Biological Classifier and Clustering Methods for High-Dimensional, Multivariate Measurement Data Sets Using the Grand Canonical Monte Carlo Ensemble

Technology Summary

Bioinformatics and genomics offer tremendous potential to revolutionize medicine, through the development of cell-based therapeutics and personalized medicine. Next-generation sequencing (NGS), microarrays, cytometry, and other techniques now make DNA sequence, mRNA transcriptome, proteomic information, morphologic features, and other types of cell data readily available. However, harnessing this data to identify cell and gene features associated with cell functioning and disease states remains a challenge. Clustering algorithms, principal component analysis (PCA), discriminant analysis, and other mathematical techniques have been used to explore cell data, but they are only mathematically accurate in spaces where the # of measurements are comparable to # of samples (e.g. hundreds of patients with tens of measurements apiece). Unfortunately, the high dimensionally of biological cell data such as NGS, compared with the comparatively low patient sample sizes, makes existing clustering and classification techniques unsuitable. Accordingly, there is a need to improve clustering and classification technologies for high-dimensional data from biological samples.

This technology describes a novel, computer-implemented method for clustering biological samples or data when dimensionality of the measurement universe vastly exceeds the sample size. Using a multi-partition grand canonical Monte Carlo ensemble technique (GCMC), samples are modeled as particles of a grand canonical ensemble and minimization of the ensemble pseudo-energy (e.g. inverse similarity) corresponds to clustering similar particles and thereby determining clusters biological samples. The biological samples can be characterized by data attributes from varying sources (e.g. NGS, other types of high-dimensional cytometric data, observed disease state) and of varying data types (e.g. Boolean, continuous, or coded sets) organized as vectors (as many as 10^9) having as many as 10^9 or more components.

Potential Commercial Applications

The sample clusters can be used for feature discovery, gene and pathway identification, and development of cell-based therapeutics. For example:

- Single-cell mRNA-seq and qPCR measurements: identify potential quality attributes for cell-based therapies
- Medical records: determination of classifiers for patients experiencing rare adverse events or identify prognostic factors for favorable/unfavorable disease outcomes or patients suitable for a therapy.
- Bio-medical data with insufficient data points classification: discoveries of hidden correlations and unknown classifications previously undiscoverable (e.g., diagnostic subcategories, complex interactions).

Competitive Advantages

- Classification when sample dimensionality vastly exceeds sample size.
- Ability to analyze up to 10^9 measurements per sample
- Uses any type of sample data source (NGS, cytometric, medical records, clinical, disease state, mRNA transcriptomes, proteomics)
- Uses any data type (e.g. Boolean, continuous, vector sets)

Development Stage: Early

Inventors: Elaine Thompson, Malcom Moos, Vahan Simonyan

Intellectual Property: PCT/US2018/061348 filed 11/15/2018

Product Area: Software, Diagnostic Screening Tool

FDA Reference No: E-2017-017

Licensing Contact:
Whitney Hastings, M.S., Ph.D.
FDA Technology Transfer Program
Email: FDAInventionlicensing@fda.hhs.gov
Phone:240-402-2232