NSF ABI INNOVATION.
A GRAPH BASED APPROACH FOR THE GENOME WIDE PREDICTION OF CONDITIONALLY ESSENTIAL GENES

1. SPECIFIC AIMS

In recent years, numerous large-scale genomics projects combined with fast sequencing techniques have generated enormous amounts of data. This has led to the identification of thousands of previously unseen genes. However, understanding the function and cellular role of genes, as well as their impact on certain phenotypes remains a challenge. Apart from the Mendelian single gene traits, a substantial portion of the phenotypes we observe in nature result from a complex interplay between numerous genes in addition to various environmental factors.

Of particular interest in the gene-phenotype landscape are the essential genes. These genes are necessary for organisms' growth and survival. The study of essential genes can shed light on the universal principle of life \cite{22951051}. Moreover, they are key players in the field of synthetic biology and can be regarded as potential targets for antimicrobial and vaccine design \cite{24209780}.

From an experimental point, the last decade has seen the development of a number of approaches for determining gene-essentiality \cite{15313213} including systematic knockouts \cite{12140549}, genetic foot-printing \cite{10591650}, and RNA interference \cite{12529635}. However their application on a genome wide scale is expensive, time demanding, and labour-intensive. Moreover, designing new functional characterization assays for genomic targets that have not been previously described is a difficult process.

A fundamental goal in computational biology is therefore to characterize which sets of genes are essential for the organisms survival in a given set of conditions.

This project aims at shedding some light on this question by computationally predicting and experimentally verifying which genes are conditionally essential under a variety of different treatments.

Our computational predictions will be based on a data driven analysis that will integrate information on two levels: phenotypical and molecular. At the *phenotypical level*, we are going to develop a machine learning method to assign phenotypic attributes to genes and identify sets of genes that share the similar phenotypic characteristics. We will then integrate these preliminary predictions with *molecular level* information, and particularly we will exploit transcriptomics data to identify the mechanisms that govern the gene activity. This will allow us to increase the reliability of our preliminary phenotypical assignments and in particular to select from the groups of genes with similar phenotypes the conditionally essential genes. Finally, we are going to experimentally validate our predictions in two yeast systems, and feedback the results into the prediction workflow in order to improve the computational models.

This project will deliver a mathematical framework that will allow scientists, to predict genes which are conditionally essential under a given condition and also suggest a hypothesis for their phenotype generation, regulation, and activity. This framework will be also implemented in a software package that will be made available to the scientific community.

This project will be developed as a collaboration between the groups of Dr Mark Gerstein at Yale University, Dr Haiyuan Yu at Cornell University, and Dr Alberto Paccanaro at Royal Holloway University of London. The three group leaders have a decade-long history of successful collaborations. Together, they have developed methods for predicting networks

from heterogeneous biological datasets, for predicting protein function, and for calculating semantic similarity between genes, among others.

This project can be subdivided into four main aims:

AIM 1: We will develop a machine-learning method to infer phenotype from network neighbourhoods. For this, we will develop semi-supervised machine-learning techniques that make use of both labelled and unlabelled data for training. In particular, we will develop specialised diffusion-based algorithms that will exploit the structure of graph models representing phenotypical associations between genes. Here we make the assumption that phenotypic attributes associated with characterized-entities can be extended to other uncharacterized entities depending on their level of "connectedness" in the graph model. Diffusion-based algorithms will thus allow us to exploit on a genome-wide scale and in an organized fashion the "guilt by association" principle, to predict the phenotypes of uncharacterized genes.

AIM 2: At high-level, genes displaying similar phenotypes are expected to exhibit similar expression and regulatory trajectories. On this premise, **we will develop a computational model that, given a set of genes (*predicted from AIM 1*), will evaluate their expression dynamics patterns, uncovering the regulatory effects that govern them.** By refining the input genes sets into groups of genes with similar dynamic patterns we will be able to pin point conditionally essential genes. This analysis will be done by decomposing the gene expression into *internal* contributions (e.g. AIM1 predictions sets), and *external* contributions (all other genes). Specifically, we will use a state space model to represent the temporal gene expression dynamics and identify principal temporal dynamic patterns. Next, we will use dimensionality reduction approaches to identify canonical temporal expression trajectories unravelling the regulatory effects from various contributors. Moreover, we will implement a scoring metric that will evaluate to what extent the expression dynamics of a group of genes are driven by their internal regulatory network. Finally, we will cluster the genes with similar regulatory dynamics and similar conditional essential phenotype.

AIM 3: We will experimentally validate our predictions in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe.* While the first two aims are strictly computational, we are going to validate our top predictions in the two model organisms under three stress conditions: oxidative stress, osmotic stress, and DNA damage stress.

AIM 4: We will integrate the computational models into a state-of-the art software suite. All theoretical models, datasets, and analysis results will be deployed online and will be made available to the larger scientific community through a web portal. The successful completion of this project will provide proof of concept methods for identifying conditionally expressed genes.

2. BACKGROUND AND PRELIMINARY RESULTS

2.1 Background on Phenotype Prediction

The concept of phenotype is defined as the set of organism observable traits such as its biochemical, physiological, behavioural properties, etc. Identifying the genes and understanding their contribution to a certain phenotype is an on-going quest for many researchers in the field of genomics. In order to address this challenge, phenotypes have been collected and systematically organised in formal, organisms-specific, phenotype ontologies. An ontology provides a conceptualization of a knowledge domain that is both human and computer comprehensible \cite{14681407}. It uses a hierarchical structure to represent relationship between concepts using a controlled vocabulary \cite{14681407,21261995}. There are a number of publicly available phenotype ontologies for human \cite{24217912}, worm \cite{21261995}, fly \cite{24138933}, mammals \cite{20052305}, yeast \cite{23658422}, etc.

Comparative genomics has been proposed for uncovering gene-trait relationships \cite{9790834,9598967}. This approach begins by constructing phenotypic profiles, which indicate which organism exhibits a particular phenotype – this is similar to the concept of phylogenetic profiles \cite{10200254}. Then causal relationships between genes and traits can be deduced from the co-occurrence of genes and phenotypes across a large number of genomes. The underlying principle is that orthologous genes involved in similar biological processes should determine orthologous phenotypes called "phenologs", across various species. These ideas were applied to predict genes involved in well characterised traits such as hyperthermophily \cite{12683966} and flagellar motility \cite{12546786}. Several approaches have been developed for this comparative analysis. For example, Tamura et al. \cite{18467347} proposed a rule-based data mining algorithm to associate Clusters of Orthologous Groups of proteins (COGs) with phenotypes. Slonim et al. \cite{6732191} proposed an information-theoretic approach to extract preferentially co-inherited clusters of genes having significant association with an observed phenotype.

However, most comparative genomics approaches do not take into account numerous clues regarding the various aspects of gene phenotype that are hidden in a vast array of gene expression, metabolite expression, and protein-protein interaction data. Biological systems are mediated by interactions between thousands of molecules. Network-based statistical models are particularly useful in unlocking the complex organization of biological systems \cite{17473168,11034217,10521342,10935628,12202830,12399590,16730024,18421347,15190252,12134151,17274682,19372386,16311037}. Although network based models have been previously used for the prediction of gene function (e.g. of GO labels), these ideas have not been exploited for the prediction of phenotype on a genome wide scale.

In this project, which builds on earlier joint works and preliminary results in the area of network analysis by Gerstein, Yu, and Paccanaro, we will develop new state-of-the-art methods for gene essentiality prediction. In particular, the methods proposed here will combine, in a unique fashion, the decomposition of temporal gene expression dynamics and diffusion-based algorithms to improve conditionally essential gene prediction.

2.2 Background on Modelling Gene Expression Dynamics Using a State-Space Model

The state-space model has been widely used in engineering \cite{Brogan}. It models the dynamical system output as a function of both the current **internal** system state, and the **external** input signal. A commonly used example in engineering is the vehicle cruise control system where the internal system state is the vehicle's speed. Based on the road conditions (external input signal), the cruise control will output the required fuel amounts in order to keep the desired speed. In biology state-space models have been used in the study of regulatory networks and in particular in the analysis of temporal gene expression data \cite{16403793,17044234,14962938}. Compared to other methods that calculate the expression correlation between individual genes, the state-space model has the advantage that it predicts the temporal-causal relationships at the system level i.e., the state at a time is determined by the state and external input at the prior time point.

One of the early adopters, Wu et al (2004) used state-space equations to model the gene expression from microarray data. The authors describe the gene expression profiles as observation variables whose values are modelled by a linear combination of the current internal state variables (e.g. regulatory elements expression levels). Factor analysis was used to identify the internal state variables and calculate their expression values. The results suggest that it is possible to unambiguously determine a gene expression dynamic pattern from a limited time-course dataset. More recently Mar and Quackenbush (2009) \cite{20041215} have used state-space models to study cell differentiation. They decompose the state-space gene expression trajectories into components one representing the changes inherent to the biological process (cellular phenotypic changes) and another component that captures the cell response to the perturbation (variation in gene expression levels).

However, these models have been able to account for only a limited number of genes. In fact, it is not feasible to use state space models for describing systems composed of thousands of genes due to the limited amount of data available, which would not allow us to learn all the required model parameters.

In this project, for the first time, we will combine state-space models with dimensionality reduction techniques in order to model the gene expression data on a genome-wide scale. Dimensionality reduction techniques will allow us to model gene expression data in terms of the expression of a few meta-genes, thus uncovering the regulatory effects.

2.3 Preliminary Results

2.3.1 Preliminary Results on Phenotype Predictions and Diffusion-Based Methods

Yu, Gerstein, and Paccanaro have developed a correlation-based method \cite{17038185} that was able to discover genotype-phenotype associations combining phenotypic information from a biomedical informatics database, GIDEON, with the molecular information contained in COGs \cite{12969510}.

Paccanaro has developed and applied graph diffusion methods in biology. In biological networks, the relations (represented by the links) are inherently very noisy and therefore algorithms that employ these links directly are prone to errors. The idea behind graph diffusion methods is to improve the accuracy of inferences from these local relations by instead considering the connections globally in a certain neighbourhood, or even in the whole graph – this is somewhat analogous to taking averages to reduce the noise. Importantly, the diffusion of information over graphs offers a natural framework for integrating datasets which are themselves graphs; that is, when diverse sets of data are available, graph diffusion allows us to combine them to obtain sound statistical inferences. In this way, the weak signals contained in each dataset are enhanced through data integration. Paccanaro has developed a new machine learning method for the diffusion of information over large weighted graphs. This method has been applied to the prediction of protein function with great success. In particular, the results obtained by his system in the recent CAFA competition (Critical Assessment of Functional Annotation) placed its performance among the very best systems in the world. The results are appearing in a Genome Biology article which is now in press \cite{arXiv:1601.00891}.

In this project, we will develop and implement analogous diffusion-based approaches to characterize and predict conditionally essential genes under a variety of different treatments.

2.3.2 Preliminary Results on Networks and State Space Models

The Gerstein, Yu, and Paccanaro labs have carried out projects in biological networks for over a decade. We have made extensive contributions in the analysis of genomic data using network frameworks \cite{14564010}. In particular, we have integrated regulatory networks with gene expression to uncover different kinds of dynamic sub-networks \cite{15372033}. We also developed methods to analyse the regulatory networks of a variety of species from yeast to human, using a wide range of data \cite{22125477,20439753,22955619,21177976}. In the following we will give more details on our earlier results in network biology that are most relevant to this project.

Biological networks, normally large in scale, are organized with topological structures in the form of interacting modules. We have previously collaborated in developing various methods to identify the functional modules of biological networks. We developed a method to extract metabolic modules from metagenomic data, enabling the identification of pathways that are expressed under different environmental conditions \cite{19164758}. We have also developed a way to identify nearly complete, fully connected modules (cliques) present in network interactions \cite{16455753}, and we have been using networks to map various kinds of functional genomics data \cite{22955619}. For example, by mapping gene-expression data onto yeast regulatory network, we identified different sub-networks that are active in different conditions \cite{15372033}.

Using biological networks, we have developed a computational approach OrthoClust \cite{25249401}, to extract meaningful new information from gene expression data. OrthoClust is a universal computational framework that integrates co-association networks of individual species using gene orthology relationships to enable the identification of functional modules formed by species-specific or conserved gene. Leveraging on the modENCODE RNA-seq data for *C. elegans* and *D. melanogaster* we used OrthoClust functional module predictions to infer putative functions of uncharacterized elements (e.g. non-coding RNAs) based on the guilt-by-association principle.

More recently, Gerstein has developed a computational approach that uses state space models to analyse the regulatory effects of evolutionary conserved vs divergent transcription factors across *C. elegans* and *D. melanogaster* using data from the modENCODE project (in preparation). Studying the regulation and expression patterns of orthologus and species specific genes we were able to characterize the regulatory systems that govern fundamental embryonic-developmental processes.

In this project, we will build upon our expertise in biological network and in the development of advanced computational approaches using state space models, to design and implement genome-wide workflows for the prediction and characterization of conditionally essential genes.

2.3.3 Preliminary Results on Growth Assays Under Stress Conditions

The Yu Lab established a stress-response interactome for the fission yeast, *S. pombe*, named StressNet \cite{23695164}. In *S. pombe*, Sty1 is activated in response to various stresses, including oxidative and osmotic stress, starvation, and other conditions \cite{17605132, \cite{8824587}. Sty1 has orthologs in *S. cerevisiae* (Hog1, with 89% sequence similarity) and human (p38, with 69% sequence similarity). Both p38 and Sty1 respond to a wide range of stresses and both are different from Hog1 in terms of function \cite{9443913}. With our stress-response interactome, we detected key interactions at every step of the MAPK signal transduction pathway and, therefore, completely recapitulated the entire Sty1 pathway. This confirmed the sensitivity and accuracy of our HT-Y2H method, especially for discovering transient interactions in signaling pathways. Among all Sty1 interactions in StressNet, those with its activator (Wis1) and inhibitor (Pyp2) were both conserved between the two yeast species, and the Sty1-Wis1 interaction interface was intact. By contrast, the interaction between Sty1 and its known target in fission yeast, Atf1, represented a rewired interaction (Fig. 1C). We also identified a previously unknown interactor of Sty1: SPBC2D10.09, a protein that we named Snr1 (Sty1-interacting stress-response protein). To confirm this interaction *in vivo,* we performed co-immunoprecipitation of tagged proteins expressed in *S. pombe* (Fig. 1D). The amount of Snr1 pulled down in the presence of Sty1 was greater than that pulled down in the absence of Sty1, indicating that the interaction with Sty1 stabilizes Snr1 (Fig. 1D). The corresponding orthologous pair of Hog1 and Ehd3 in *S.*

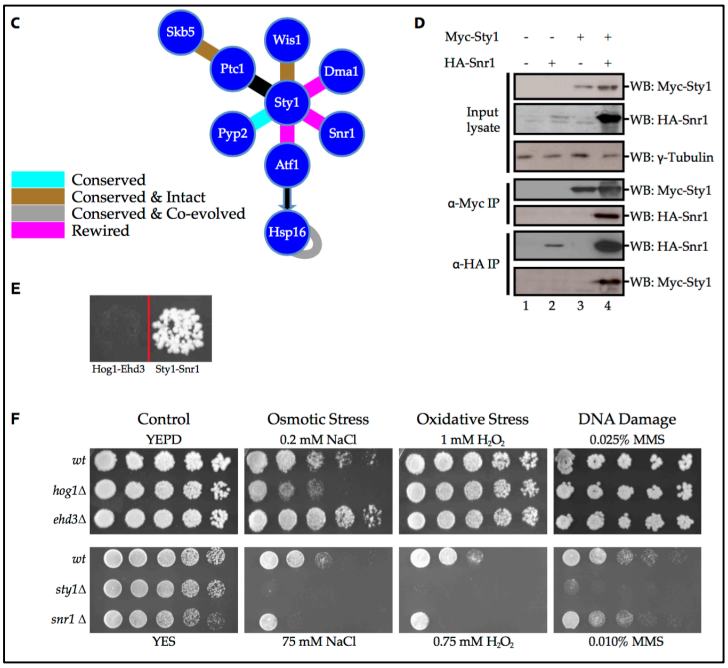

Figure 1 Gene expression under stress conditions in *S. pombe*

cerevisiae did not interact by Y2H (Fig. 1E). Cells lacking *snr1* (*snr1*Δ cells) grew slower under stress, similar to *sty*1Δ cells (Fig. 1F), whereas growth of *ehd3*Δ cells was not compromised. These results suggested that Snr1 is a component of the Sty1 pathway and that its functions diverged from its budding yeast counterpart. Moreover, *snr1* also has a human ortholog, HIBCH, further investigation of which may expand our knowledge of the human p38 MAPK pathway.

In this project, we will build upon our expertise in large-scale gene deletions and growth assays under stress conditions to validate a large number of predictions made in AIMs 1 and 2.

3. RESEARCH PLAN AND METHODS

3.1 AIM 1: Inferring Phenotypes Through Diffusion on Biological Networks

The vast array of available data brings a fresh perspective in the area of gene phenotype prediction. By integrating various datasets we believe that we can make statistically significant large-scale phenotypical inferences.

The data available for phenotype prediction can be divided into two categories. While some types of data translate directly into a probability of a given phenotype, other types of data instead describe a "relatedness" in the phenotypes associated to two genes in the same genome. For example, detecting a phenolog (i.e. an orthologous phenotype between organisms) amounts to assigning a probability P that a certain gene is involved in phenotype F. On the other hand, finding a certain correlation between the profiles of the expression of genes X and Y amounts to assigning a certain probability Q that the two genes have related phenotypes. We will refer to these two types of data as "unary relations" and "binary relations" respectively (Table 1).

Table 1. Phenotype prediction input data.

Data Type	Example
UNARY RELATIONS	Experimental evidence Phenolog
BINARY RELATIONS	Gene expression Protein expression Protein-protein interaction Genetic interaction Pathway information

Binary relations have a natural representation as graphs. Recently, there has been a lot of interest in the machine learning community in methods for making inferences on graphs. We propose to leverage on these ideas and develop theoretical graph-based methods for large-scale phenotypical inference. The approach makes use of the phenotypical labels associated with some genes to infer phenotypes of uncharacterized ones (semi-supervised learning).

In a typical situation, for a given genome there will be genes that have already been associated with a given phenotype, and genes whose associated phenotype is still unknown. We begin by constructing graphs, in which the nodes represent the genes and each edge represents a (binary) relation between the two connected genes, i.e. co-expression. Each edge is labelled with a value that quantifies the relation it represents (i.e. their level of co-expression); similarly each node is labelled with its known phenotypical assignment or "NA" otherwise.

The two different types of relations described above will be treated differently for inference: binary relations will allow the characterization of the unknown genes by *diffusing* the information of the labelled nodes over the graph, through the links; while unary relations will be thought of as representing a "tendency" (or a prior probability) of a gene to be associated with a given phenotype.

Here we provide an intuition for how the diffusion process works. Let us imagine the graph as having a physical implementation as a network of water wheels connected by underground pipes in which water flows: for each node (gene) we have a wheel, and for each edge (binary relation) we have a pipe connecting the corresponding wheels. The pipes have different sizes

according to the edge label, thus allowing different amounts of water to flow through them, depending on the strength of the relation. Each different phenotypical assignment of genes in the dataset is represented by a salt (dye) of a specific colour. When a salt is dropped in a wheel, it colours the water in it, and we will assume that waters of different colour don't mix. The diffusion process consists in dropping the coloured salt of each known gene in its corresponding wheel, and then letting the coloured water be transported by the pipes. No salt is dropped in the wheels corresponding to the uncharacterized gene. However, the water in these wheels will also eventually become coloured due to the coloured waters coming from the pipes. After the coloured waters have been allowed to circulate in the pipes for some time, the amounts of different coloured waters arriving at such unlabelled wheels will provide the basis for a probabilistic distribution of assignments over the phenotypical classes for the corresponding uncharacterized genes. It is important to notice that the whole process can naturally take into account genes having multiple phenotypes, as salts of different colours can be poured into the same wheel.

From this analogy we can see that the diffusion of information over graphs offers a natural framework for integrating datasets which are themselves graphs. This process produces evidence for phenotypical assignments that can be further integrated with the evidence coming from the unary relations using a statistical method, such as the Bayesian model. *The strength of the methodology proposed here lies in its ability to use diverse sets of noisy data, and to combine them to obtain sound statistical inferences of gene phenotypes; the weak signals contained in each dataset are enhanced by integrating the data.*

3.1.1 Algorithm Development

The phenotype inference method will contain several parameters that will be learned from the data. Here we assume that, for a given genome, this will be done by applying various machine learning techniques (as described below) to subsets of genes for which the phenotypic assignment is known (training sets). The method development will have to solve two main issues: (i) how to integrate information coming from different experimental sources; and (ii) how to properly diffuse the information over the graphs. The study of solutions for these two problems will constitute most of the algorithmic research of AIM 1. In the remainder of this section we will analyse each one in turn, proposing some possible ideas for their solution.

(i) Integration of Information from Different Experimental Sources. As anticipated earlier, a possible method for integrating the various types of information is using a statistical Bayesian model. Using the Naïve Bayes assumption, we can rewrite the likelihood of the combined vector of evidences given the phenotype as a product of each evidence given the phenotype. That is, the posterior probability distribution of the phenotypic assignment given the evidence, $P(F_i \mid E_1 \ldots E_n)$, is defined as:

$$P(F_i \mid E_1,\ldots,E_n) = \frac{P(E_1,\ldots,E_n \mid F_i)\cdot P(F_i)}{\sum_j P(E_1,\ldots,E_n \mid F_j)\cdot P(F_j)}$$

and can be approximated by:

$$P(F_i \mid E_1,\ldots,E_n) = \frac{P(E_1 \mid F_i)\cdot \ldots \cdot P(E_n \mid F_i)\cdot P(F_i)}{\sum_j P(E_1 \mid F_j)\cdot \ldots \cdot P(E_n \mid F_j)\cdot P(F_j)}$$

where $(E_1 \ldots E_n)$ is the combined vector of n different evidences or features (E_j), and F_i represents the i-th phenotypical assignment. Here, each E_j represents evidence coming either from a unary relation (i.e. a phenolog) or a binary relation (i.e. co-expression). Since unary and binary relations must be treated differently, their likelihood model $P(E_j|F_i)$ will be built in a different way from the training set.

For unary relations, the likelihood models, $P(E_j|F_i)$, can be approximated directly by using maximum likelihood estimates, that is by using the frequencies of the features in the training set (or alternatively using more robust "smoothed" estimates).

In order to estimate a likelihood model for a given binary relation we first need to build a graph, and then we need to run the diffusion process (described in the next sub-section). The graph will have a node for each gene. The values for the edges controlling the diffusion process will be a non-linear mapping of the experimental data that will be learned[1] \cite{16554755} from the training set using, for example, Support Vector Machines. Thus, for each binary relation there would be a different graph and the diffusion process would be carried out separately. The result of each diffusion process, corresponding to the amount of different phenotypic labels, will constitute the feature for that binary relation. The likelihood models for the binary relations will be approximated by the frequencies of these features in the training set. The prior probabilities of phenotypic assignment, $P(F_i)$, will also be approximated by the relative numerosity of the different phenotypic classes in the training set. Thus, having obtained likelihood models for both unary and binary relations and estimates for the priors, we can obtain a phenotypical assignment by computing the numerator of the above equation (notice that the denominator is independent on the phenotypical class).

The Bayesian model outlined here is not the only possible way to integrate the information coming from the different types of data. In this project we will evaluate a number of different machine learning techniques in order to find the optimum method for solving our porblem. In general, data from unary relations can be included directly, while for each binary relation we would go through the additional step of the diffusion process. However, once the diffusion process has generated a feature for a binary relation, then all the features can be collected into a vector and a unique probability distribution of phenotypical assignments can be obtained as a non-linear mapping of this vector. Such non-linear mapping would also be learned from a well-characterized training set.

(ii) Diffusion of Experimental Information for Phenotypical Assignments. Here we describe three methods for diffusing the phenotypic label information over the graphs that we will evaluate in this project:

Method 1. This approach consists of simply diffusing the phenotypical labels by simulating Markov random walks on the graph. Given a graph, we can derive the Markov transition matrix that controls the Markov diffusion process, and use it to diffuse the normalized vectors of known phenotypic assignments over the graph. Using similar approaches, Paccanaro has recently obtained excellent results clustering protein sequences \cite{16547200}.

Method 2. This approach projects the nodes of the graph onto points in a (low dimensional) space in such a way that the distance between any two points is related to how well connected the two nodes are in the original graph. In other words, we project the nodes in such a way that for any two nodes, the higher the number of short paths existing between them in the original graph, the smaller their distance in the projected space (here the length of a path in a graph is defined as the sum of the values that label the edges along the path). Once the genes have been projected into this space, we need to discriminate between the distinct phenotypical classes. This could be done by learning an appropriate discriminative function using some training data; or by learning a separate probabilistic model for the points in each phenotypical category. This type of projection, sometimes called *Diffusion Maps*, has recently been successfully applied to solve problems from Computer Vision: lip-reading and image-sequence alignment \cite{15899970}. We have used these ideas with very good results for predicting protein-protein interactions using the topological properties of networks of interactions observed experimentally \cite{AlbertoPac}.

---

[1] This technique for building the graph is similar to the method that we (Gerstein and Paccanaro) have already successfully applied to obtain a unique protein-protein interaction network from several independent protein-protein interaction datasets obtained using different experimental techniques in yeast [46].

Method 3. Finally, a third approach is to map the problem of phenotypical assignment onto that of learning a particular classification on a Riemannian manifold. This approach has been shown to be very successful in a variety of classification problems, in the context of semi-supervised learning, by Belkin et al \cite{Belkin}. The authors modelled the manifold where the data lies as a weighted graph G. Next, they showed that any function on G can be decomposed as a weighted sum of eigenfunctions of the graph Laplacian L, and they learned such coefficients from the training data. For the problem of phenotypical assignment, data from binary relations are already in the form of graphs, and therefore we need to learn the values for the weights for the eigenfunctions of the graph Laplacian. This can be seen as another way to diffuse information, as the Laplacian matrix is related to the Markov random walk \cite{16547200}.

Details on Method Development and Validation. We will develop and validate our proof-of-concept using publicly available data on *S. cerevisiae* and *S. pombe*. Phenotype ontologies as well as genotype-phenotype associations are available for these organisms \cite{23658422,16982638}. Our algorithms will be trained using training sets composed of known gene-phenotype associations, and their performance will be evaluated by means of test sets (by "cross-validation"). However, the general approach presented in this proposal does not have any species specificity, and for this reason the methods developed here on yeast should also perform well in different organisms. The performance of the algorithms will be evaluated "*in silico*" by cross-validation.

Impact and Innovation. *The methods developed in AIM 1 provide, for the first time, a principled computational approach to functionally annotate the phenotypes of previously uncharacterized genes on a genome-wide scale.*

3.2 AIM 2: Identification of Gene Canonical Expression Patterns Using State-Space Models and Biological Networks

The results of AIM 1 will provide us with sets of genes sharing similar phenotypes, and will inform us with respect to genes sharing the similar conditional essential characteristics. However, questions regarding how phenotypes emerge and are regulated, and which subset of genes are conditionally essential remain unanswered. Thus in AIM 2 we will provide a molecular characterization of phenotype by looking at gene expression and uncovering the regulatory effects that govern it. Using as input the gene clusters from AIM 1, we will develop a novel computational method for decomposing the gene expression into contribution from internal (within the same group) and external (all other genes) regulatory components using state-space models and dimensionality reduction techniques. Finally we will introduce a scoring function that will measure the similarity in the expression dynamics for genes within the same phenotypic cluster allowing us to obtained a refined set of conditionally essential genes.

FILT
REFINE

3.2.1 Development of a State Space Model for Large-Scale Gene Expression Data

A gene regulatory network is composed of a variety of smaller regulatory sub-systems that define each a particular regulatory function \cite{12840046,16738561}. Given a group of genes of interest in a subsystem, their expression levels are controlled by internal interaction within their subsystem and by external interactions with regulatory factors from other sub-systems. Both the internal and external regulatory factors control the gene expression patterns in a dynamic fashion (e.g. the regulatory signal at time t will be affect the gene expression at time $t+1$). Thus a state-space model can be used to formulate the temporal gene expression dynamics for the group of genes of interest as a linear combination between the internal and external interactions.

Let X be the gene group of interest and U a set of external regulators (Fig. 2A). The state-space model of gene expression dynamics is:

$$X_{t+1}=AX_t+BU_t$$

where the vector $X_t \in \mathbb{R}^N$ consists of the expression levels of each the N genes from group X at time t, and the vector $U_t \in \mathbb{R}^M$ contains the expression levels of each of the M regulatory genes in group U at time t. The system matrix $A \in \mathbb{R}^{N\times N}$ captures the internal causal interactions among genes in X (e.g. A_{ij} describes the contribution from the j^{th} gene expression at time t to the i^{th} gene expression at time $t+1$) which instantiates a gene regulatory network. The control matrix $B \in \mathbb{R}^{N\times M}$ captures the external causal regulation from the U genes to the X genes (e.g. B_{ij} describes the contribution from the j^{th} gene expression in U at time t to the i^{th} gene expression in X at time $t+1$). $\mathbb{R}$ represents the real number domain.

Unfortunately, the above equation cannot be applied directly to large-scale gene expression data. In fact, gene expression experiments normally have limited time samples (for example, there may only be a dozen time points), which are far less than the time samples needed to estimate the large matrices A and B, when X and U are composed of hundreds or thousands of genes. Our idea for solving this problem is to project the experimental high dimensional expressions data onto a much lower dimensional space in which the expression of a few meta-genes accounts for most of the variance in the original expression data. We will attempt to achieve this using a dimensionality reduction technique, such as Principal Component Analysis or Locally Linear Embedding \cite{11125150} (Fig. 2B). Having reduced the dimensionality of the problem, we will be able to model the resulting small scale system composed of the few meta-genes using the above space-state equation – the expression data is now sufficient to learn the smaller set of required parameters $\tilde{A}$ and $\tilde{B}$ (Fig. 2C).

The learned model provides us with a way to decompose the contributions of internal ($\tilde{A}$) and external ($\tilde{B}$) meta-gene regulatory factors into canonical dynamic patterns. This can be done, for example, by applying the eigenvalue decomposition to the $\tilde{A}$ and $\tilde{B}$, in which case we would obtain canonical patterns.

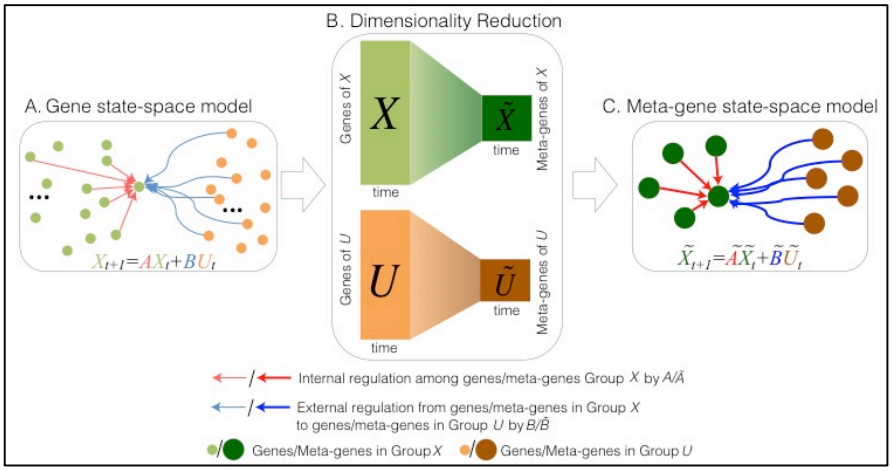

Figure 2 Decomposing the high dimensional experimental gene expression data into contributions from the internal and external regulatory

This will allow us to quantify the contribution of each of the genes in the external or internal set to the expression of the genes in the internal set. In other words, for every gene in the dataset, we will understand what its contribution is, in terms of canonical patterns, to the dynamics of the internal set.

Moreover, we will develop a scoring function to measure the degree to what extent the expression dynamics of a group of genes are driven by their internal regulatory network, e.g. an "internal score". This metric will work under the assumption that the higher internal score is, the more degree of their expression dynamics are driven by internal genes themselves. The mathematical definition of internal score, $\boldsymbol{s_{int}}$ can be proposed as follows: given a group of genes or meta-genes (internal group X), the internal score, $\boldsymbol{s_{int}}$ is the distance of their time-series expression matrix and the product of the internal principal dynamic pattern (***iPDP***)

coefficient matrix (***c(i,j)***) and ***iPDP*** time-series expression matrix, normalized by their time series expression matrix; i.e.,

$$s_{int} = \frac{\|[X_1, X_2, \dots, X_t] - [c(i,j)] \cdot [iPDP]\|_L}{\|X_1, X_2, \dots, X_t\|_L}$$

where ***L*** can be different norms for matrix distance such as Frobenius norm \cite{Frobenius}.

3.2.2 Identification and Refinement of the Conditionally Essential Gene Sets

In order to identify with high confidence the set of conditionally essential genes, we are going to use the phenotypic assignment of AIM 1 to cluster the genes into phenotypic groups. Each group containing genes associated with a conditionally essential label will be used as input for the internal group X and be subjected to the analysis of its internal and external regulatory dynamics patterns, while the remaining genes will form the external group U. Following the expression and regulatory analysis workflow as described in **3.2.1.**, we will be able to obtain a refined subgroup of conditionally essential genes that will be further validated experimentally as described in **AIM 3.**

Details on Method Development and Validation. We plan to validate our state-space methods for gene expression pattern decomposition using publicly available data for two yeast species *S. cerevisiae and S. Pombe* as described in AIM 3. In particular, we will test our methods by analysing the gene expression dynamics patterns during yeast development. As such we will use the conditionally essential genes as the *internal* group *X* and species specific non essential transcription factors as the *external* group U.

Impact and Innovation *The outcome of AIM 2 will provide a state-of-the-art approach to characterize temporal expression data and differentiating the contributions from Internal and External regulatory factors. This general approach will allow us to compare the dynamic expression patterns of multiple datasets. By integrating this results with phenotypical characterization we will be able to identify conditionally essential genes.*

3.3 AIM 3: Experimental Validation of Conditional Essential Gene Predictions in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*

We will validate the conditionally essential gene predictions from AIM 1 and 2 through wet lab experiments in two yeast species.

3.3.1 Experimental Design for Essential Gene Validation in S. cerevisiae Under Three Stress Conditions

The budding yeast, *S. cerevisiae* is one of the best-characterized model organisms with a vast amount of associated functional genomics data. In the past decade, budding yeast has been often used to study gene expression under a variety of stress conditions \cite{27074556, 26888869,26596469,27305947,26849847}. Therefore, we are now able to make highly accurate predictions regarding gene expression and organism viability in various growth/stress conditions without relying on functional genomics data from other species. In this project we will test our top 15% of conditional essentiality predictions under oxidative stress, osmotic stress, and DNA damage stress conditions (top 5% of predictions for each condition).

Deletion strains will be ordered from the Stanford Yeast Deletion Library. For each strain, we will verify the deletion upon arrival through PCR by using a primer specific to the 3' end of the gene and a primer specific to the region downstream of the gene. A pair of primers specific to the KanMX4 cassette will also be used to detect the deletion cassette. For strains where genes of interest are not deleted correctly, we will generate our own deletion strains using a

PCR-based strategy. Briefly, primers with 50 bp homology to the regions immediately upstream and downstream of the gene will be synthesized for PCR of the pFA6a-KanMX6 cassette. This deletion cassette will be transform into *S. cerevisiae* BY4741 strain, and transformed yeast cells will be selected on yeast extract peptone dextrose (YEPD) media plates containing 300 mg/L G418. The deletion strain will be verified by PCR as described above. In our hands, 50 bp gene-specific regions upstream and downstream of gene is enough for deletion of most genes in *S. cerevisiae*. For certain difficult genes, we will increase the gene-specific region to 100 bp.

For the growth assay under stress conditions, *S. cerevisiae* cells will be grown in medium. All yeast strains are initially grown as a starter culture overnight at 30°C. From the starter culture, yeast cells are diluted into fresh medium to an initial OD_{600nm} = 0.2. The cultures are grown to mid-log phase (OD_{600nm} = 0.7). The *S. cerevisiae* strains are serially diluted 4-fold in sterile water and spotted onto YEPD plates, respectively, containing various stressors. Spotted plates were incubated at 30°C and yeast growth was assessed after 3 days.

3.3.2 Experimental Design for Essential Gene Validation in S. pombe Under Three Stress Conditions

The fission yeast, *S. pombe* does not have nearly as much functional genomics data available under diverse stress conditions, compared to *S. cerevisiae*. Therefore we will rely on evolutionary analysis and orthology mapping from *S. cerevisiae* to inform the predictions in *S. pombe*. Although the two yeast species are often considered related, their divergence is estimated to be more than 1 billion years apart \cite{12415314}. Thus *S. pombe* will serve as a proof of principle model organism, allowing us to test our prediction pipeline for species where not a lot of functional genomics data are available, and we have to rely on data from distant species. We will test our top 15% of the conditional essentiality predictions under oxidative stress, osmotic stress, and DNA damage stress conditions (top 5% predictions for each condition).

Deletion strains will be ordered from the the Bioneer *Schizosaccharomyces pombe* Genome-wide Deletion Library. For each strain, we will verify the deletion upon arrival through PCR by using a primer specific to the 3' end of the gene and a primer specific to the region downstream of the gene. A pair of primers specific to the KanMX4 cassette will also be used to detect the deletion cassette. For strains where genes of interest are not deleted correctly, we will generate our own deletion strains using a PCR-based strategy. Briefly, in the first round of PCR, primers with 20 bp homology to the regions upstream and downstream of the gene of interest, respectively, will be synthesized for PCR of the pFA6a-KanMX6 cassette. Primers with 20 bp homology to the pFA6a-KanMX6 will be synthesized to PCR ~300 bp upstream and ~300 bp downstream of the gene of interest. The three PCR products will be stitched together sequentially with a second round of PCR. Stitch PCR of the upstream region and pFA6a-KanMX6 and of the downstream region and pFA6a-KanMX6 are carried out separately. In the third round of PCR, both upstream and downstream stitched PCR products are further stitched together to produce a final product of pFA6a-KanMX6 flanked on the 5' and 3' ends by ~300 bp that are homologous to the upstream and downstream chromosomal regions of the gene of interest. The final PCR product was transformed into *S. pombe* 972h- canonical wild-type (ATCC). The deletion strain will be verified by PCR as described above. In our hands, ~300 bp gene-specific regions upstream and downstream of gene is enough for deletion of most genes in *S. pombe*. For certain difficult genes, we will increase the gene-specific region to ~500 bp.

For the growth assay under stress conditions, *S. pombe* cells will be grown in medium. All yeast strains are initially grown as a starter culture overnight at 30°C. From the starter culture, yeast cells are diluted into fresh medium to an initial OD_{600nm} = 0.2. The cultures are grown to mid-log phase (OD_{600nm} = 0.7). The *S. pombe* strains are serially diluted 4-fold in sterile water and spotted onto YEPD plates, respectively, containing various stressors. Spotted plates were incubated at 30°C and yeast growth was assessed after 3 days.

Impact and Innovation *The results of AIM 3 will provide an in-depth experimental validation of the predicted, both known and previously uncharacterised, conditional essential genes as well as their molecular characterization of gene activity for two yeast species.*

3.4 AIM 4: Development of a Software Package for Essential Genes Annotation and Analysis

In this project we will design and the implement a suite of software tools for the identification and characterization of phenotypes. All algorithms will be developed using publicly available data for model organisms *S. cerevisiae*, and *S. pombe*. The performance of the algorithms will be evaluated "*in silico*" by means of test sets (using cross-validation). The successful completion of this project will provide a proof of concept workflow for identifying conditionally essential gene.

All software tools will incorporate all the algorithms developed as described in AIMs 1 and 2. The algorithms will be prototyped using MATLAB and R as well as scripting languages such as Python. Once refined, we will develop a robust implementation in C/C++/Java. Full unit tests and documentation of the code will be provided to facilitate future improvements and development.

We will create a web portal for this project that will allow the larger scientific community to freely access both the implementation of our algorithms as well as the results of all our phenotype predictions and characterization.

---

4. BROADER IMPACTS OF THE PROPOSED WORK

4.1 Integration of Research into Education

We propose to integrate the above described research activities into graduate and undergraduate education.

Mark Gerstein is the Co-Director of the Computational Biology and Bioinformatics (CBB) PhD program (cbb.yale.edu) at Yale University, and he has been designing and teaching graduate courses in bioinformatics, genomics, and data mining for almost 20 years. These activities could easily be translated into class projects, which may help recruit undergraduates into Yale labs.

In addition, we will take full advantage of the **Yale program for students of underrepresented groups called "Science, Technology and Research Scholars" or STARS (science.yalecollege.yale.edu/stars-home),** which includes Computer Science, Bioinformatics, and Genomics components. **As part of this grant, we are going to design research projects for the STARS undergraduates with the potential of recruiting them for graduate degree programs (MSc & PhD).**

All the tools developed for phenotype prediction will be integrated into Computational Biology and Bioinformatics 752 (*Bioinformatics: Practical Application of Simulation and Data Mining*), a course directed by Dr Gerstein, and taught to undergraduates and graduate students. The course is an introduction to the computational approaches used for addressing questions in genomics and structural biology. The function and phenotype component of the course can be substantially improved by introducing the students to innovative tools to predict gene phenotypes using a variety of data. **This resource represents the integration of many facets of bioinformatics, including functional data, biological network analysis, programming, as well as sets of algorithms applied to address questions about phenotype discovery and gene essentiality**. It will also be integrated into final year projects, and as part

of these projects, students will develop online phenotype libraries and essential gene repositories. The students will also have the opportunity to exchange ideas and expand their networking skills by attending the invited lectures and seminars that will be offered by Dr Paccanaro during his work visits at Yale.

The students will have for the first time, the chance to take part in **in-class Kaggle-like competition projects** (https://inclass.kaggle.com/) focused on designing and developing new machine learning algorithms for phenotype annotations for previously uncharacterized genomes. Also following a positive student feedback we will proceed on extending in-class Kaggle project at a university wide level.

Haiyuan Yu has been an active participant and contributor to the **annual Career Explorations Conference organized by the New York State 4-H Youth Development**. 4-H started over 100 years ago. Currently, with ~500,000 teen and adult volunteers and over 7 million youth members, 4-H is the largest out of school youth program in the US. In the State of New York, Cornell University hosts and organizes New York State 4-H. As the land-grant university of New York, Cornell is committed to community service and has established scores of outreach programs. In 2011, Cornell University earned one of the nation's top recognitions – The Carnegie Foundation for the Advancement of Teaching designated Cornell as an "institution of community engagement." **Dr Yu, worked tirelessly on developing a new focus program, "A new age of biology: working with robots", for the annual 4-H Career Explorations Conference, aiming to expose youth to academic fields and career choices, to develop leadership skills, to provide hands-on experience in a college setting and to introduce youth to Cornell University.**

4.2 Conferences and Workshops

As a "tool is just as useful as the consumer's ability to effectively use it" we plan to reach out to the scientific community and popularize our newly developed methods using reach media interactions such as webinars and hands-on workshops. Also, we aim to present the developed algorithms at scientific conferences as well as at "Open Day" events.

As part of numerous consortia (i.e. Kbase, exRNA, 1000 Genomes, ENCODE), Dr Gerstein will also have the opportunity to disseminate the research findings and make available the developed tools to all his consortia colleagues and collaborators.

We will set up one-day, free attendance, Machine Learning in Bioinformatics workshops that will be led by Dr Paccanaro and will be hosted at Yale University. These workshops will be dedicated to both computer science students as well as experimental biologist that would like to learn more about *"in-silico"* analysis of biological data. All the seminars as well the instruction material will be also made available online following the workshop.

5. PROJECT MANAGEMENT PLAN

The research will be conducted by graduate students and early career personnel under the supervision of Dr Mark Gerstein at Yale University, Haiyuan Yu at Cornell University, and Dr Alberto Paccanaro at Royal Holloway University of London.

In leading this collaborative project, we will draw on considerable experience we have had with other integrative collaborative projects. In particular, Dr Gerstein has been an integral part of the ENCODE Project as well as the modENCODE Project since its inception. Within these he has had a number of leadership roles, as he has co-directed the Networks/Elements Group.

This project will integrate the biological networks expertise of Dr Gerstein with the machine learning and software development expertise of Dr Paccanaro and the experimental assay development of Dr Yu, brining a fresh new perspective to conditionally essential genes prediction. The three group leaders have been collaborating for over ten years on many

network-based approaches for problems in biology. To some degree the collaboration between the three labs will be cemented through knowledge exchange and work visits. As such Dr Sisu (Yale) will have a visiting scientist appointment in Dr Paccanaro's lab and will work closely with his team to integrate the network analysis tool with essential genes predictions. Dr Sisu will also be the project manager and will be the contact person between the three labs. Dr Paccanaro is already scheduled to spend a period of time as visiting professor at Yale University in Dr Gerstein's Lab in the next three years. During this time at Yale he will contribute invited lectures to the computational biology and bioinformatics course led by Dr Gerstein. He will also take this opportunity to visit the lab of Dr Yu at Cornell University.

The three group leaders will have scheduled monthly **conference calls** to exchange details on the project progress and development. Dr Gerstein will also contribute **invited talks** to both Cornell University and Royal Holloway University of London.

Dr Paccanaro will be involved in the design and development of gene essentiality and phenotype prediction tools associated with AIM 1. Dr Gerstein will be responsible for the coordination, designing and development of tools associated with AIM 2 created by Dr Sisu at Yale. As lead of AIM 3, Dr Yu will lead the experimental validation of the functional predictions resulted from AIMs 1 and 2. While AIMs 1, 2 and 3 are led by each lab mostly independently, all three groups will collaborate towards their completion. As such, the Paccanaro group will help with model development and implementation for AIM 2, while the Gerstein group will help with assessment of data quality, standardization and biological interpretation of AIM 1 results'. Both the Gerstein and Paccanaro groups will work closely together to facilitate the implementation of AIM 4 and improve their predictions methods following experimental validation feedback. All three groups will be involved in the design of the experimental validation as described in AIM 4.

The overall progress of the project is summarized in milestones as follows:

[[CSDS to update the milestones]]

Year 0-1.5 The Gerstein lab will work on the development of algorithms for decomposing expression dynamic patterns in contribution for internal and external regulators using biological networks network analysis (AIM 2). Dr Paccanaro will provide technical support for the correct implementation and optimization of the algorithm.

Year 1.5-2. We (Gerstein and Paccanaro labs) will extend the model validations from the model organism (worm, fly) to other more complex systems organisms (i.e. human, mouse, primates, etc.) and thus improve the proposed algorithms accordingly. We will also combine our efforts to implement AIM 3.

Year 3. The work in both labs will be focused on completing AIM 3 and AIM 4. Together, we will also develop a robust and friendly interface for the phenotype prediction and characterization tools. The third year will also be dedicated to publishing collaborative papers describing the newly developed tools as well as the scientific advances resulting from their use.

The three groups will also coordinate the analysis and writing of collaborative manuscripts. To achieve this, we plan to implement regular conference calls between the three groups, and also open them to the larger networks, functional genomics and computer science research community.

We will also take advantage of the plethora of tools available to facilitate collaboration. To this end the software development between the three labs will be hosted on a communal version control system, **github**. In order to guarantee a high standard of our tool, we will employ regular code reviews. Similarly, we will use google drive and online whiteboard tools on a regular basis to enhance the sharing of ideas between the three groups.