitable scores for classification of SL. First, we considered the union of both networks irrespective of the interaction type. The term “mixed” denotes the resulting coefficients: *j.mixed.coeff*, *mm.mixed.coeff*, *g.mixed.coeff*, *hg.mixed.coeff*.

Then, we followed the terminology introduced by Kelley et al. (6) who distinguished two types of SL interactions. According to their work, a small fraction of all SL relationships occurs within the same protein complex or pathway, bridging proteins that are densely connected with each other in the protein network. The “within pathway”-model relates to these SL interactions. The other larger part of SL occurs between pathways or protein complexes and bridges genes whose proteins are densely connected with other proteins (clusters or complexes): the “between pathway”-model. Consistent with this concept, we derived within- and between scores using protein and genetic network topology. For the between-score, common interaction partners of a pair of genes are only counted if they are connected to both genes over exactly one protein and one genetic interaction, with the protein interaction linking proteins in the same pathway and the genetic interaction bridging two pathways. For the within-score, genes connected in this two-way manner were only counted if there were also direct protein interactions to both proteins and thus, all three proteins would belong to the same physical pathway. This is consistent to the between and within models of Kelley et al. (6) as illustrated in Figures 4 and 5. Note, that the between-score is also a conceptual extension of the 2hop physical-SSL feature used by Wong et al. (7). In their approach, a single two-way link with a common interaction partner of a gene pair involving a protein and a genetic interaction resulted in a strong input feature for SL-prediction. We consider a weighted score of multiple such links in the derived ”between”-features.

According to the used terminology, the scores are: *j.between.coeff*, *mm.between.coeff*, *g.between.coeff*, *hg.between.coeff* and *j.within.coeff*, *mm.within.coeff*, *g.within.coeff*, *hg.within.coeff*.

As additional features we used a set of correlation coefficients calculated on different data sets.

Gene ontology correlations

We used the similarity score based on the Gene Ontology information as proposed by Lord et al. (19) to assess the correlations of the GO annotations between the genes in each pair. Considering all three GO-categories, namely biological process (p), cellular component (c) and molecular function (f), this resulted in the real-valued GO-similarity (GOS) scores *GOS.p*, *GOS.c* and *GOS.f*.

Gene expression

To include gene expression information in the inputs, we calculated the Pearson correlation for each gene pair based on two large compendia of gene expression profiles, the Hughes-data set (h, (11)) and the Mnaimneh-data set (m, (12)) (see input Section 1.2). These two inputs scores *mRNA.coef.h* and *mRNA.coef.m* assume continuous values between −1 and 1.

Amino acid content

To include correlations in amino acid composition among the genes, we also calculated the Pearson correlation of the amino acid profiles of the proteins corresponding to each gene a pair. This resulted in the input *aa.coef*.

Sequence-based features

As in Section 2.1.1, the molecular weight, the protein length and other coefficents that characterize a protein product, such as codon usage (*CAI*, *FOP*, *Codon.bias*) and the fraction of aromatic or hydrophobic residues were included (*AROMA*, *GRAVY*). To assess the similarity between both genes in each pair, we took the absolute difference (diff) and the log-ratio of the absolute difference between the respective values (ratio) for each gene pair. This results in the following list of integer and continuous input scores: *MW.diff*, *MW.ratio*, *PI.diff*, *PI.ratio*, *CAI.diff*, *CAI.ratio*, *Prot.length.diff*, *Prot.length.ratio*, *Codon.bias.diff*, *Codon.bias.ratio*, *FOP.diff*, *FOP.ratio*, *GRAVY.diff*, *GRAVY.ratio*, *AROMA.diff*, *AROMA.ratio*

2.2.2 Cross-validation and feature selection

As in the regression approach, we applied leave-one-query-out cross-validation for assessing the model performance (see Section 5.1). In the scenario presented here, all data from a given query gene was omitted from the data set involving all SL interactions where the given query was actually used as a query in a systematic screen, because this represents the unknown information that we were trying to predict. However, we kept interaction data from other screens in which the given gene was found as a target as well as synthetic sick interactions and results from small scale experiments, because the latter information is independent from the htp assay with the given query gene. As in the first approach, we selected the best list of input features using the Random Forest variable importance and the classification error as obtained from leave-one-query-out cross-validation. The list of selected features contains the following 37 inputs: *j.gene.coeff*, *mm.gene.coeff*, *g.gene.coeff*, *hg.gene.coeff*, *j.mixed.coeff*, *mm.mixed.coeff*, *g.mixed.coeff*, *hg.mixed.coeff*, *j.between.coeff*, *mm.between.coeff*, *g.between.coeff*, *hg.between.coeff*, *GOS.p*, *GOS.c*, *GOS.f*, *mRNAcoef.h*, *mRNAcoef.m*, *aa.coef*, *MW.diff*, *PI.diff*, *CAI.diff*, *Prot.length.diff*, *Codon.bias.diff*, *FOP.diff*, *GRAVY.diff*, *AROMA.diff*, *MW.ratio*, *PI.ratio*, *CAI.ratio*, *Prot.length.ratio*, *Codon.bias.ratio*, *FOP.ratio*, *GRAVY.ratio*, *AROMA.ratio*, *C.ltp*, *C.htp*, *C.all*.

Figure 6 shows, that the between-features are the most important inputs followed by the mixed features. From the remaining features, the genetic clustering coefficient *hg.gene.coeff* is most important. This is consistent with the approach reported by Wang (7), in which the features based on genetic interactions or combinations of genetic and protein interactions resulted in the best inputs. All the features derived from gene expression, amino acid profiles and sequences have intermediate importance in the model and are kept in the selection of inputs. The inputs based on protein interactions only and the within-scores were omitted from the model due to their very low importance. However, the combined physical connectivity score also derived from the protein interaction network is a powerful predictor for SL interaction. Considering shared partners in the protein network apparently does not help to predict SL but the combined physical connectivity calculated from a protein network with htp and ltp interactions ($C.all$) is a good predictor. This supports the hypothesis that protein network hubs are also hubs in the genetic network.

2.2.3 Statistical Assessment

Random Forest classification yields conditional probabilities for synthetic lethality as a function of the input variables. We considered the different cutoffs 0.5, 0.75 and 0.9 as thresholds for the conditional probability of synthetic lethality. Gene pairs with an estimated probability of equal or higher than these values were classified as positives and the rest as negatives. The results for all three cutoffs are shown in Additional File 3. We found that a cutoff of 0.9 resulted in realistic network degrees with minimal deviation of predicted to observed network degrees and therefore yielded smallest overall mean-squared error (MSE). With this setting, we can predict network hub with the highest accuracy. We further knew, that the estimated number of SL interactions is only a small fraction of the total pool of single-null mutants, so it was reasonable to apply a stringent cutoff. For the prediction of hubs from the list of novel query genes, we classified any gene pair with a conditional probability of 0.9 or more as SL and any value below this cutoff was predicted as a non-SL interaction. This resulted in a list of predicted genetic network hubs as shown in Additional File 4.

2.3 Integrated result

In order obtain a robust prediction of genetic network hubs, we integrated the results of regression for network degree and classification of SL interaction. We combined the results of the two methods by applying a maximum-rank procedure as described in Methods (Section 5.2). This reduces possible biases inherent to one method alone and guarantees a robust ranking. A scheme of the complete modeling procedure is outlined in Figure 7. Ranking of the top 10, top 20 and top 30 genes is highly significant compared to the ranking of the genes that would be expected by chance (p-values$< 10^{-5}$ for all three cutoffs top10, top20 and top30). The fist 30 positions of the final list are shown in Table 1 together with the predicted degree and the ranks in both lists. The complete list can be found in Additional File 5. Inspection of the final list shows that genes with very important functions in the cell rank very highly. It features genes with known roles in mitotic exit (CDH1, LTE1), transcription and mRNA metabolism (PAT1, LSM1), endocytosis (MON2) and protein sorting (PMR1). We expect that these genes will have a high number of SL interactions.

3 Discussion

We pursued a twofold modeling strategy, using regression and classification, to predict new query genes for SL screening that are potential genetic network hubs. Focussing on these query genes in future screens will increase the information gain of the genetic network, because informative experiments with an above-average number of novel interactions will be performed. We implemented both approaches with Random Forest models. While the regression approach directly yielded the desired output, we derived the network degree of a gene in the classification approach after predicting binary SL gene pairs and counting them. This was done using a stringent cutoff criterion. Both approaches rely on suitable genomic and proteomic input features: statistical selection of the best features was done by optimizing the predictive capacity of the method.

The best input features were the degrees from genetic and mixed networks in the regression approach and clustering coefficients in these networks in the classification approach. The fact that the mixed and the between-scores ranked high in the model shows that combining protein interaction and genetic interaction information is very useful for assessing and predicting SL interactions. The power of the between-features gives further support to the "between-pathway model" for synthetic lethality proposed by Kelley et al. (6), whereas the fact that the within-scores were unimportant in our approach questions the relevance of the "within-pathway model". Contrary to the approach followed by Wong et al. (7) who found that two-way genetic interactions (2hop SSL-SSL) with a third gene are most predictive for SL, we found that a combination of protein and genetic interactions with a topology consistent with the between-pathway model are better predictors than two-way genetic interactions (e. g. *hg.between.coeff* vs. *hg.gene.coeff*). Since the most informative features are derived from genetic network information, the method tends to identify suitable query genes for which a large number of genetic interactions are already known from other experiments. So densely connected genes in pathways that have not been targeted by SL screening up to now, might not be detected. They will only be found by the model if they have a high score with other important features not derived from the genetic network. Nevertheless, the genetic interaction information is very valuable in comparison to the remaining genomic data types, as can be seen by the importance in the model (Figure 3 and 6). Thus, it is a valid approach to construct a model based on these features, in particular, because we omit all known htp SL information in the input variables when predicting the interactions to be found in a screen with a given query gene. Furthermore, new screens based on the results of this method will reveal new interactions also for genes currently not present in the list. This will fill in the gaps which are currently present in the genetic interaction landscape. Feeding this information back into the model will increase the predictive power and the genome coverage of the method.

The approach is an extension of Wong et al. (7) who predicted synthetic sick or lethal gene pairs using binary decision trees. Here, we use Random Forest models. This has a few advantages over the method by Wong et al. Due to randomization and ensemble averaging, Random Forests yield better results than decision trees (8). In addition, our combined approach using regression and classification is more robust for prediction of hubs. Most importantly, we do not confine ourselves to the prediction of SL pairs but we use the model to designate optimal query genes for future screens. For the ongoing effort to complete the genetic network using htp SL screens, the predictions can be used to direct and plan future SL screens. The resulting list of novel query genes contains genes with roles in very important cellular processes that are highly recommended for htp SL screening.

4 Conclusions

Correlations between different systematic data sets obtained from studies of the yeast interactome and genome can be exploited to predict genetic interactions. Using two independent models guarantees a robust and unbiased outcome and very likely increases the biological relevance of the result. Assmbling and compiling the predictions helps to prioritize candidate genes for further comprehensive synthetic lethality screening experiments. The presented modeling approach is a substantial tool to efficiently complete the SL network providing high information gain and minimizing experimental cost. Moreover, it is not restricted to genetic networks in yeast but is applicable for hub prediction in any network and any organism where sufficient input data is available.

5 Methods

5.1 Leave-one-query-out cross-validation

For assessing the model performance we apply a special cross-validation strategy. We leave out the information resulting from an experiment with a given query gene, fit the model to the remaining data, predict the output for the left out observations and assess the deviation from the predicted to the real observed network degree. This procedure is repeated to leave out every query gene once and the deviations (errors) are averaged over the repetitions. Leaving out information affects all observations belonging to a given query gene. For the regression analysis, we delete one observation per query gene; in the classification model, all gene pairs with the given query gene are omitted. Furthermore all interactions resulting from the screen with a given query gene are omitted in the genetic networks used for feature generation on the input side. This mimics the conditions for a novel screen, where this kind of information would be unknown, as well. Thus, in both regression and classification models, the values for the input features derived from genetic network information are different for every given left-out query gene and are calculated on a reduced genetic network. It would be too optimistic to use the complete genetic network information because then input and output of the model would contain strongly redundant information. However, we do not exclude interactions with the given query gene that are known from independent screens with other query genes. The variables *SSL.htp.deg*, *SSL.ltp.deg*, *SSL.deg*, *Mixed.htp.deg*, *Mixed.ltp.deg* and *Mixed.deg* in the regression approach and the features *j.gene.coeff*, *mm.gene.coeff*, *g.gene.coeff*, *hg.gene.coeff*, *j.mixed.coeff*, *mm.mixed.coeff*, *g.mixed.coeff*, *hg.mixed.coeff*, *j.within.coeff*, *mm.within.coeff*, *g.within.coeff*, *hg.within.coeff*, *j.between.coeff*, *mm.between.coeff*, *g.between.coeff*, *hg.between.coeff* in the classification approach are calculated anew for each query gene observation in the leave-one-query-out cross-validation.

5.2 Max-rank scoring scheme

For integration of both gene lists at the final stage, we apply a maximum-rank scoring procedure. We combine the two lists by using the maximal rank

$$rank(i) = max(rank(i, regr), rank(i, class))$$

This means, that we sort the list of genes according to their maximum rank in both lists. This ensures that the top of the list only contains genes with high positions in both rankings thereby reducing biases coming from one single method.

6 Authors contributions

DS carried out the data analysis. CD, PB helped with statistical advice and discussions. WG and PB supervised the work.

7 Acknowledgements

Daniel Schöner and Corinne Dahinden have been supported by the Competence Center for Systems Physiology and Metabolic Diseases. All authors acknowledge the ETH Zurich for financial support. All authors have no competing interests.

References

1. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CWV, Bussey H, et al.: **Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants**. *Science* 2001, **294**(5550):2364–2368.

2. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, et al.: **Global mapping of the yeast genetic interaction network**. *Science* 2004, **303**(5659):808–13.

3. Pan X, Ye P, Yuan D, Wang X, Bader J, Boeke J: **A DNA integrity network in the yeast saccharomyces cerevisiae**. *Cell* 2006, **124**(5):1069–81.

4. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**(6833):41–42.

5. Ozier O, Amin N, Ideker T: **Global architecture of genetic interactions on the protein network.** *Nat Biotechnol* 2003, **21**(5):490–491.

6. Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks**. *Nature Biotechnology* 2005, **102**:561–66.

7. Wong SL, Zhang LV, Tong AHY, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey, et al.: **Combining biological networks to predict genetic interactions**. *Proc Natl Acad Sci USA* 2004, **101**(44):15682–15687.

8. Breiman L: **Random Forests**. *Machine Learning* 2001, **V45**:5–32, [http://dx.doi.org/10.1023/A:1010933404324].

9. Liaw A, Wiener M: **Classification and Regression by randomForest**. *R News* 2002, **2**(3):18–22, [http://CRAN.R-project.org/doc/Rnews/].

10. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucl. Acids Res.* 2006, **34**(suppl 1):D535–539.

11. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett H, Coffey E, Dai H, He YD, et al.: **Functional discovery via a compendium of expression profiles**. *Cell* 2000, **23**(5):109–26.

12. Mnaimneh S, Davierwala AP, Haynes J, Moffat J, Peng WT, Zhang W, Yang X, Pootoolal J, Chua G, Lopez A, Trochesset M, Morse D, Krogan NJ, Hiley SL, Li Z, Morris Q, Grigull J, Mitsakakis N, Roberts CJ, Greenblatt JF, Boone C, Kaiser CA, Andrews BJ, Hughes TR: **Exploration of essential gene functions via titratable promoter alleles.** *Cell* 2004, **118**:31–44.

13. Consortium GO: **The Gene Ontology (GO) database and informatics resource**. *Nucl. Acids Res.* 2004, **32**, [http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D258].

14. Beyer A, Bandyopadhyay S, Ideker T: **Integrating physical and genetic maps: from genomes to interaction networks.** *Nat Rev Genet* 2007, **8**(9):699–710.

15. Sharp PM, Li WH: **The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15**(3):1281–1295.

16. Batada NN, Hurst LD: **Evolution of chromosome organization driven by selection for reduced gene expression noise.** *Nat Genet* 2007, **39**(8):945–949.

17. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci U S A* 2003, **100**(8):4372–4376.

18. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS: **Gene function prediction from congruent synthetic lethal interactions in yeast**. *Mol Syst Biol* 2005, **1**:E1–E12.

19. Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19**(10):1275–1283.

8 Figures

8.1 Figure 1

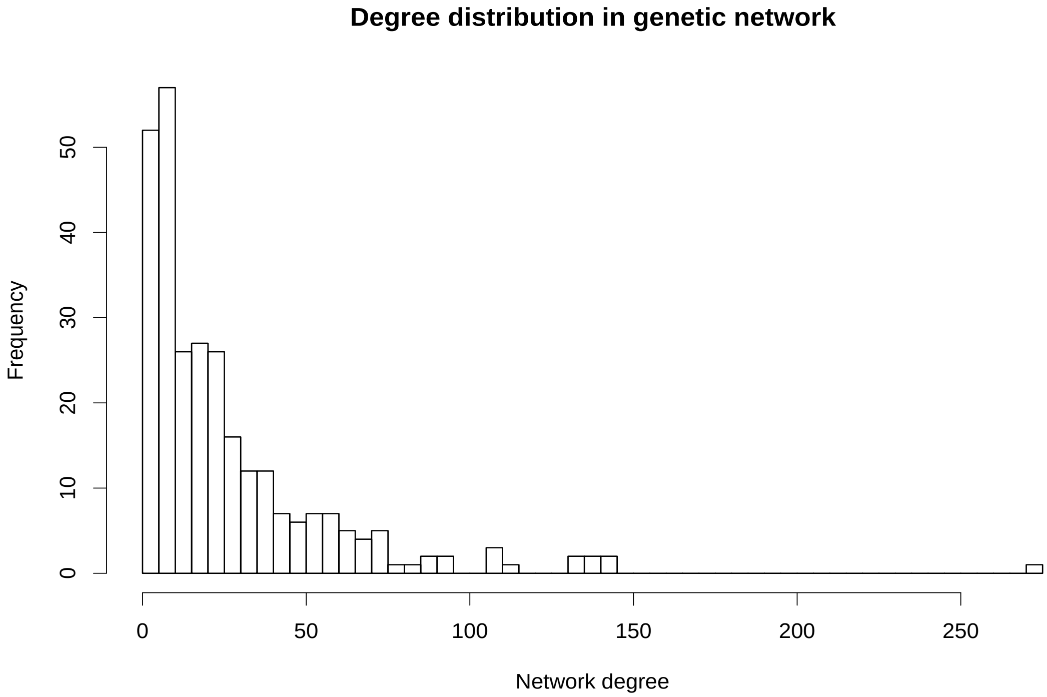

Figure 1: The degrees of the query genes in the genetic network follow a power-law distribution. Many query genes have a low number of connections and a few ones are highly connected (degree $>$ 100, 4800 possible targets).

8.2 Figure 2

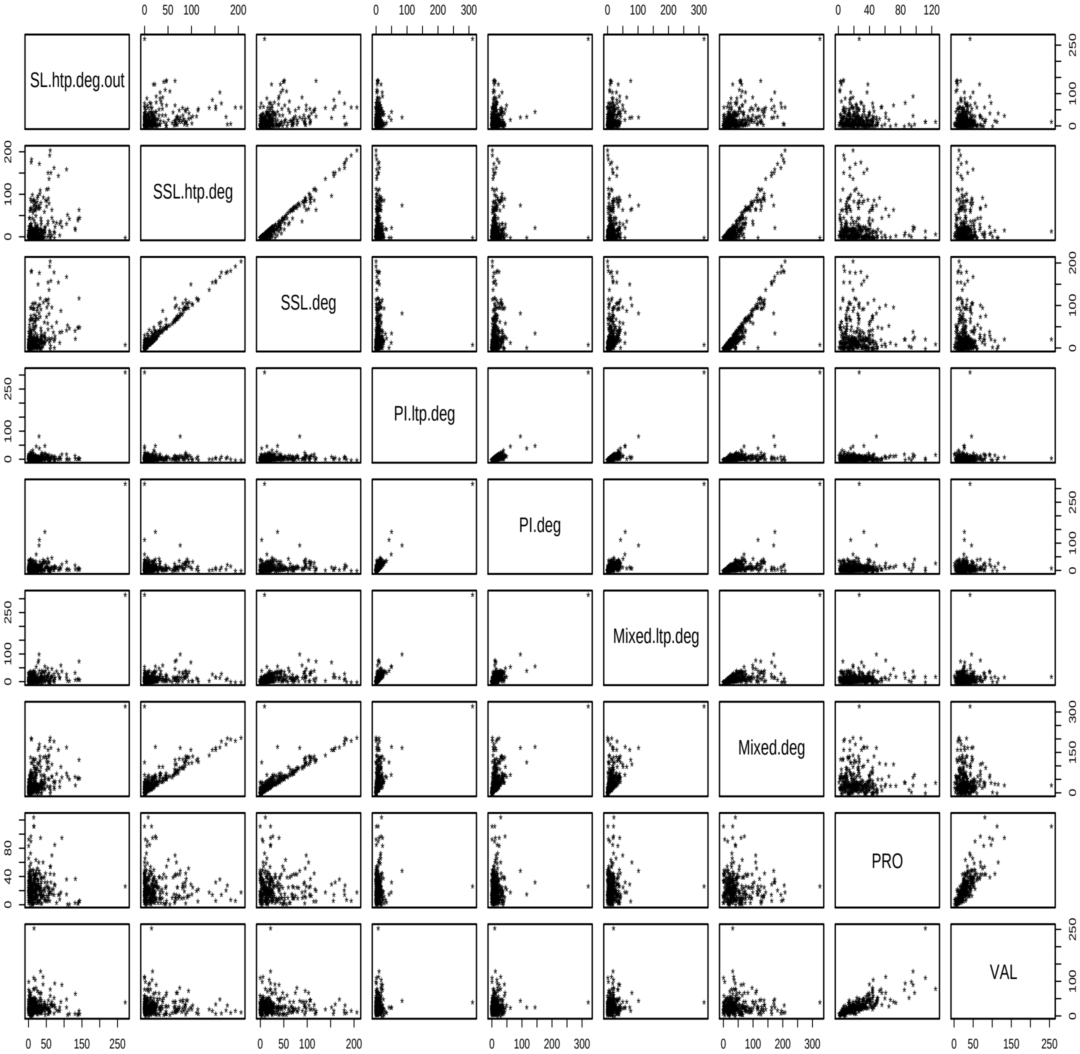

Figure 2: The selected input features show correlations with the response variable *SL.htp.deg.out*. Especially the inputs *SSL.deg*, *SSL.htp.deg* and *Mixed.deg* exhibit marginal correlations to the response variable.

8.3 Figure 3

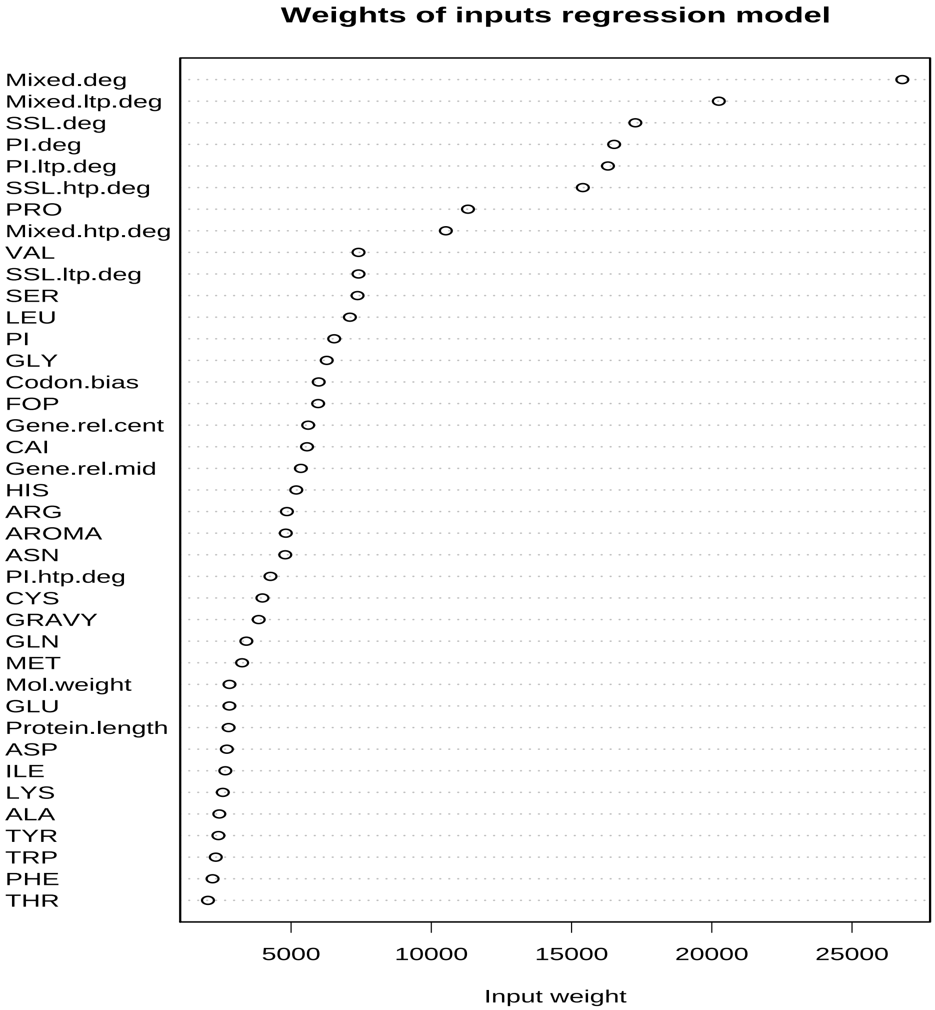

Figure 3: The input features derived from a mixed genetic and protein interaction network show the highest weights in the model, followed by separate genetic and protein networks. Also important are some amino acids, such as proline and valine. The sequence-based inputs, the chromosomal location and most of the remaining amino acid inputs play a minor role.

8.4 Figure 4

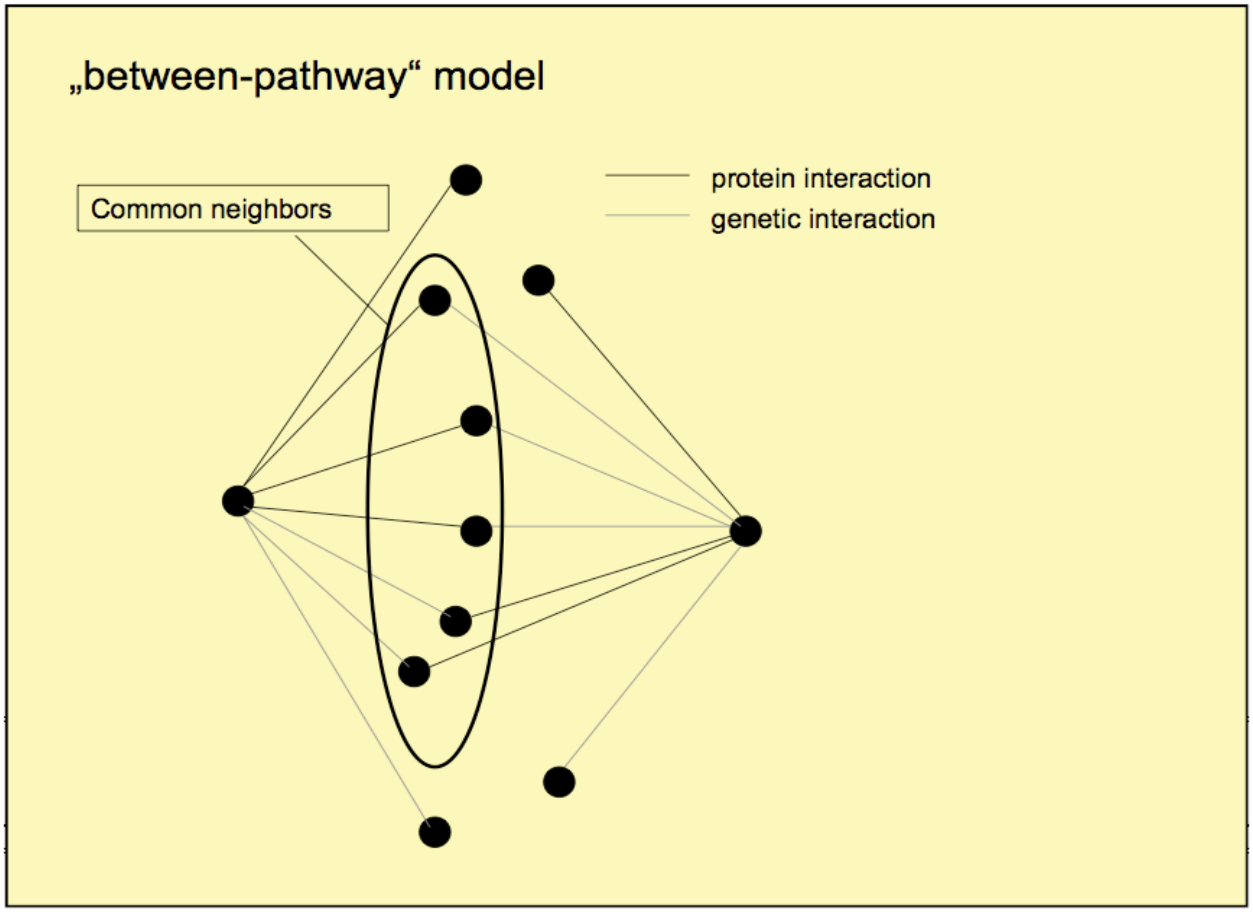

Figure 4: The "between-pathway" model for synthetic lethality by Kelley et al. (6) illustrates that a part of all synthetic lethality interactions occur between groups of proteins densely connected by physical protein interactions. We derive different "between-scores" from the amount of common neighbors of two genes in mixed networks that exhibit exactly one protein and one genetic interaction to each gene.

8.5 Figure 5

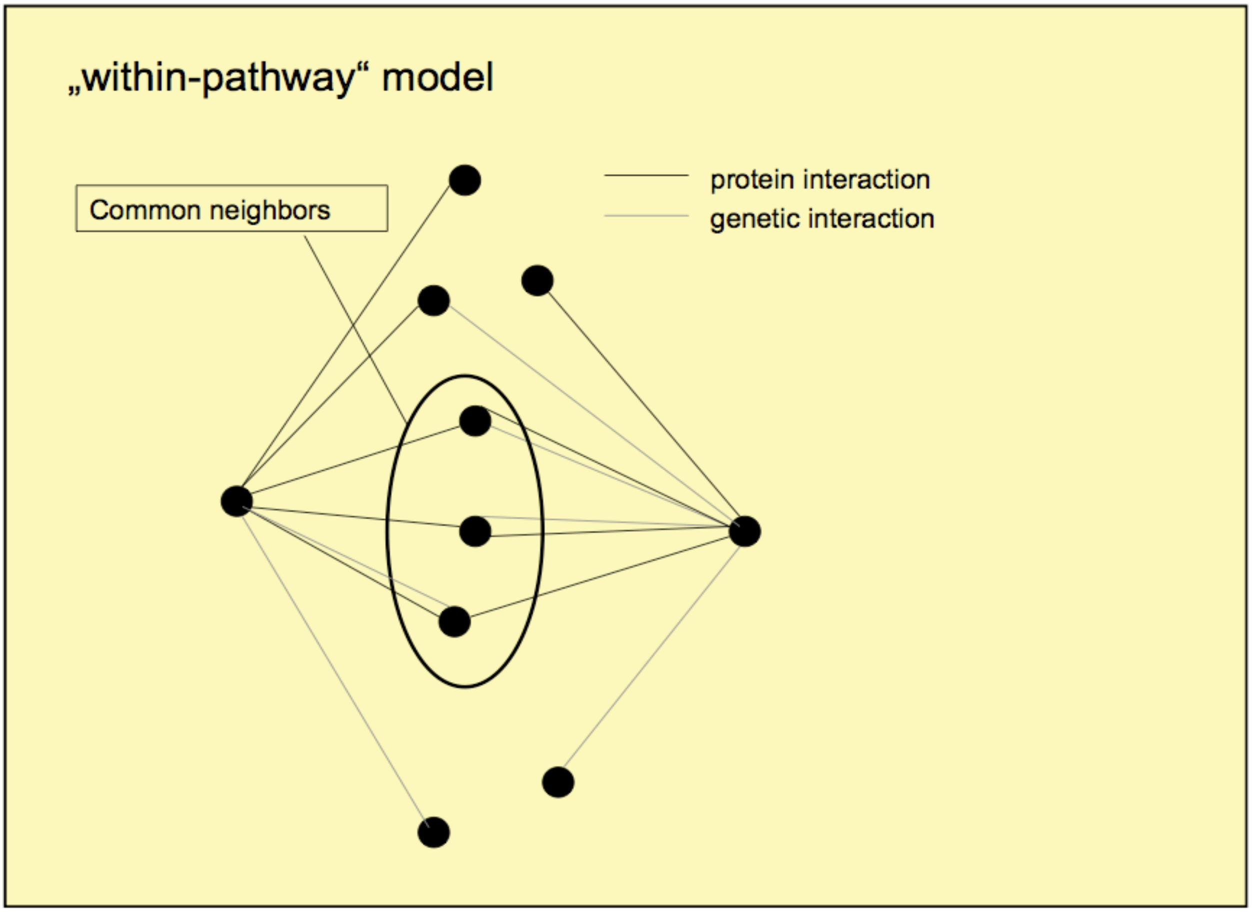

Figure 5: The “within-pathway“ model for synthetic lethality by Kelley et al. (6) illustrates that a part of all synthetic lethality interactions occur within groups of proteins densely connected by physical protein interactions. We derive different “within-scores” from the amount of common neighbors of two genes in mixed networks that exhibit a protein interactions to one gene and both a protein and a genetic interaction to the other gene, so that the synthetic lethality link clearly occurs within and not between the pathway defined by all connected genes.

8.6 Figure 6

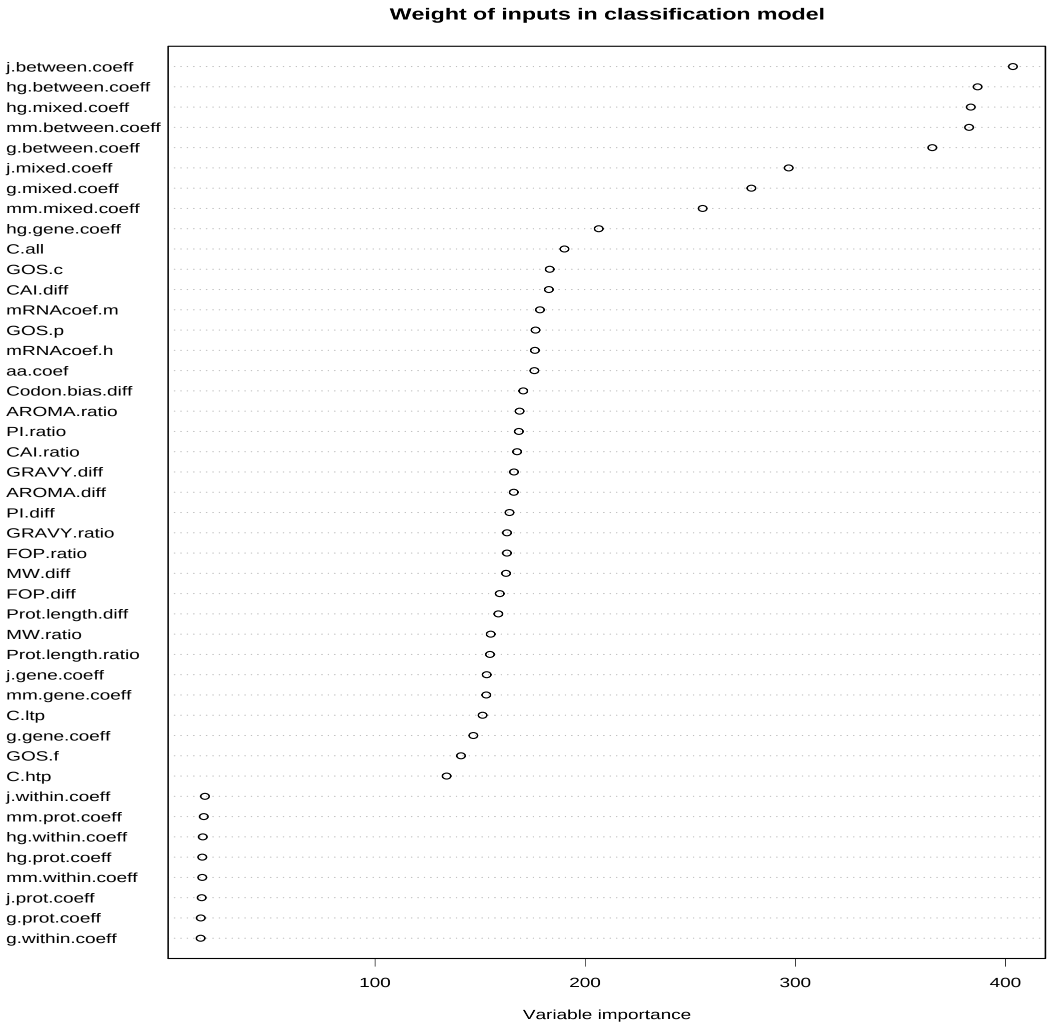

Figure 6: The inputs derived from the mixed network containing genetic and protein interactions clearly show the highest weights, while the features ranging from genetic network clustering, the combined physical connectivity, gene expression correlations to the sequence-based features have lower weights in the model. The features derived from the protein network and the within features have very low importance.

8.7 Figure 7

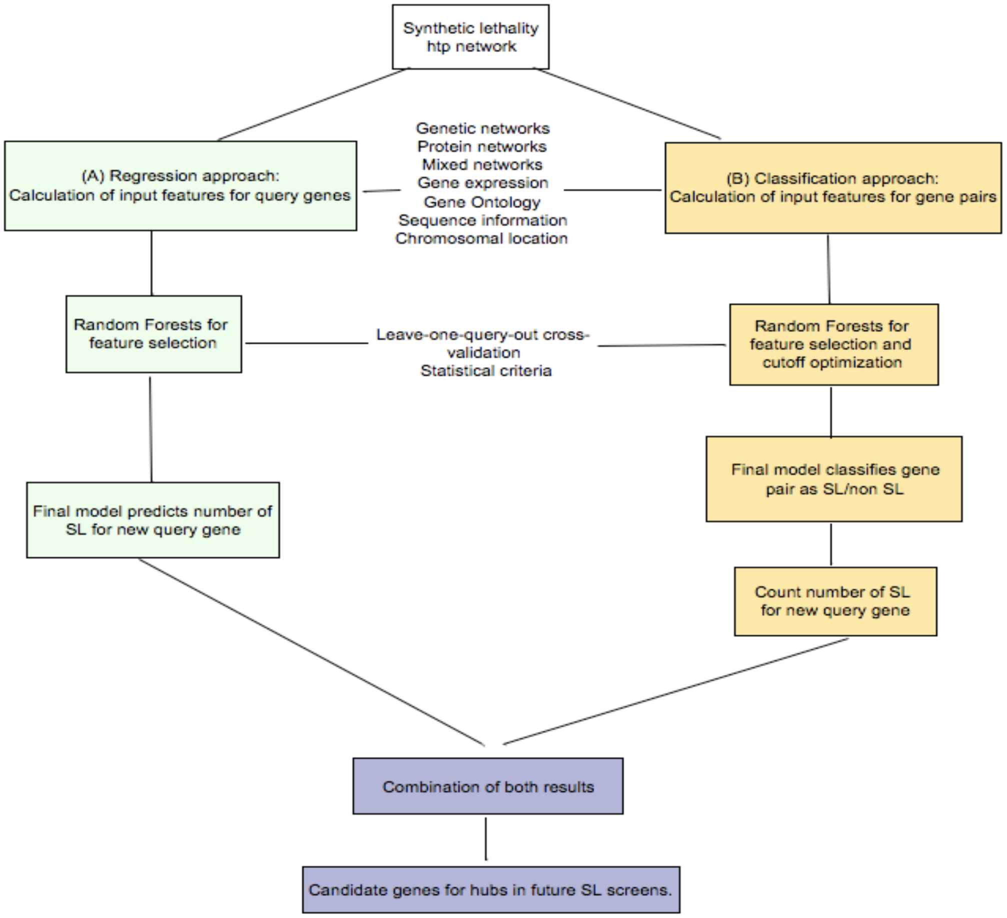
Synthetic lethality htp network
(A) Regression approach: Calculation of input features for query genes
Genetic networks
Protein networks
Mixed networks
Gene expression
Gene Ontology
Sequence information
Chromosomal location
(B) Classification approach: Calculation of input features for gene pairs
Random Forests for feature selection
Leave-one-query-out cross-validation
Statistical criteria
Random Forests for feature selection and cutoff optimization
Final model classifies gene pair as SL/non SL
Final model predicts number of SL for new query gene
Count number of SL for new query gene
Combination of both results
Candidate genes for hubs in future SL screens.

Figure 7: Genes with a high potential of being genetic network hubs are predicted in a parallel approach using two different models based on inputs derived from genomic and proteomic data. A regression model trained on single-gene features directly predicts the number of synthetic lethality interactions partners for a of considered possible future query genes. A classification approach trained on gene-pair characteristics predicts the probability of finding a synthetic lethality interaction between the query genes and a set of target genes. Setting a cutoff on this probability results in a number predicted synthetic lethal connections. The results of both approaches are integrated and ranked to give a robust estimate of future genetic network hubs.

9 Tables

9.1 Table 1

Table 1: Final list of new genetic network hubs obtained from integrating the results of both approaches. The first 30 positions are shown which are highly significant (see text above). Their position in the final list is a result of the ranks in both single lists. The table also shows the degree of both single approaches, which translates to the number of expected SL hits in an experiment.

Rank	ORF name	Gene name	rank regression	degree regression	rank classification	degree classification
1	YGL003C	*cdh1*	8	35.93	10	77
2	YAL024C	*lte1*	13	34.51	14	66
3	YCR077C	*pat1*	15	33.53	20	54
4	YNL297C	*mon1*	5	37.64	22	51
5	YJL124C	*lsm1*	1	52.75	25	48
6	YGL167C	*pmr1*	2	40.91	28	46
7	YDL059C	*rad59*	34	29.57	21	54
8	YNL054W	*vac7*	4	37.74	38	38
9	YER087W	YER087W	26	31.02	42	37
10	YBR088C	*pol30*	46	28.69	5	91
11	YCL008C	*stp22*	35	29.56	48	35
12	YNL107W	*yaf9*	51	27.91	2	98
13	YPR120C	*clb5*	52	27.87	16	64
14	YMR231W	*pep5*	6	37.08	52	34
15	YJL095W	*bck1*	59	27.13	64	31
16	YLR240W	*vps34*	66	26.74	30	46
17	YPR119W	*clb2*	11	35.23	69	31
18	YER155C	*bem2*	14	33.97	74	29
19	YLR182W	*swi6*	82	25.43	9	79
20	YGL173C	*kem1*	38	29.33	84	26
21	YGR092W	*dbf2*	85	25.4	18	61
22	YDR378C	*lsm6*	90	25.2	59	33
23	YNL147W	*lsm7*	91	25.1	41	37
24	YMR094W	*ctf13*	41	29.25	96	23
25	YDR432W	*npl3*	3	38.67	97	23
26	YER019W	*isc1*	65	26.77	105	22
27	YBR097W	*vps15*	63	26.8	106	22
28	YMR198W	*cik1*	109	24.27	11	75
29	YBL007C	*sla1*	89	25.22	109	21
30	YJL090C	*dpb11*	110	24.2	107	21

Additional Files

9.2 Additional File 1

Additional File 1 shows the results of the loqo-CV procedure for the regression model. It lists the number of hits determined by SL screening, the predicted number of hits, the squared difference, the random forest out-of-bag estimate and the overall MSE at the bottom of the data matrix.

9.3 Additional File 2

Additional File 2 lists the predicted number of SL interactions for future gene candidates obtained by the regression model. The gene name, the ORF name and the predicted number of hits are listed.

9.4 Additional File 3

Additional File 3 shows the results of the loqo-CV procedure for the classification model. It lists the experimentally found number of hits and the predicted number of hits for the cutoffs 0.5, 0.7 and 0.9 together with the corresponding squared errors. At the bottom of the data matrix the MSEs for all 3 cutoffs are shown.

9.5 Additional File 4

Additional File 4 lists the number of SL interactions for future gene candidates obtained by the classification model. The gene name, the ORF name and the predicted number of hits are listed.

9.6 Additional File 5

Additional File 5 shows the integrated list of results of both approaches ranked by maximum rank. This is the final list of recommended future SL screening candidate genes.