**NSF EAGER Proposal: “Neighborly Network Inference”**

PI: Dennis Shasha (NYU Courant), Gloria Coruzzi (NYU Center for Systems Biology)

**Key problem and Opportunity**. As more and more species are sequenced across all of biology, the questions immediately arise: 1. What does each gene do; 2. How do different genes interact? In a world of unlimited resources, we would want to study each species with the same thoroughness lavished on model species like Arabidopsis, yeast, and C elegans. In practice however, we will have to try to learn as much as we can about each new species s, by a mixture of experiments in s, and inference from species that are closely related to s. That is, we want to learn a lot about newly sequenced species using as few experiments as possible. If successful, this would give more insight for less expense.

**Approach**: Orthology can give us information about gene function. We want to accomplish the same for interactions among genes - gene networks- including expression correlation, protein-protein interaction and so on.

**The question is challenging**: Suppose that we have interactions in one species. Can we learn about interactions in another? If the species are very close, we would think so. If not, maybe or maybe not.

**Innovation**: Apply machine-learning techniques to discover how inference of interactions depends on the closeness of a newly sequenced species s to one or more well-studied species. This approach must answer many challenging questions (and success is far from certain) including: How much does network inference depend on individual gene orthology between genes in s and interacting genes in other species? How much does inference depend on the strength of the interaction(s) in the well-studied species? If genes g1 and g2 in s are orthologous to many interacting pairs of genes in other species, how do we weight the different contributions of those other pairs?

**In-silico Validation**: We will train our machine learning algorithms on well-studied species for which we have some notion of a gold standard. We hope to find robust coefficients in our machine-learning model for network inference.

**Impac**t: Our focus is on plants, but our impact will be on any field of biology in which new species are being sequenced at a rapid rate.

**PI notes to PO on history of Network Inference Proposal:** As you can see from our track record, we are obsessed with the network inference project, because we believe (and two panels agreed) it is novel, sorely needed by the community, will save money, and will advance science.

**1) Plant Genome (Jan 2010, #1025989)**: Reviews said cross species network inference was a novel and sorely needed tool but we were rejected because the panel objected to details of the experimental section. Okamuro suggested we pitch the computational Aim to ABI.

**2) Cross-species to ABI (July 2010, ABI #1062434)**: Got even better reviews, but we were rejected because we proposed to deliver something concrete, so we were told we were too development-like. PO (Beaman) suggested lower budget for higher risk, high pay-off project with a refocus on innovation.

**3) This latest review at ABI (July 2011, ABI # 1147061)** “**Neighborly network inference”** got very mediocre reviews because we are now too vague and they did not like Aim 3. To present our algorithms clearly, we need to do a lot of the work. Machine learning by its nature must fit parameters and explore methods. An EAGER would permit us to explore the solution space enough to formulate a precise direction forward.