**SPECIFIC AIMS**

Early clinical interventions can often dramatically curtail the trajectory of diseases, effectively reducing morbidity in a wide array of diseases. However, frequently clinical diagnoses occur following significant, often irreversible disease progression. Early disease prediction from genetics and molecular biomarkers offers to improve the management of individuals with pre-clinical disease states. The overarching goals of this study are to develop robust, molecular-based predictive models for eight chronic, complex diseases using a combination of exome genetic data, key circulating inflammatory and metabolic proteins, and the results from an antigen screen. By doing so, we will gain insight on how these biomarkers and genetic variants contribute to the pathogenesis of these diseases. The diseases studied were selected for having considerable genetic and inflammatory/metabolic components and were well-defined in the Marshfield Clinic Biobank as part of a previous study. Using extensive longitudinal electronic health records linked to this biobank, we are able to assay biomarkers in samples obtained prior to clinical diagnosis of the diseases studied. From these multi-omics data, we will employ and develop a Sure Independence Screening approach followed by an Elastic Net to select informative signals from large molecular datasets. Bayesian Networks will be used to coalesce the predictive signals from genetic variants and protein biomarkers into disease classifiers. The diseases studied are rheumatoid arthritis, systemic lupus erythematosus, relapsing-remitting multiple sclerosis, premature myocardial infarction, chronic lymphocytic leukemia, obsessive compulsive disorder, autoimmune thyroid disease, and axial spondyloarthritis. Further, we will validate the classifer built for rheumatoid arthritis in an independent set of individuals. As the definitive clinical diagnosis of rheumatoid arthritis is typically prolonged, delaying treatment and risking irreversible impairment of mobility, chronic pain and structural damage to joints, an effective molecular prognostic will serve as a useful tool for rheumatologists. **Importantly, the methods and approaches developed for this study will inform and enable other researchers to perform complementary studies.**

**Specific Aim 1: To systematically characterize the levels of 48 key circulating inflammatory and metabolic proteins and exome variants in eight common diseases and matched control groups. Additionally screen antibodies to 800 autoantigens and pathogen epitopes in rheumatoid arthritis and systemic lupus erythematosus.** To understand the level of baseline systemic inflammatory activity, metabolic dysfunction and immune tolerance in eight sets of individuals with disease and controls by measuring (i) exome variants and GWAS-significant polymorphisms (ii) circulating levels of 42 widely-studied inflammatory cytokines characteristic of NFkB-signaling, TH17 signaling, Treg activity, innate immunity, complement activity and Th1/Th2 balance, (iii) antibodies to 800 selected autoantigens and viral/bacterial antigens, and (iv) six metabolic proteins in the insulin-glucose axis in biobanked plasma collected on a 20,000-person cohort selected from a homogeneous Central Wisconsin population.For the disease groups, we will measure all genetic polymorphisms and molecular biomarkers in samples stored prior to a clinical diagnosis. We will also determine which genetic polymorphisms and molecular biomarkers exhibit statistically significant effects in distinguishing between cases and controls. ***We hypothesize that significant differences in the genetics and protein biomarkers exist between each set of cases and their respective controls.***

**Specific Aim 2: To optimally classify cases and controls using genetics and circulating proteins in DNA and plasma obtained prior to clinical diagnosis.** Using the genetic and protein features exhibiting disease signals, we will select informative sets of these genetic and protein biomarkers and develop a Bayesian Network classifier and Penalized Logistic Regression classifier for each of the diseases listed. Robustness of each of the disease classifiers will be assessed using cross validation procedures. ***We hypothesize that the combination of selected (i) triggering event signals from antibodies to antigens, (ii) exome variants, (iii) GWAS-significant SNPs, (iv) levels of inflammatory cytokines in plasma, and (v) levels of metabolic proteins in plasma are capable of predicting disease states in the Central Wisconsin cohort for the eight phenotypes studied.***

**Specific Aim 3: To validate the predictive model for rheumatoid arthritis in an independent sample set.** For the top performing rheumatoid arthritis predictive model developed in Aim2, we will measure all component genetic polymorphisms and protein biomarkers in an independent set of samples from 97 physician-identified rheumatoid arthritis individuals and 200 matched controls. We will calculate the classifier and assess the performance using standard measures of diagnostic utility. ***We hypothesize that the rheumatoid arthritis classifier has the predictive capacity and robustness to effectively delineate cases from matched controls in an independent sample set, thereby validating the classifier.***