

# **BILATERAL NSF/BIO-BBSRC. ABI INNOVATION. LARGE SCALE IDENTIFICATION AND MOLECULAR CHARACTERISATION OF PHENOTYPES**

In 2014, we submitted a grant to the Bilateral NSF/BIO-BBSRC, ABI Innovation call, proposing to develop a methods gene function prediction using biological networks. We thank the reviewers for their constructive reviews that made us completely rethink our proposal. **This new, original proposal focuses on phenotype characterization rather than function prediction.** In the following we address the reviewer comments from last year proposal and show how we were able to improve our submission thanks to their feedback.

First, the reviewers commented that our original proposal offered little preliminary data. This proposal contains a new section that highlights our preliminary results that support our new proposed work. Second, the reviewers pointed out that our aims could have benefited from a better justification and motivation for our methods. The current proposal presents the state of the art in the area of phenotype predictions, highlighting its current limitations, and describing the impact and advancement that our proposed methods will bring in this area. In this way we also responded to the reviewers' comments on properly outlining impact and innovation of the proposed work. Thirdly, the reviewers criticised the originality of our aims and remarked that we had not clearly stated how a positive outcome of the project could have been quantified. The current proposal describes three original aims. For each of them we describe their relevance, impact and innovation and we outline specific "*in silico*" experiments. Importantly, we now included wet lab experiments to further assess the validity of our methods. These will be carried out through the support of our experimental collaborator Dr Haiyuan Yu's lab at Cornell University NY (see Letter of Collaboration). Finally, the reviewers made observations on our brevity in addressing the specifics of the proposed outreach activities, the innovation in the educational plan, and the potential for broader impact. In our new proposal, we clearly indicate the creation of bioinformatics workshops. We also introduce for the first time the concept of "in-class Kaggle"-like competition projects as part of the bioinformatics course, and highlighting throughout the grant the broader-impact that our methods have in extending the field of phenotype prediction and characterization.

## **1. RESULTS FROM PRIOR NSF SUPPORT**

Results from prior NSF support are not applicable since it has been over five years since the end of the last NSF support titled "Development of an Arabidopsis Proteome Chip" (2/1/2008-1/31/2010; DBI 0723722; awarded amount \$335,817.00) that was awarded to the Principal Investigator Dr Mark Gerstein. The Gerstein lab was involved in the microarray analysis, and the development of an online repository for the expression clones, protocols and reagents, that is available to the scientific community. The work from this project resulted in a successful publication [1].

## **2. 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 (including protein coding and non-coding RNAs). For these newly identified genes, as well as for genes whose molecular function and cellular roles are already known, understanding their role in affecting a certain phenotype 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.

From an experimental point of view applying known phenotypic assays on genomic wide scale is expensive and unfeasible. Moreover, designing new phenotypic assays for genomic targets that have not been previously described is a difficult process. At the same time, complex traits are hard to predict and the development of methods for uncovering genotype-phenotype relationships has been identified as one of the major post-genomic challenges [2].

**A fundamental goal in computational biology is therefore to characterize which sets of genes give rise to a given phenotype and how this phenotype is controlled. This fundamental question has never been addressed before on a genome-wide scale, and this is exactly what this project aims to do.**

We are going to look at a systems-level phenotypic characterisation describing the effects of a group of genes rather than that of individual elements. Our goal is to build a state-of-the-art system that, given a set of genes (protein coding and ncRNAs), or even an entire genome, can detect their phenotypes and suggest a hypothesis for their generation, regulation and activity that can then be validated experimentally.

**Our main idea is to conduct a data driven analysis that will integrate information on two levels: molecular (mechanistic) and phenotypical.** At the molecular level we are going to identify the mechanisms that govern gene activity by exploiting expression data. At the phenotypical level, we are going to assign phenotypic attributes to genes and identify sets of genes that share the similar phenotypic characteristics. Finally, since information in these two layers is complementary, we will be able to synergistically combine them to characterise mechanistically, for the first time, phenotype predictions on a genome-wide scale.

This project will be developed as a bilateral collaboration between the groups of Dr Mark Gerstein (US NSF PI) at Yale University and Dr Alberto Paccanaro (UK BBSRC PI) at Royal Holloway University of London. The two PIs have a 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.

This project can be subdivided into three main aims:

**AIM 1: We will develop a computational 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:** While AIM 1 will provide phenotype predictions, it will not inform us on how phenotypes emerge and are regulated, i.e. their molecular characterization. At high-level, genes displaying similar phenotypes are expected to exhibit similar expression trajectories. On these premises, **we will develop a computational model that, focusing on a given groups of genes, will evaluate their expression dynamics patterns, thus uncovering the regulatory effects that govern them.** This will be done by decomposing the gene expression into *internal* contributions, from genes within the group, and *external* contributions, from genes outside the group. Specifically, we propose to use a state space model to represent the temporal gene expression dynamics and identify principal temporal dynamic patterns. In order to handle limited time samples, we plan to use dimensionality reduction approaches that will identify canonical temporal expression trajectories (e.g. degradation, growth, damped oscillation). In this way, we will untangle the regulatory effects from various contributors.

**AIM 3:** While the first two aims are independent of each other, in order to obtain the genome wide molecular characterization of the phenotype prediction, we need to **synergistically combine the phenotypic assignments from AIM 1 with the mechanistic description of gene expression from AIM 2.** For this, we will develop a metric for quantifying the similarity between phenotypes. This will allow us to detect groups of phenotypically related genes as predicted in AIM 1. We will then focus on each of these groups in turn, to identify their corresponding canonical expression patterns and the regulatory effects that govern them by applying the methods developed in AIM 2. In other words, we will combine the two aims in order to provide a molecular characterization of all predicted phenotypes that can be validated experimentally. In AIM 3 we will also integrate all the developed algorithms into a comprehensive software package.

All algorithms will be developed using publicly available data for model organisms including *S. cerevisiae*, *S. pombe*, *C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens*. The performance of the algorithms will be evaluated “*in silico*” by means of test sets (using cross-validation). Furthermore, Prof Haiyuan Yu at Cornell University, has agreed to experimentally validate 5% of our phenotypic predictions in yeast (*S. cerevisiae*).

### 3. BACKGROUND AND PRELIMINARY RESULTS

#### 3.1 Background on Phenotype Prediction and Biological Networks

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 [3]. It uses a hierarchical structure to represent relationship between concepts using a controlled vocabulary [3, 4]. There are a number of publicly available phenotype ontologies for human [5], worm [4], fly [6], mammals [7], yeast [8], etc.

Comparative genomics has been proposed for uncovering gene-trait relationships [2, 9]. This approach begins by constructing phenotypic profiles, which indicate which organism exhibits a particular phenotype – this is similar to the concept of phylogenetic profiles [10]. 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 [11] and flagellar motility [12]. Several approaches have been developed for this comparative analysis. For example, Tamura et al. [13] proposed a rule-based data mining algorithm to associate Clusters of Orthologous Groups of proteins (COGs) with phenotypes. Slonim et al. [14] 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 [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. 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 and Paccanaro, we will develop new state-of-the-art methods for phenotype 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 obtain coherent functional modules to improve gene phenotype prediction.*

### **3.2 Background on Modelling Gene Expression Dynamics Using a State-Space Model**

The state-space model has been widely used in engineering [28]. 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 [29, 30, 31]. 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) [32] 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.*

### **3.3 Preliminary Results**

#### **3.3.1 Preliminary Results on Phenotype Predictions**

Paccanaro and Gerstein have developed a correlation-based method [33] that was able to discover genotype-phenotype associations combining phenotypic information from a biomedical informatics database, GIDEON, with the molecular information contained in COGs [34]. Evolutionary relationships between species are also used in methods that predict phenotypes by exploiting the concept of phenolog. The idea behind these methods is to associate a given gene with the phenotype which is most common among its orthologues [35].

Recently, much research has also been carried out to characterize and predict disease phenotypes. Inherited diseases that are phenotypically similar share disease-associated cellular components: they are linked by common molecular machinery whose normal functioning is somehow perturbed [36]. In other words, the disease modules of phenotypically similar diseases should be located closely on the interactome. The Paccanaro

lab has recently developed a measure that quantifies the distance between diseases at the molecular level by using only their phenotype (the textual description of the diseases), taking advantage of the abundance of high-quality descriptions of disease phenotypes. Briefly, our method mines this extensive biomedical literature to produce an accurate, compact and structured description of the disease based on an ontology framework. This description allows a systematic comparison of pairs of diseases resulting in an accurate similarity score. We have tested our measure by correlating it with the experimentally verified disease similarities at a molecular level, and we showed that it performs significantly better than the current state of the art. Importantly, our method proves that textual descriptions of phenotypes combined with well-structured vocabularies from ontologies provide valuable and under-exploited information for a systematic analysis of phenotype.

*In this project, we will exploit and build upon this important result in order to characterize and predict general organism phenotypes.*

### **3.3.2 Preliminary Results on Networks**

The Gerstein and the Paccanaro labs have carried out projects in biological networks for over a decade. We (Gerstein and Paccanaro) have made extensive contributions in the analysis of genomic data using network frameworks [37]. In particular, we have integrated regulatory networks with gene expression to uncover different kinds of dynamic sub-networks [38]. We also developed methods to analyse the regulatory networks of a variety of species from yeast to human, using a wide range of data [39, 40, 41, 42]. In the following we will give more details on our earlier results in network biology that are most relevant to this project.

**Networks and Function.** Biological networks, normally large in scale, are organized with topological structures in the form of interacting modules. We (Gerstein and Paccanaro) 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 [43]. We have also developed a way to identify nearly complete, fully connected modules (cliques) present in network interactions [44], and we have been using networks to map various kinds of functional genomics data [41]. For example, by mapping gene-expression data onto yeast regulatory network, we identified different sub-networks that are active in different conditions [38].

More recently we have developed a computational approach OrthoClust [45], to extract meaningful new information from gene expression data using biological networks. 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.

**Integrating Networks with Biological Data.** A central problem in computational biology is how to integrate different types of biological data in order to obtain more reliable predictions. We (Gerstein and Paccanaro) have developed several techniques in this area. One method, which is particularly relevant to this project, combined a widely heterogeneous set of biological networks, ranging from co-expression relationships to similar phylogenetic profiles, to predict genome-wide protein-protein interactions. Moreover, we explored the limits of genomic data integration, assessing the degree to which the predictive power increases with the addition of more features [46].

*In this project, we will build upon our expertise in biological network analysis for genome-wide prediction and characterization of phenotypes.*

## 4. RESEARCH PLAN AND METHODS

### 4.1 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 provides a probability  $P$  that a gene has phenotype  $F$ . On the other hand, finding a certain correlation between the profiles of the expression of genes  $X$  and  $Y$  gives 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 (see 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.*

#### 4.1.1 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.

**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,  $P(F_i | E_1 \dots E_n)$ , is defined as:

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

and can be approximated by:

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

where  $(E_1 \dots E_n)$  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  $P(E_j | F_j)$  will be built in a different way from the training set.

For unary relations, the likelihood models,  $P(E_j | F_j)$ , 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<sup>1</sup> [47] 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 problem. 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.

**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 [48].

**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 [49]. We have used these ideas with very good results for predicting protein-protein interactions using the topological properties of networks of interactions observed experimentally [50].

**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 [51]. 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

---

<sup>1</sup> 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].



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 [48].

**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 [8, 52]. 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”).

Next we will extend and test our method on different model organisms such as *C. elegans*, *D. melanogaster*, *M. musculus* and *H. sapiens*. We will use phenotype ontologies and genotype-phenotype associations which are available for these organisms [4, 6, 52, 53]. Clearly, this will involve re-training/tuning the algorithms for each of the species, as we expect the parameters to depend on the different species. 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.*

---

---

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

While AIM 1 will provide phenotype predictions, it will not inform us on how phenotypes emerge and are regulated, i.e. their molecular characterization. At high-level, genes displaying similar phenotypes correspond to small regulatory sub-systems and thus are expected to exhibit similar expression trajectories. In this aim, we will provide a characterization of gene expression uncovering the regulatory effects that govern it. We will develop a novel computational method for decomposing the gene expression into contribution from external and internal regulatory components using state-space models and dimensionality reduction techniques. This will allow us to accurately model the canonical temporal expression dynamic patterns.

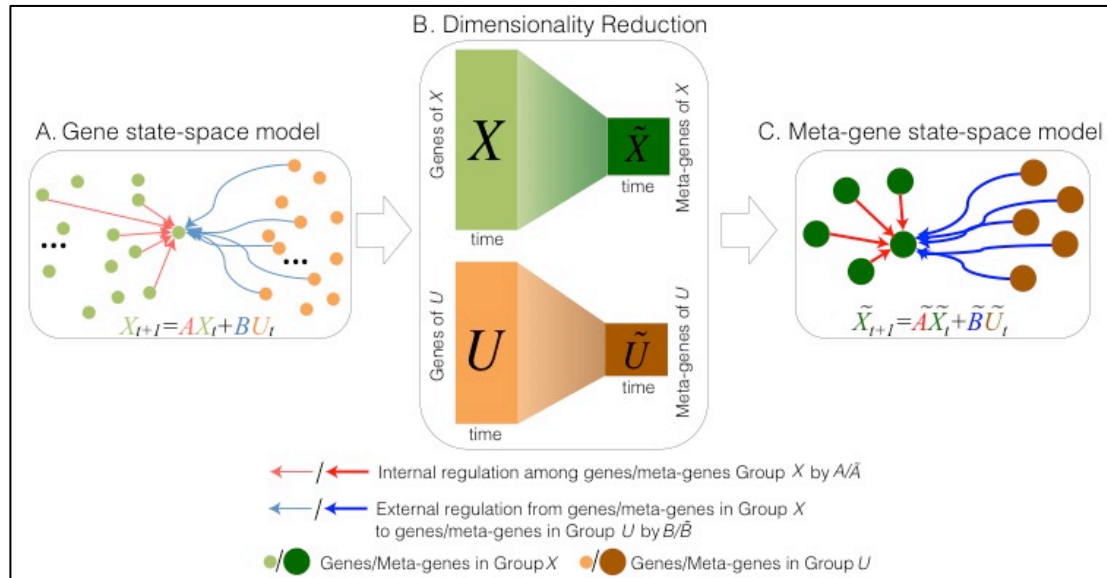
### 4.2.1 Development of a State Space Model for Large-Scale Gene Expression Data

A gene regulatory network is composed of a variety of smaller regulatory sub-systems that define each a particular regulatory function [54, 55]. 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 (Figure 1A). 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  $j^{\text{th}}$  gene expression at time  $t$  to the  $i^{\text{th}}$  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  $j^{\text{th}}$  gene expression in  $U$  at time  $t$  to the  $i^{\text{th}}$  gene expression in  $X$  at time  $t+1$ ).  $\mathbb{R}$  represents the real number domain.



**Figure 1** An outline of our plan for decomposing the high dimensional experimental gene expression data into contributions from the internal and external regulatory components.

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 [56] (Figure 1B). 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  $\tilde{B}$  (Figure 1C).










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 as those exemplified in Table 2.

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.

**Details on Method Development and Validation.** In AIM 2 we plan to develop and validate our state-space methods for gene expression pattern decomposition using publicly available data on worm (*C. elegans*) and fly (*D. melanogaster*). In particular, we will test our methods by analysing the gene expression dynamics patterns during embryonic development of worm

and fly. As such we will use the worm-fly one-to-one orthologs as the *internal* group  $X$  and the worm [57] and respectively fly-specific [58] transcription factors as the *external* group  $U$ .

**Table 2** Examples of canonical temporal expression trajectories obtained by eigenvalue decomposition.

PDP eigenvalue	Real					
	<-1	-1	(-1,0)	(0,1)	1	>1
Canonical temporal expression trajectory (initial)	Vibrating late (VL) 	Vibrating (V) 	Vibrating early (VE) 	Decreasing (D) 	Flat (F) 	Increasing (I) 
PDP eigenvalue	Complex (radius)					
	<1	1	>1			
Canonical temporal expression trajectory (initial)	Damped oscillation (DO) 	Oscillation (O) 	Underdamped oscillation (UO) 			

We will also use the corresponding temporal gene expression levels during embryonic development as described in the modENCODE project [59]. Next, we will use our state-space model and dimensionality reduction techniques to identify and characterize the canonical temporal expression trajectory for both the orthologous genes and species-specific TFs. Leveraging on the available knowledgebase on the dynamics of gene expression during the worm and fly embryogenesis [60, 61], we would consider a positive outcome if our pipeline results would indicate a variation in the level of abundance for mRNAs from genes involved in DNA replication. In the same time we would also expect the DNA replication to be governed by a highly conserved regulatory network.

**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 also allow us to compare the dynamic expression patterns of multiple datasets.*

### 4.3 AIM 3: Molecular Characterization of Predicted Phenotypes

In order to obtain the genome wide molecular characterization of the phenotype prediction, we need to synergistically combine the phenotypic assignments from AIM 1 with the mechanistic description of gene expression from AIM 2.

We will develop a measure of distance between phenotypes that can be used to define sets of genes with similar phenotypes. This will allow us to use the predicted phenotypes from AIM 1 to identify sets of genes with similar phenotypes. Then, the methods developed in AIM 2 will be used to characterize each of these sets of genes in turn.

We will apply this approach to obtain genome wide phenotype prediction and molecular characterization for various model organisms. The results will be validated through wet lab experiments in yeast.

Finally, all the algorithms developed in this project will be built into an efficient and robust software package that will be made freely available to the scientific community through the project website.

#### 4.3.1 *Development of a Measure of Phenotypic Distance*

Our earlier result on disease phenotypes (see section 3.3.1) proves that the textual descriptions of phenotypes combined with well-structured vocabularies from ontologies can be extremely effective in characterizing phenotypes. Here, we will build upon this important result in order to characterize and quantify the similarity between any two phenotypes in a given organism.

To do this, leveraging on the available phenotype ontology, for each term, we will begin by extracting its textual description. This will be followed by a text mining analysis in which we will apply standard pre-processing techniques (stemming, stop word removal, etc.) in order to get a list of terms from MESH ontologies. In this way, a given phenotype will be represented by a set of MESH terms. A distance between two given phenotypes can then be calculated as a semantic distance on the MESH ontology between the sets of terms describing them. A possible variation of this approach will consist in learning a weighting for the different MESH terms. This will allow us to discount often used, and thus non-informative, terms.

**Details on Method Development and Validation.** We will develop our method for *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens*. To evaluate our measure we will exploit the assumption that genes with similar phenotypes are more likely to interact. This will allow us to reduce the problem of evaluating our measure to a binary classification problem, where disease similarity scores are used to predict protein-protein interactions. The performance of our measure will then be assessed through ROC curve analysis.

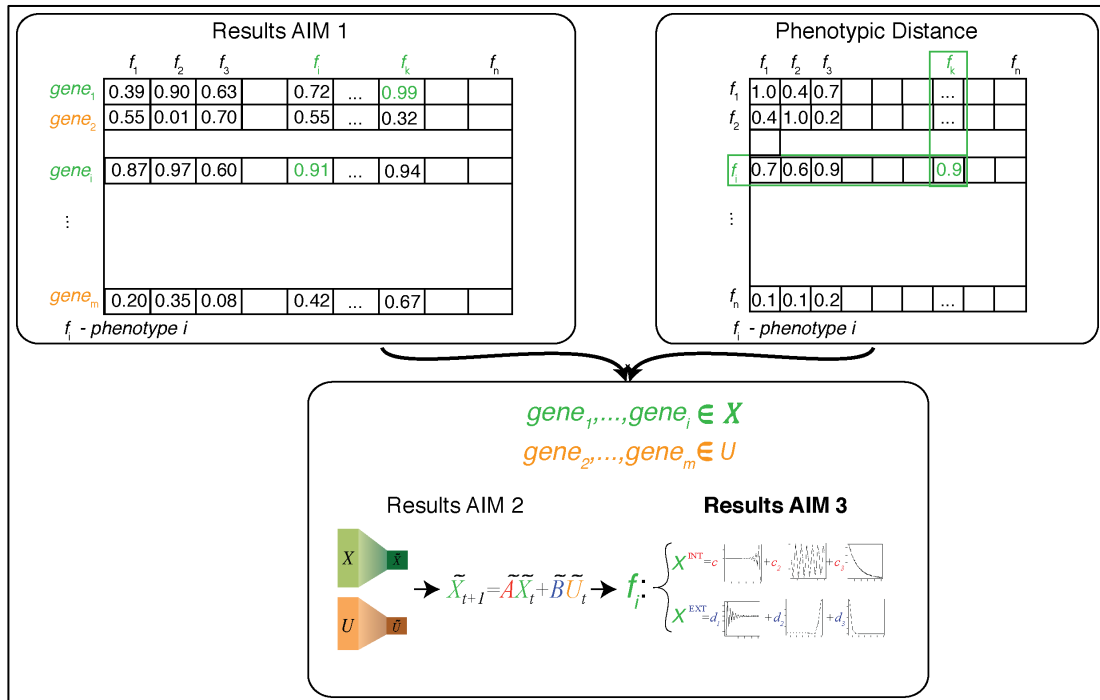
#### 4.3.2 *Molecular Characterization of Predicted Phenotypes on a Genome Wide Scale*

The method developed in AIM 1 will be able to provide a phenotypic score indicating the likelihood that a chosen phenotype can accurately characterize the gene of interest.

In order to apply the methodology developed in AIM 2 for mechanistically characterizing the phenotypes, we need to identify sets of genes that share similar phenotypes (which will constitute  $X$ , the *internal* set of genes of interest). We will begin by choosing the set of genes associated with phenotype  $f_i$  – this will involve setting an organism-dependent threshold which will vary according to the overall distribution of the phenotypic scores obtained from AIM 1. At this stage, we can use our phenotypic distance to expand the internal set with genes whose phenotypes are closely related to  $f_i$ . This will provide us with different granularity on the groupings allowing us to study the phenotypic landscape at various levels. The genes not selected in the internal set will constitute the *external* group  $U$ .

The results will give us an indication of the functional and regulatory activity associated with genes of a particular phenotype. A schematic description of the proposed workflow is shown in Figure 2.

**Details on Method Validation.** We will run our method for *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens*. We will make all our phenotype prediction and characterization available from the project portal (see section 4.3.3). Furthermore, we will experimentally validate our predictions as well as the mechanistic characterization of phenotypes in the context of the studied gene set in yeast. For this we are going to take advantage of our (Gerstein and Paccanaro labs) history of fruitful collaborations with Haiyuan Yu's lab at Cornell University. Dr Yu has extensive knowledge in both experimental and integrative computational biology and has kindly agreed to support our studies by offering to experimentally validate our results (see Letter of Collaboration). Specifically we are going to analyse the genes in yeast regulatory network. We will randomly select 10% of newly predicted phenotypes. For each phenotype we will select the top 5% of associated genes that are best described by that particular phenotype. We will use knockout experiments to validate the gene-phenotype relationship. We will use RT-PCRseq to calculate gene expression levels and knockout studies to test whether the contributions of internal and external regulators to the expression of the gene of interest follow the trend predicted by our methods.



**Figure 2.** Integration of AIMs 1 and 2 into the AIM 3.

### 4.3.3 Software Implementations

In this project we will design and the implement a suite of software tools for the identification and characterization of phenotypes. These tools will incorporate all the algorithms developed as described in AIMs 1, 2, and 3. 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.

**Impact and Innovation** *The results of AIM 3 will provide, for the first time, a genome wide mechanistic description of phenotypes for both known and previously uncharacterized genes for model organisms. This characterization, together with the developed software tools, will be made freely available on the project portal.*

## 5. BROADER IMPACTS OF THE PROPOSED WORK

### 5.1 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](http://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 focus on students of underrepresented groups through a Yale

program called “Science, Technology and Research Scholars” or STARS ([science.yalecollege.yale.edu/stars-home](http://science.yalecollege.yale.edu/stars-home)), which includes Computer Science, Bioinformatics, and Genomics components.

**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. It will also be integrated into final year projects, and as part of these projects, students will develop online phenotype libraries. 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.

## 5.2 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.

## 6. PROJECT MANAGEMENT PLAN

The research will be conducted by graduate students and early career personnel under the supervision of Dr Mark Gerstein (US NSF PI) at Yale University, and Dr Alberto Paccanaro at Royal Holloway University of London (UK BBSRC PI).

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 thus will bring a fresh new perspective to phenotype prediction. Dr Gerstein and Dr Paccanaro have been collaborating for over ten years on many network-based approaches for problems in biology. To some degree the collaboration between the two 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 phenotype predictions. Dr Sisu will also be the project manager and will be the contact person between the two labs. Also Dr Paccanaro will visit the Gerstein Lab and contribute invited lectures to the computational biology and bioinformatics course led by Dr Gerstein.

Dr Paccanaro will be involved in the design and development of 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. While these two aims are led by each lab mostly independently, both 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'. The two groups will work closely together to facilitate the implementation of AIM 3.

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

**Year 0-1.5** The Paccanaro lab will focus on developing machine learning methods for phenotype predictions (AIM 1). The Gerstein lab will provide scientific feedback and validation of the prediction results. In the same time, 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. The successful development of this method will be assessed by a pilot study on worm and fly embryogenesis.

**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. 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 two groups will also coordinate the analysis and writing of collaborative manuscripts. To achieve this, we plan to implement regular conference calls between the two 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 two 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 two groups.

We will also work closely with other investigators from the UK and US to identify additional biological network datasets for integrative analysis, and coordinate the sharing of information with the larger scientific community. On a regular basis, the project results will be disseminated to a broad audience (from senior researchers to middle and high school teachers) through conferences, public workshops and webinars.