Bio-Inspired Metaheuristic Optimization Algorithms for Biomarker Identification in Mass Spectrometry Analysis

Syarifah Adilah Mohamed Yusoff, Universiti Teknologi MARA, Malaysia
Ibrahim Venkat, Universiti Sains Malaysia, Malaysia
Umi Kalsom Yusof, Universiti Sains Malaysia, Malaysia
Rosni Abdullah, Universiti Sains Malaysia, Malaysia

ABSTRACT

Mass spectrometry is an emerging technique that is continuously gaining momentum among bioinformatics researchers who intend to study biological or chemical properties of complex structures such as protein sequences. This advancement also embarks in the discovery of proteomic biomarkers through accessible body fluids such as serum, saliva, and urine. Recently, literature reveals that sophisticated computational techniques mimetic survival and natural processes adapted from biological life for reasoning voluminous mass spectrometry data yields promising results. Such advanced approaches can provide efficient ways to mine mass spectrometry data in order to extract parsimonious features that represent vital information, specifically in discovering disease-related protein patterns in complex proteins sequences. This article intends to provide a systematic survey on bio-inspired approaches for feature subset selection via mass spectrometry data for biomarker analysis.

Keywords: Bio-Inspired Algorithms, Bioinformatics, Biomarker, Feature Selection, Mass Spectrometry, Metaheuristic, Proteomics, Swarm Algorithm

INTRODUCTION

Analysis of biomarkers based on their diagnostic and prognostic potentials has been growing as an active area of bioinformatics oriented cancer research (Madu & Lu, 2010). Well known mass spectrometry soft-ionization techniques such as Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry (MALDI-TOF-MS) and Surface-Enhanced Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry (SELDI-TOF-MS) generate high throughputs of proteomics patterns, structures of proteins, from complex mixtures.
such as serum, urine, nipple aspirate fluids and so on. This valuable information paves upon the exploration of facts in proteomics studies viz., characterization of regulatory and functional networks, investigation of precious molecular defect in biological fluids and identification of symptoms of various stages of a disease via development of reagents (Celis & Gromov, 2003). Apart from such valuable explorations, it also provides functional insight pertaining to the development of clinically significant drugs.

Basically the output of any typical Mass Spectrometry (MS) analysis yields a spectrum, which can be represented as a typical xy-graph in terms of ratio of mass to charge ratio (m/z) versus ionization intensities. Significant information of the spectrum comprises of peaks of the intensities with proportional m/z values. Concerning to intensities of peaks that represent protein expression level for certain molecules of peptides, it leads on discovering new biomarkers for a particular disease in different stages. However MS data bears high dimensionality and makes significant numbers of m/z values are correlated or noisy. It implicitly demands the application of robust pattern recognition techniques that can cope up with large amounts of redundant data.

Feature selection, a process of selecting a subset of original features according to certain criteria, is an important and frequently used dimensionality reduction technique for data mining (Guyon & Elisseeff, 2003; Liu & Modtoda, 1998). It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications: thereby speeding up data mining algorithms, and improving mining performance such as predictive accuracy and comprehensibility of results. In biological context, the technique is also called as discriminative gene selection, which detects influential genes based on DNA micro-array experiments. In MS analysis, feature selection plays two vital roles; (1) It aids to construct a feature selection search, which seeks for significant features to discriminate diseases from control samples; and (2) It helps to construct an appropriate classification model that enables the identification of potential biomarkers for further analysis.

In general, algorithms pertaining to feature selection can be typically classified into two categories viz., feature ranking and subset selection. Feature ranking uses all features inherent on the datasets based on primarily rank-listing them using a metric and then discarding those features that falls below a predefined threshold. The threshold is usually set as a substantial score derived from the ranks. In contrast, subset selection searches the set of possible features for the optimal subset. That is, it evaluates a subset of features as a group for suitability.

Further, subset selection algorithms can be classified into three categories viz.: Wrappers, Filters and Embedded (Guyon & Elisseeff, 2003). Wrappers and filters are both most popular feature subset methods applied in order to achieve dimensionality reduction. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a learning model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. However, this drawback can be reduced by injecting some heuristic techniques in the search process to achieve an optimal subset and apply cross-validation techniques to avoid over fitting.

Filters are similar to wrappers in the search approach, but instead of evaluating against a model, a simpler filter-based strategy is evaluated. Filter-based feature ranking techniques rank features independently without the involvement of any learning algorithms. Feature ranking consists of scoring each feature according to a particular method, and then selecting features based on their scores. Filter methods are the most commonly applied techniques in bioinformatics studies since they have proven to be computationally simple, fast and independent of other analysis algorithms. Also they allow features to be quantified and prioritized according to the scores, which is particularly important for biological interpretation. Their main drawback is that they are not optimized to be used with a particular classifier as they are completely independent of the classification
Related Content

Propagation of Front Waves in Myelinated Nerve Fibres: New Electrical Transmission Lines Constituted of Linear and Nonlinear Portions
[www.igi-global.com/article/propagation-front-waves-myelinated-nerve/52977?camid=4v1a](www.igi-global.com/article/propagation-front-waves-myelinated-nerve/52977?camid=4v1a)

The Dendritic Cell Algorithm for Intrusion Detection
Feng Gu, Julie Greensmith and Uwe Aicklein (2012). *Biologically Inspired Networking and Sensing: Algorithms and Architectures* (pp. 84-102).
[www.igi-global.com/chapter/dendritic-cell-algorithm-intrusion-detection/58302?camid=4v1a](www.igi-global.com/chapter/dendritic-cell-algorithm-intrusion-detection/58302?camid=4v1a)

PSO-CGO: A Particle Swarm Algorithm for Cluster Geometry Optimization
Nuno Lourenço and Francisco Baptista Pereira (2014). *Natural Computing for Simulation and Knowledge Discovery* (pp. 1-21).
[www.igi-global.com/chapter/pso-cgo/80054?camid=4v1a](www.igi-global.com/chapter/pso-cgo/80054?camid=4v1a)
Simulation and ABC to Improve the Performance of Emergency Department
Abbas Al-Refaie, Mohammed Shurrab and Ming-Hsien Li (2012). *International Journal of Artificial Life Research* (pp. 15-31).