CDI Type II: Learning stochastic geometric models for automatic phenotype quantification
From Data to Knowledge K Barnard, K Lin, R Palanivelu, B Pryor, and R Wing U Arizona

Quantifying biological structure is critical for enabling discoveries that link form and function, understanding adaptation to the environment, and explaining organism evolution. Previously, structure analysis has been minimally productive due to difficulties in obtaining biologically relevant data through laborious and time consuming manual measurements. We propose a novel strategy for acquiring morphological data through automated development and exploitation of structure models from image data. This will enable accurate high-throughput extraction of quantified morphology—numbers from structure—that can be efficiently linked to molecular, environmental, and fitness data. The proposed project will minimize painstaking and inaccurate human data collection and/or software development targeted at extracting specific quantities. The resulting broad access to the detailed structure will support discovery in diverse sciences including ecology, developmental biology, and the genetic basis of biodiversity.

To catalyze this transformation we will develop an innovative methodology for learning stochastic geometric structure models for phenotype groups from image data. Using image data coupled with a learning approach is essential for endowing the models with accurate real world statistics. Further, we will develop an approach for rapid extension of models with minimal human input into phenotypes with novel characters. We will develop widely applicable methodologies in the context of three important model organisms with distinct morphologies—*Alternaria* (a fungus), *Arabidopsis* (a model plant), and *Oryza* (rice, a critical crop). These domains will provide a test bed for demonstrating the scientific benefit of easily available structural information.

Intellectual Merits. Stochastic geometric models for organism geometry will enable a more comprehensive study of the role of structural form than what is currently possible. Most imaginable measures of morphology are easily extracted from such models when fit to image data. These quantities can be assembled alongside other complex data sets to computationally identify and explore important hypotheses, thereby **opening up bioinformatics to organism structural form**. Further, high volume screens based on phenotype will be effective because analyzing the data will be automated, and the results will be informatively rich. Thus the project has significant potential to enable a better understanding of the complex relationship between structure and genetics (i.e., phenomics), structural responses to environment, and how it can be experimentally manipulated for basic research and applications in medicine and agriculture.

The project will also advance computer vision and machine learning. The domain provides an ideal test bed for inference on models that combine unknown, constrained, discrete, model topology with parameterized components. Such models apply to much data, but are not well studied. We will also develop a broadly applicable novel approach for automatically translating user sketches into model schemas that can be refined using Bayesian inference.

Broader Impacts. This research will have significant long term impact by enabling discovery in basic biological research and related applied sciences. For example, this project will specifically establish linkage between the morphology of airborne allergenic fungal spores and spore dispersal, which may help predict conditions of high allergy/asthma potential in downwind urban centers. This project will also specifically develop systems to rapidly access morphological variation among closely related plant species, which will have a direct impact on studies in speciation and evolution. Similarly, this project will advance automated accurate quantification of structure of whole plant, which will open up new opportunities for high throughput analysis of environment/genetic effect on morphology that can be exploited to increase production of food and fiber. Finally, this project opens up new methods for automatic species identification that is relevant to biodiversity, biosecurity, and ecosystem management.

The proposed work integrates research and education, and explicitly addresses cross-disciplinary research training at multiple levels to enhance participation in, and appreciation of, computational and mathematics sciences for answering biological questions. The project will disseminate robust software tools for modeling structure and fitting such models to data. The software will enable interested parties anywhere to model organisms, compare models with image data, and extract numbers from structure and link those to other complex data sets.