**PROJECT SUMMARY** “**ABI Innovation: Neighborly Network Inference**”

**1. Senior personnel**

 **PI:** Dennis Shasha (NYU Courant Institute of Mathematical Sciences)

 **Co-PIs:** Gloria Coruzzi & Manpreet Katari (NYU Biology, Center for Genomics & Systems Biology)

 **Senior Personnel:** Arthur Goldberg (Memorial Sloan-Kettering Cancer Center)

 **Collaborators:** Rodrigo Gutierrez, Catolica Universita de Chile.

**2. Intellectual merit of the proposed activity** Species are being sequenced at a vastly increasing rate. When embarking on the study of a newly sequenced species, researchers would benefit from tools that infer gene interaction networks from experiments on phylogenetically neighboring species. We propose to develop such “Neighborly Network Inference (NNI)” tools. Our vision is to construct species-specific interaction networks of many kinds (transcription factor-binding networks, protein-protein, siRNA-RNA, metabolic etc) for many species *synergistically*. In this vision, every experiment in species *s* will contribute to inferences on *s* and all related species. The experimental bases of our inference will vary from steady state wild type experiments to time series experiments to mutant experiments all on 21 plant species of great practical and research importance (e.g. Arabidopsis, rice, soy). This project will leverage the facilities of the current VirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) developed under an NSF Arabidopsis 2010 Grant (DBI-0445666) including Arabidopsis multinetwork data, analysis, integration and manipulation tools [1]. As output, we will provide a pipeline of tools for Neighborly Network Inference to the community. These tools will include the inference of gene interaction networks in a target species based on measured results in a source species and an approach to select which experiments are likely to be most helpful. While NNI is here described with respect to plants, the framework and basic algorithms extend to any under-analyzed species. This work will achieve one of the main goals of Systems Biology – predicting network states under untested conditions.

**We divide the work into three aims:**

**Aim 1.** **Development of the Neighborly Network Inference Model on Expression Data.** Omic-scale xpression correlation is the basis for clustering, transcription factor-target inference, and many other goals. This aim explains a machine learning framework to infer correlation in a little-studied target species based on experiments in a well-studied source species. The aim also shows a workflow for inferring causal networks when perturbation and time series data, in addition to steady state wild type data, are available.

**Aim 2.** **Proof-of-principle verification of Neighborly Network Inference (NNI) on heterogeneous data** This aim extends the previous model to infer non-expression edges even among distant species. Our preliminary results show high precision inference of metabolic and protein:protein networks in Rice inferred from Arabidopsis The approach uses a fixed set of parameters. After showing the promise of this approach, the aim explains how to choose the parameters in a principled way using machine learning.

**Aim 3. Predicting experimental “Pay-off”: Framework to Determine the Next Best Experiment to Perform** This aim proposes a tool to help experimentalists determine how they should spend their assay resources. The basic idea is to define a notion of the pay-off of a set of experiments. Then the tool tries removing existing experiments to see which already done experiments give the highest payoffs until now. The experiments giving the highest payoff then guide the selection of future experiments.

**3. Broader impacts of the proposed research** This project is the result of a long-standing and highly successful collaboration between biologists at NYU and elsewhere, and computer scientists at NYU's Courant Institute of Mathematical Sciences. The Systems Biology tools and pipelines resulting from this project will empower biologists to use genomic data to predict a spectrum of gene networks in biology with broad applications to agriculture, the environment, and health. In addition to scientific results, this collaboration extends to joint training of graduate students in Systems Biology.