metabolites

MDPI

Review

Computational Methods for the Discovery of Metabolic Markers of Complex Traits

Michael Y. Lee [1] and Ting Hu [2,*]

[1] Faculty of Medicine, Memorial University, St. John's, NL A1B 3V6, Canada; mylee@mun.ca
[2] Department of Computer Science, Memorial University, St. John's, NL A1B 3X5, Canada
* Correspondence: ting.hu@mun.ca; Tel.: +1-709-864-6943

Received: 1 February 2019; Accepted: 1 April 2019; Published: 4 April 2019

Abstract: Metabolomics uses quantitative analyses of metabolites from tissues or bodily fluids to acquire a functional readout of the physiological state. Complex diseases arise from the influence of multiple factors, such as genetics, environment and lifestyle. Since genes, RNAs and proteins converge onto the terminal downstream metabolome, metabolomics datasets offer a rich source of information in a complex and convoluted presentation. Thus, powerful computational methods capable of deciphering the effects of many upstream influences have become increasingly necessary. In this review, the workflow of metabolic marker discovery is outlined from metabolite extraction to model interpretation and validation. Additionally, current metabolomics research in various complex disease areas is examined to identify gaps and trends in the use of several statistical and computational algorithms. Then, we highlight and discuss three advanced machine-learning algorithms, specifically ensemble learning, artificial neural networks, and genetic programming, that are currently less visible, but are budding with high potential for utility in metabolomics research. With an upward trend in the use of highly-accurate, multivariate models in the metabolomics literature, diagnostic biomarker panels of complex diseases are more recently achieving accuracies approaching or exceeding traditional diagnostic procedures. This review aims to provide an overview of computational methods in metabolomics and promote the use of up-to-date machine-learning and computational methods by metabolomics researchers.

Keywords: metabolomics; complex diseases; biomarker discovery; machine learning; feature selection; classification; ensemble learning; artificial neural networks; genetic programming

1. Introduction

Over the past decade, the metabolome has been deemed the final frontier for broad, biochemical databases of organismal information among the well-established fields of genomics, transcriptomics, and proteomics [1]. Metabolomics is the study of quantifying metabolites and mapping their complex interactions within this domain, which is comprised of the total set of small molecules (<1500 Da) present in cells, tissues, organs and biological fluids [2,3]. It is the final downstream component of the biochemical stages, involving genes, RNA, proteins and environmental factors, ultimately yielding phenotypic changes in an organism [2]. Since metabolism crucially involves important physiological processes that diseases often alter, metabolomics analyses can be used to detect disease-driven changes from the levels of thousands of metabolites, allowing for enhancements in current diagnostic methods and discoveries of specific, perturbed metabolic networks. The advantage of using metabolomics is therefore derived from its provision of a functional readout of the physiological state of an organism. This is because metabolites act as direct signatures of biochemical activity, whereas genes and proteins may be affected by epigenetic regulation and post-translational modifications. In other words, genomics reveals what may have occurred, whereas metabolomics reflects what certainly occurred.

Importantly, metabolomics may hold the key to tackling the challenges associated with complex diseases, which are caused by an intricate interplay between an individual's genes, environment and lifestyle [4]. Interestingly, most diseases lie under this umbrella term, which include, but are not limited to cancer, cardiovascular disease, diabetes, arthritis, obesity and dementia. It is well known that the classical Mendelian patterns of inheritance are not observed within these illnesses. Rather, expression of certain correlational genes may increase risk of contraction, but does not guarantee incidence; instead, toxins from the environment, drugs consumed over one's lifetime, poor diet and lack of exercise in combination with such genes would likely lead to disease onset. Therefore, researchers of complex diseases must identify methods to overcome the challenges of deciphering the quantitative influence of risk-associated genes in comparison to non-genetic factors. Metabolomics offers a solution to this by allowing the individual influences of genetics, environment and lifestyle to converge onto the metabolome as a terminal downstream domain of products. This holistic approach allows metabolomics researchers to discover biomarker signatures that capture the multiple major factors driving the complex disease. Ultimately, these panels can help to diagnose at-risk complex disease patients in the clinic and even predict onset years before symptoms arise using prodromal metabolomes [5]. In addition to the clinical benefits, it grants researchers a useful visualization of how the complex metabolic networks differ with and without disease influence. Research for metabolic marker discovery spans a fast-growing array of prevalent disease areas, such as breast cancer, osteoarthritis and Alzheimer's [5–7].

Although rich quantitative datasets may contain valuable information, the extents of their utilities are limited by the appropriateness of the selected statistical and computational methods of analysis. Since these datasets contain hundreds of features, the value of an appropriate method would be derived from its ability to account not only for the effects of each metabolite in isolation, but in a multivariate manner with consideration of interaction-based effects. Thus, while recent advancements in analytical chemistry techniques, such as nuclear magnetic resonance (NMR) and mass spectrometry (MS), have made it possible to quantify hundreds of metabolites within a reasonable time frame, these techniques must be coupled with fitting statistical and computational algorithms to translate the data into a practical application in the clinic [8]. Unfortunately, the majority of metabolomics studies historically have not employed optimal methods for biomarker discovery, perhaps due to a lack of statistical and computational expertise among metabolomics researchers, which has spurred the publication of instructional and guideline-setting papers in the field [9–12]. Today, the Human Metabolome Database reports the existence of over 100,000 metabolites in the human body [13]. As analytical methods improve with regard to their discriminatory ability and efficiency, the quantifiable metabolome and its associated datasets will continue to grow, raising the relevance of powerful, heuristic computational methods to the forefront and placing a greater importance on their delineation to biological researchers.

The aims of this review are three-fold. First, we will outline a general workflow of the steps required from a biological question to metabolic marker discovery. Second, the current authors will provide an overview of current metabolomics research within prominent complex diseases to identify gaps and trends in computational methods' use. Third, this review will discuss the benefits and costs of using different computational methods and highlight more recently-applied, promising machine-learning algorithms, including ensemble learning, artificial neural networks and genetic programming.

2. Metabolic Marker Discovery

The general workflow of metabolic marker discovery (Figure 1) typically involves forming the biological questions, extracting metabolites from cells, tissues or organs, quantifying metabolites using NMR or MS, preprocessing data to remove irrelevant biases, selecting a biomarker panel and constructing a predictive model through feature selection and classification, typically using machine learning algorithms.

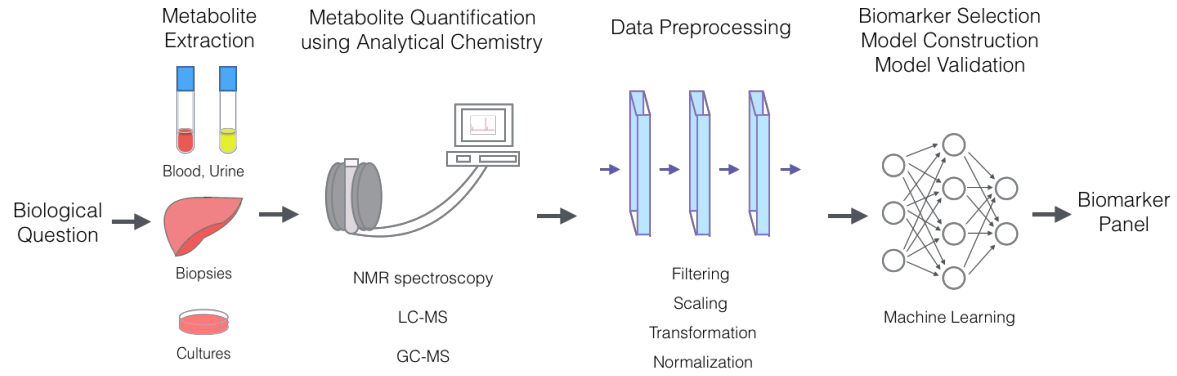

Figure 1. General workflow of the metabolic marker discovery process. Metabolite extraction often can be performed from cells, tissues, organs and fluids, including blood and urine. Metabolite quantification is performed using analytical chemistry techniques, such as LC-MS, GC-MS and NMR, which provide concentration values for each metabolite in solution. Biomarker selection and model construction are conducted using various machine learning algorithms.

There are generally two types of metabolomics studies: targeted and non-targeted. In non-targeted studies, the global metabolic profile is assessed. Thus, all detected metabolites in a sample are given the opportunity to be included in the biomarker panel. Hypotheses are not tested in this approach, but rather formed. In contrast, targeted studies focus on a selective group of metabolites to enhance specificity, precision and accuracy in testing a specific hypothesis. Furthermore, a targeted approach is useful in validating results from a global metabolic profiling (non-targeted) study [14]. It is worth noting that quantifying the data in a non-targeted manner is far from a reality due to the limitations of NMR and MS and the broad diversity of metabolite structures. However, today's technology provides sufficient data for powerful, multivariate computational algorithms to classify people for current and future disease states with accuracies approaching or exceeding current diagnostic measures in various disease areas [5,15,16].

The two most common metabolite quantification techniques used in metabolomics are MS and NMR spectroscopy. In MS, the metabolite is ionized before analysis. Using charged ion modes, the ion signal is converted into mass spectra. By examining the resulting peaks, molecular mass is determined with the mass-to-charge ratio. There is a diversity of MS techniques that exist for complementary purposes, such as gas chromatography MS (GC-MS) and liquid chromatography MS (LC-MS). Each type takes advantage of certain physicochemical properties of the assessed metabolites to separate samples into their constituents. Metabolites with low molecular weights are analysed using GC-MS, whereas LC-MS is capable of evaluating a higher weight range. MS is able to provide quantitative data with high sensitivity and selectivity. In contrast, NMR spectroscopy employs a magnetic field to exploit angled spins and properties of atoms in the molecule of interest, which consequently absorb and re-emit electromagnetic radiation. Detectors on the NMR device use this signal to produce results that are comparably more quantifiable and reproducible and do not destroy measured samples. Furthermore, although MS can detect a broader range of metabolites, it is less time efficient in comparison to NMR; however, both methods are used to examine a combined, wider range of metabolites [14].

Data preprocessing is the general preparatory stage of the data, which ensures that the resulting dataset can be analysed without major issues. Each step of the process contributes to an overall removal of biases and incomplete features. For reference, patients in a metabolomics study belong broadly to the category of samples (or rows), whereas metabolites are classified under features (or columns). First, normalization is one critical step in this process and includes various variable types such as sample normalization, probabilistic quotient normalization, and quantile normalization [17,18]. Sample normalization simply requires tying all samples to a standard value for one particular metabolite in an attempt to resolve significant discrepancies across entire samples due to varying fluid dilution levels. Second, filtering is notably valuable for its ability to eliminate metabolites

that have an excessive percentage of missing values across samples and a concentration constancy independent of group classification. A tolerable percentage of present data would typically be 80%, and the maximum relative standard deviation for constancy evaluation would be approximately 15% [9]. Regarding transformation, it is common to observe replacement of all values with their logarithmic outputs (i.e., x becomes $\log(x)$). The logarithmic transformation improves the normality of the data distribution, since metabolomics data have been shown to mostly follow a log-normal distribution [17]. This also conveniently provides the benefit of working with values within a much narrower range, improving pattern visualization and interpretability for datasets containing extremely large values. Finally, scaling provides a way to address large differences in metabolite levels across patients through standardization down each feature. Typically, the sample mean is subtracted from each data point and divided by the sample standard deviation, removing potential biases related to absolute quantities [9]. Upon completion of the appropriate preprocessing methods, metabolomics researchers may proceed to statistical and computational analysis.

Population-based metabolomics looks for metabolic markers that can provide the best discriminating power between the diseased cases and healthy controls. These metabolic markers in turn can help us develop highly-cost-efficient and effective drugs that target enzymes involved in key processes for better disease treatments, as well as construct a computational model to predict the clinical outcome for new patients [9,19–21]. Biomarker discovery in omics science usually follows a three-step scheme. In the following text, we discuss the objectives, most commonly-used methodologies and the challenges of each step.

The first step is biomarker selection, or attribute selection, where only the most relevant bio-attributes are identified. The necessity of attribute selection is due to the high-dimensionality of most omics data, where hundreds to a million attributes can be considered for their potential association with diseases. Removing irrelevant attributes reduces the computational overhead of downstream analyses, simplifies the learned model and guides the search for biomarkers since the true signal in the data will be more predominant after the noise is removed [22,23].

For attribute selection in omics, most existing studies use univariate tests, such as Wilcoxon, Kruskal-Wallis tests, or additive multivariate analyses, such as logistic regression, least squares regression (LSR) or discriminant analysis. Attributes are examined separately or combined additively for their association with the disease outcome, and only those with significant main effects are usually selected. Such analyses inherently overlook the synergistic non-linear interactions among multiple attributes. Given the complexity of human diseases, it is more plausible that multiple factors interact synergistically, and one factor's effect on the disease depends on others'. However, looking for combinations of attributes exposes us to an enormous search space since the total number of all possible combinations with all orders for n attributes is 2^n. Even a small number of $n = 10$ translates to 1023 subsets of attributes to be tested. For $n = 1000$, enumerating all combinations of only the orders of two and three (i.e., pairwise and three-way synergy) requires 499,500 and 166,167,000 tests, respectively. If we used all the computers currently known on this planet, it would still be impossible to exhaustively enumerate and test all possible combinations of attributes of all orders by a meaningful deadline. Therefore, powerful heuristic search algorithms are needed [24,25].

The second step is model construction. This step uses identified important biomarkers to construct a classification model that can predict a new subject with a high or low risk of developing the disease. Model construction is usually carried out using machine-learning algorithms through a training process on population-based omics data [9]. Although biomarker selection and model construction depend on each other since only the selection of the most relevant attributes can yield an accurate and general prediction model, they are usually done separately. A filter method is usually employed to select the attributes, and the subsequent classification algorithm for model construction is independent of the filter method. These two steps can also be wrapped in an iterative process to further refine their results. That is, a heuristic search algorithm provides a subset of attributes, and model construction

trains a model and feeds back to attribute selection in the next round. However, such an iterative method imposes a large computational cost due to repetition.

The third step in biomarker discovery is biological interpretation and validation of the discovered biomarkers and the constructed predictive model. This is particularly challenging since many machine learning algorithms produce a "black box" model, which can be uninformative for interpretation and validation [26,27]. In many other data-rich fields, such as information technology and finance, prediction accuracy alone may often be sufficient for decision making. However, model interpretation would be particularly important in bioinformatics since the biological mechanisms underlying the model must be understood for us to transfer knowledge to clinical applications.

Model validation in metabolomics typically involves a random splitting of metabolomics samples into 80% training and 20% test sets; however, this ratio may depend on the number of patients and metabolites available for analysis. Furthermore, cross-validation is a useful technique that repeats this process multiple times with different training and test sets, ultimately utilizing the average of the evaluated model validity measures. Additionally, it is important to be aware of the risk of achieving a local optimum rather than the global, but the loss value of such an event may depend on the complexity of the generated model (particularly in artificial neural networks) and may be mitigated with cross-validation, ensemble learning and population-based model search [28]. Feature selection and model construction use training datasets, and the evaluation of a predictive model should always be reported using the unseen test set. This ensures the generalization of the trained model that can translate to future incoming data. In addition to statistical validation, for bioinformatics research on metabolic marker discovery, it is crucial to use independent data in order to replicate the findings and to include follow-up biological experiments to further validate the mechanistic hypotheses generated by the informatics studies.

3. Current Research in Metabolomics, Complex Diseases and Computational Methods

The current literature regarding the investigation of complex diseases using computational methods in metabolomics is rapidly growing. Metabolites are small endogenous or exogenous molecules that play a direct role in energy homoeostasis, macromolecule synthesis, waste elimination and biological regulation [2,3]. These molecules exist in cells, tissues, organs and fluids, including cerebrospinal fluid (CSF), blood and urine. Methods of metabolite extraction are especially important for future clinical applications of metabolomics since safety and cost are likely to influence the adoption of a new diagnostic test. Metabolomics offers a useful new entry in the world of diagnostics, as acquiring blood or urine samples is minimally invasive to non-invasive. Furthermore, metabolomics has a unique capacity to provide insights into an individual's physiological state and capture the multi-causal nature of a complex disease. Today, the majority of complex disease research has shifted from univariate and additive multivariate techniques to newer, more powerful multivariate methods. This review sets its focus on Alzheimer's, breast cancer and osteoarthritis, as complex diseases with substantial metabolomics research history and recent advancements in the application of computational methods to the disease area.

3.1. Alzheimer's Disease

Alzheimer's disease (AD) is a neurodegenerative disorder that is characterized by progressive cerebral atrophy and hypometabolism [29]. Clinical symptoms include memory deficits, language challenges and personality changes. The AD population in the U.S. is projected to triple by mid-century, highlighting the urgency and utility of having concrete developments toward an effective treatment or cure [30]. Currently, amyloid beta (Aβ) plaques and neurofibrillary tau tangles are believed to be the primary neuropathological substrates contributing to pathogenesis [31]. This leads to a slow buildup of plaques and tangles over the course of the prodromal (20 years) and clinical phases (8–10 years), ultimately ending in mortality [31]. However, recent clinical trials that have successfully cleared Aβ plaques from AD brains have failed to demonstrate symptomatic improvements [32].

With only four non-curative FDA-approved drugs and an abysmal failure rate for clinical trials (i.e., 99.6%), efforts to investigate the early disease stages and develop novel diagnostic measures for preclinical prevention and management have become increasingly present and valued in the field [33].

One early study in AD metabolomics revealed that sulfatide species (a class of myelin-specific sphingolipids) were reduced significantly in brain tissue lipid extracts from patients with mild AD [34]. Furthermore, ceramide levels were discovered to be increased three-fold in white matter. These findings were determined using linear correlations and analysis of variance (ANOVA), which provided enough information to support that there was an associative relationship between one particular class of species and dementia severity score, but did not account for the combined effect of multiple metabolites. Years later, Han et al. showed that it was possible to achieve a similar finding from a blood sample quantification of over 800 different lipid species using shotgun lipidomics [35]. A metabolic signature was formed using this "broad-stroke" method of metabolomics study. This analysis revealed the value of analysing large datasets to establish a metabolic signature. Yet, Wilcoxon rank sum tests, a univariate method, were used to uncover the significant differences between the AD and control groups amongst the complex combinatorial and interactive possibilities of these hundreds of metabolites.

Wang et al., addressed this statistical limitation in their study, which examined a relatively large cohort of 172 individuals with an additional third group of mild cognitive impairment (MCI; considered to be an early form of AD) patients [36]. The researchers quantified the concentrations of 238 small-molecule core metabolites (including fatty acids, amino acids, nucleic acids and carbohydrates) from plasma samples and subsequently employed several machine-learning methods for multivariate analyses, such as logistic regression, principal components analysis (PCA) and partial least-squares-discrimination analysis (PLS-DA) to determine that six metabolites, including arachidonic acid, *N*,*N*-dimethylglycine, thymine, glutamine, glutamic acid and cytidine, accounted for the differences between AD patients and controls. A similar analysis was performed for the MCI group against the healthy controls, unveiling five important metabolites, three of which were shared for the AD comparison. The area under the curve (AUC), a measure of classifier performance, yielded high scores of one and 0.998 in a training set, for the AD and MCI groups against controls, respectively. This study demonstrated the utility of an analysis that could capture the multi-factorial nature of AD within a single computational model, leading to a more accurate biomarker panel.

Most recently, Varma et al. analysed serum samples from two longitudinal cohorts of 767 prodromal individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and 207 preclinical AD samples from the Baltimore Longitudinal Study of Aging (BLSA) [5]. Two machine-learning methods, support vector machine (SVM) and random forest (RF), were used to identify a 26-metabolite panel that performed with 83.33% accuracy, 86.67% sensitivity and 80% specificity. The longitudinal nature of this study allowed for an evaluation of a host of metabolite correlates with MRI measures of brain atrophy, AD pathology (i.e., Aβ concentrations in CSF), conversion risk to incident AD and cognitive performance over time. Enhanced blood levels of sphingolipid species were found to be correlated with post-mortem AD pathological severity and preclinical disease progression. Importantly, the uncovered 26-metabolites were involved in various AD-related pathways, including tau phosphorylation, Aβ metabolism, acetylcholine biosynthesis and apoptosis. This was a multi-centre study with regard to collection and analysis, which allowed for the acquisition of the largest sample size to date for an AD metabolomics study. The use of SVM and RF, which are superior multivariate methods to PLS-DA and other additive algorithms when working with large, highly-complex datasets, allowed these authors to extract more accurate information from the ADNI and BLSA data [5].

Overall, AD research is trending in the direction of using computationally-robust, multivariate methods to develop new diagnostic tests that are safer, more cost-efficient, and have similar or greater accuracy rates than current neuropsychological and imaging techniques (which are estimated at 77%) [37].

3.2. Breast Cancer

Although cancer incidence in the U.S. has declined since 1991 by 26%, it continues to be one of the 10 leading causes of death with an estimate of over 600,000 in 2018 [38]. Early detection of cancer continues to be a viable prevention strategy, but has several issues. One particular problem is that researchers must grapple with cancer's paradoxical nature to occur commonly across a lifetime, but rarely present itself at a single point in time, which magnifies the challenge of achieving acceptable test performance [39]. Current biomarkers, such as prostate-specific antigen (PSA) and carcinoembryonic antigen (CEA), have been less useful than expected either due to a low positive predictive value (PPV) or lack of survival benefit [40,41]. With the complex array of contributors to cancer, it would be prudent to consider a multivariate method for screening and diagnostic purposes.

Across cancers, metabolism, the intricate collection of intertwined pathways of energy substrates and enzymatic regulation, is dramatically altered. The central shift involves a phenomenon termed the Warburg effect, which represents the change toward the use of aerobic glycolysis to generate adenosine 5′-triphosphate (ATP) and lactate, even though it is less efficient than oxidative phosphorylation [42]. However, beyond this cellular shift, the hallmarks of cancer, including selective growth and proliferative advantage, altered stress response favouring overall survival, vascularization, invasion and metastasis, metabolic rewiring, an abetting microenvironment and immune modulation, reveal a deeply physiological nature to the disease regardless of the cancer's origin [43]. With metabolomics at the forefront of data-driven physiological research, the field has been an important area of research for cancer with notable developments in recent years.

In particular, breast cancer is the source of 25% of all cancer cases and causes over 500,000 deaths annually [44]. Survival rates depend greatly on early detection, which is often times expensive with imaging methods, such as mammography and magnetic resonance imaging (MRI) [45]. Furthermore, mammography has been shown to miss approximately 15% of breast masses, and surgical biopsies are necessary to confirm definitively the malignancy of the tissue [46]. Testing for metabolic markers of breast cancer may allow for the development of less expensive, less invasive, and more accurate diagnostic techniques.

Similar to AD, cancer research began with univariate statistical analyses during its early years of biomarker discovery. Several studies used such techniques to infer biomarker significance, examining only a handful of metabolites [47]. Furthermore, most of these studies do not provide values supporting the utility of the biomarker, such as sensitivity, specificity and AUC [9]. As analytical chemistry techniques have improved in scalability and efficiency over the years, more studies have begun to utilize multivariate methods of analysis. In 2011, Hilvo et al. used ultra-performance LC-MS to assess the lipids in normal breast tissues compared to those that were cancerous [48]. In particular, the kernel-based orthogonal projections to latent structures (K-OPLS) method was used to generate a predictive model for estrogen receptor (ER) status based on altered lipid concentrations. Furthermore, a validation cohort was used showing similar results. The AUC was found to be 0.94 and 0.88 in the training and validation sets, respectively. In 2015, Huang et al. utilized GC-MS to profile serum samples from patients with malignant and benign breast tumour, as well as healthy controls. They then employed random forests (RF), a machine-learning multivariate technique, to assess the quality of their identified relevant serum metabolites [49]. The prediction accuracy, sensitivity and specificity between malignant breast cancer patients and healthy controls were 100%, 97% and 98%, demonstrating the assessed metabolites' high performance as predictors.

Perhaps one of the most computationally-forward-looking studies to date has been a recent study published in 2018 on using deep learning, a subset of the neural network category of machine-learning methods, to predict accurately estrogen receptor status in breast cancer samples [6]. In the study, feed-forward networks, a framework utilizing deep learning, was compared to other machine learning techniques, including RF, SVM, prediction analysis for microarrays (PAM), generalized boosted models (GBM), recursive partitioning and regression trees (RPART) and linear discriminant analysis (LDA), with data from 162 metabolites. The results demonstrated that deep learning with its AUC of 0.93 was

superior to all other tested methods. Moreover, they were able to uncover eight pathways involved in breast cancer, including central carbon and glutathione metabolism, which were not revealed from the other analyses. Overall, these researchers showed that deep learning has utility within the scope of medium-sized databases and can offer network knowledge between metabolites that other machine learning techniques may lack the capability to provide.

Across research studies for breast cancer, there appears to be evidence of a shift toward more computationally-powerful tools to understand large, complex cancer metabolomics databases. As one of the primary recommendations of the current review, this shift is necessary to improve our current biomarker panels and adapt to the rapidly-increasing number of metabolites and patients available for analysis. Utilizing the latest algorithms, such as feed-forward networks and select heuristic techniques, will promote the development of more representative models with greater potential for translation to practical, clinical applications. In a field like cancer where early diagnosis may mean the difference between a slim one-year survival rate and a simple surgical resection, advancements in metabolomics and subsequent knowledge translation efforts in diagnostics will contribute notable differences to the lives of those soon to be afflicted.

3.3. Osteoarthritis

Among the global population of individuals over the age of 60, osteoarthritis, a degenerative disease of joint cartilage and underlying bone, has a notable prevalence of 10% [50]. It is the most common type of arthritis and the leading form of disability in developed countries [51]. In particular, knee osteoarthritis, which accounts for over 80% of the disease burden being the primary cause of mobility-based disability, has doubled in prevalence in three generations [52]. Molecular theories explaining the pathogenesis of osteoarthritis primarily involve the aging process in association with inflammation, senescence, mitochondrial dysfunction and oxidative stress and changes in energy metabolism and cell signalling [53]. Other important predictors include old age, female sex, overweight status and obesity, muscle weakness, knee injury, frequent joint use, bone density and possibly dietary factors [54]. Therefore, metabolomics can play an important role in quantifying the multi-faceted character of osteoarthritis, especially since the combination of genetic markers and epidemiological factors, such as age, sex and BMI, has been shown to produce a relatively low AUC (i.e., 0.668) [55].

In 2014, Zhang et al. conducted a seminal study on the classification of osteoarthritis into subtypes based on metabolomics data [56]. This study used synovial fluid samples in an effort to enhance the connection between the analysed metabolome and physiological reality. With 80 osteoarthritis patients, the researchers employed PCA, cluster analysis and PLS-DA to perform a multivariate analysis on 168 quantified metabolites. These methods yielded results demonstrating that osteoarthritis was actually composed of two distinct groups, as a result of differences in the levels of 86 unique metabolites. This novel finding provides a deeper understanding of the disease phenotype, which could be further investigated in physiological and cellular studies to identify differences in molecular targets and ultimately improve drug specificity. The lack of curative drugs for the disease underscores this need for more knowledge-building, data-driven investigations in contrast to drug development efforts.

Another study examined knee osteoarthritis, developing global serum profiles for 60 individuals (including a control group) using a dataset of 106 metabolites [57]. With PCA and PLS-DA, the researchers identified a 14-metabolite signature, involved in the metabolism of energy, purines, amino acids, fatty acids, lipids and glycolysis. It was found to have an accuracy measure of 0.662. One limitation of the study was the use of serum as opposed to synovial fluid, which may have provided improved sensitivity and specificity values. However, providing evidence for the use of serum may be beneficial since it has safety and cost benefits over other fluids, despite its potential shortcomings in accurately reflecting the products of the disease process.

In 2018, Hu et al. demonstrated the utility of genetic programming (GP), a heuristic multivariate evolutionary machine-learning technique, in osteoarthritis metabolic marker discovery [7,58].

The authors applied the process of evolution on computer models, generating hundreds of potential models, selecting for the best-performing ones, breeding them together to produce children and repeating the process. Iterating through hundreds of generations with a dataset of 389 samples and 167 metabolites ultimately led to the discovery of nine key metabolites, specifically arginine, C16, C18:1, isoleucine, nitrotyrosine, ornithine, taurine, threonine and tyrosine, several of which had not been reported previously in the literature. Furthermore, genetic programming was found to perform more highly than logistic regression (a non-heuristic method) with AUC values of one and 0.91, respectively.

Overall, similar to other complex disease areas, osteoarthritis appears to be trending toward the use of multivariate, heuristic approaches that are apt at generating high-performing predictive models for the endless ways in which the upstream pathways of genes, RNA and proteins may converge.

4. Advanced Learning Methods for Metabolic Marker Discovery

The previous section delineated the use of more recent and advanced machine-learning methods, revealing significant improvements on metabolic marker selection and predictive model construction as a result. Table 1 summarizes the most commonly-applied classification algorithms for metabolic marker discovery, as well as a set of less utilized, but potentially powerful methods that we will spend more length explaining in this section.

Table 1. Machine-learning algorithms and their example applications to metabolic marker discovery.

Algorithm	Description	Examples
Logistic regression (LR)	Use a logistic function to fit a regression model for categorical outcome prediction.	[59]
Partial least squares-discriminant analysis (PLS-DA)	Find a linear subspace of high-dimensional explanatory variables to maximize the covariance between the input variables and the class label.	[60]
Support vector machine (SVM)	Use various similarity measures of training samples (also known as kernel functions) to perform linear or non-linear separation of two classes.	[61]
Random forest (RF)	Construct an ensemble of decision trees to classify training samples, as well as to assess the variable importance in the classification.	[25]
Gradient boosting machine (GBM)	Build an ensemble of decision trees in a step-wise fashion using boosting and gradient descent algorithms.	[62]
Artificial neural network (ANN)	Construct multi-layered networks of neurons to learn highly non-linear functions that map the explanatory variables to the class label.	[63,64]
Genetic programming (GP)	Use natural evolution mechanisms to automatically search for the most relevant features and classification models.	[7,65]

Logistic regression (LR), partial least squares-discriminant analysis (PLS-DA) and recently support vector machine (SVM) are among the most currently-used statistical tools for metabolic marker discovery and predictive model training. This is likely a result of the extensive methodological research on these methods and the abundant availability of analysis packages in various programming languages including R and Python. The curation of large-volume, high-dimensional big data across multiple disciplines has been driving the methodological development of machine-learning algorithms [66,67]. Recent advanced learning algorithms have seen increasing applications to computer vision, natural language processing, pattern recognition, social sciences and medicine. Their potential has not been fully explored in the youngest member of omics, metabolomics, but they are very naturally suited to tackling the metabolic marker discovery task, which can be easily formulated as a typical feature selection and classification problem in machine learning. Here, we introduce three types of advanced learning algorithms, including ensemble learning, artificial neural networks, and genetic programming, and discuss their potential applications for metabolic marker discovery.

4.1. Ensemble Learning

In more complex datasets, the trained predictive models are often found to have high instability, a phenomena characterized by Breiman [68], where many distinct models involving different feature subsets can achieve comparably good training or testing prediction accuracy. Ensemble learning was proposed to address the issue by aggregating over a large set of competing base learners. Base learners are predictive models trained separately or sequentially and are often weighted based on their prediction performance. The final prediction is thus decided through majority voting for classification and averaging for regression tasks.

Base learners are usually generated from training data by one or multiple learning algorithms, resulting in a homogeneous or a heterogeneous ensemble. The learning algorithms can be any classification or regression algorithm. In the most common ensemble learning algorithms, a homogeneous ensemble is comprised of diverse classification and regression trees (CART) [69,70]. CART typically use internal nodes to represent features and use the best feature value cut-offs to spit samples into branches to reach leaf nodes representing the class labels (for classification) or target variable (for regression).

There are different mechanisms that can effectively construct the ensemble of base learners. The most commonly-used ones are bagging and boosting. Bagging is short for bootstrap aggregating and uses bootstrapped samples of the training data to train independent decision trees [69]. A bootstrapped training set is obtained by randomly sampling the training data with replacement. Therefore, a training sample may have multiple copies or not be present in a bootstrapped training set. Each bootstrapped training set is used independently to derive one decision tree. Bagging then decides the final prediction/regression outcome by majority voting or averaging these decision trees. The random forests (RF) algorithm is the most well-known ensemble learning method that employs bagging.

Boosting, on the other hand, constructs an ensemble of base learners by deriving a new learner through improving the previous one in a sequential fashion [71,72]. Boosting in fact refers to a class of such iterative ensemble techniques, among which gradient boosting machine (GBM) is a very popular and powerful boosting algorithm [73]. In GBM, at iteration i, a new decision tree approximation F_i is derived by adjusting the decision tree approximation F_{i-1} using the gradient of the loss function $\nabla L(y, F_{i-1})$, where y is the expected outcome.

Ensemble learning has been reported to have stronger generalization abilities in comparison to other machine-learning algorithms that use single predictive models [70]. The search for a single optimal model might be imperfect especially for complex, noisy and incomplete training data, and thus, using multiple separately trained or sequentially evolved models may give a good approximation of the true nature of the data. Ensemble learning has seen increasing applications to a variety of machine learning problems and could be a powerful analysis tool for metabolic marker discovery given the complexity, high-dimensionality and incompleteness of metabolomic data.

4.2. Artificial Neural Networks

Artificial neural networks (ANN) refer to a collection of learning algorithms inspired by the nervous systems and the human and animal brain. They loosely mimic how a large number of neurons process information and communicate in a highly-parallel style [74,75]. Each neuron is a computing unit that takes inputs (dendrites) from other neurons, and is "activated" if the aggregative inputs reach a certain condition. An activated neuron sends information (activation signals) to others through the connection (synapses) between its output (axon) and other neurons' inputs.

In a typical ANN, neurons are organized into several layers with the first being the input and last being the output layer. The input layer includes neurons that take explanatory inputs (e.g., metabolite concentrations), and the output layer gives one or multiple prediction outcome (e.g., the disease risk). The intermediate layers are called "hidden layers" that do not interact directly with the "environment" (i.e., either input or output) and are used to construct complex relationships that combine input

variables to compute the outcome. In general, more hidden layers lead to more complexity and more accurate modelling [67].

ANNs are represented by directed graphs where each node denotes a neuron and directed edges connecting neurons representing possible communication of activation signals. ANNs typically compute using a feed-forward mechanism, where neurons of a certain layer take outputs of the predecessor layer, compute the aggregative signals and use the activation functions to generate their own outputs that will be sent to the successor layer. ANNs are usually initialized randomly and then trained using an error-back-propagation mechanism. The current classification/regression error of an ANN is calculated as the absolute discrepancy of the expected and the computed outcome. The parameters of the ANN are then updated starting from the output layer to each predecessor layer based on the gradient of the cost function.

ANNs have seen much interest in research and applications in the past decades given their superior abilities of highly-accurate function approximations [76,77]. They have been exceptionally successful in tackling complex learning tasks in computer vision, natural language processing and recently recommender systems. There is a variety of network structures proposed in order to suit various learning problems, including convolutional neural networks [78], Boltzmann machine networks [79] and generative adversarial networks [80], just to name a few.

ANNs, especially deep ANNs that employ multiple hidden layers, can be powerful learning tools to construct highly-accurate predictive models for metabolomic research on complex diseases. However, they are often regarded as less "visible", or more difficult to interpret, especially for bioinformatics research, in terms of extracting mechanistic explanations from the learned complex models [27]. Research on designing ANN structures that are more amenable for mechanistic explanations is thus needed for a better utilization of this powerful and advanced machine-learning algorithm.

4.3. Genetic Programming

Many well-known machine-learning algorithms gradually adjust predictive models using the gradient of the cost function, typically defined as the prediction error (i.e., the distance of the expected and actual output of a model). Genetic programming (GP) improves randomly-generated predictive models using mechanisms borrowed from natural evolution. GP is located at the intersection of machine learning and evolutionary computing and is a lesser known, but potentially powerful algorithm for complex and incomplete modelling problems.

Evolutionary algorithms define a collection of meta-heuristic optimization and modelling algorithms inspired by natural evolution [81–83], and have been applied to bioinformatics on various optimization and modelling problems [23,24,84–90]. Evolutionary algorithms employ the trial-and-error problem-solving strategy and borrow ideas from how living organisms adapt through evolution. Various branches of evolutionary algorithms have been developed over the past decades, and they encode the solution to a problem differently. Specifically, genetic algorithm (GA) and evolution strategy (ES) solve optimization problems and typically represent candidate solutions using binary strings or real-valued vectors. As a machine-learning algorithm, GP solves modelling problems whose evolutionary individuals are regression or classification models, typically represented using expression trees or imperative programs [91,92].

Figure 2 shows the general workflow of evolutionary algorithms. An evolutionary algorithm maintains a population of diverse candidate solutions, or individuals, typically initialized randomly. These candidate solutions are compared to the desired outcome, and a fitness value can be calculated for each candidate solution based on how close it is to the desired outcome. Fitter candidate solutions will have higher probabilities of being selected for reproduction. During the reproduction process, slight and stochastic changes are applied to parent solutions, defined as mutations. Parents also swap portions of their encodings to form related, but distinctive offspring, defined as recombination. Fitter candidates survive to the next generation, and less fit ones die out. Then, through multiple

generations of selection, variation and reproduction, the selection criterion (the relative distance from the desired outcome) leads the population to produce increasingly fitter solutions.

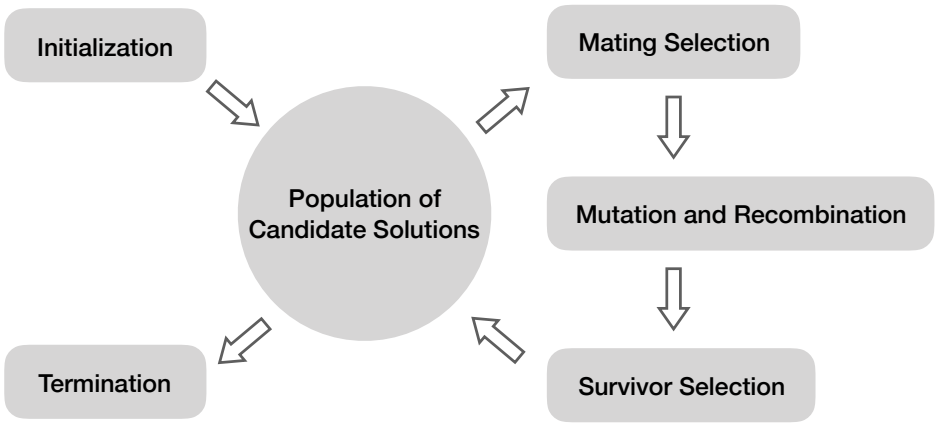

Figure 2. General workflow of an evolutionary algorithm. Typically, a population of candidate solutions to a problem is randomly initialized. Then, through the iterative evolution process, fitter solutions are more likely to be picked for reproduction and survival. Similar to living systems, random changes can be introduced to reproduction, such as mutation and recombination. The algorithm terminates once satisfactory solutions are observed or the computational limit (e.g., the maximal number of generations) has been reached.

In GP, candidate solutions are symbolic models, taking the form of a syntax tree (tree GP) or a symbolic computer program (linear GP) [82,92] that map the input variables to the output. Therefore, the fitness can be naturally characterized as the prediction accuracy of such a symbolic model. Mutations can be the alteration of elements of a symbolic model, and recombination swaps sections of two symbolic models in the hope of producing better child models. In the context of metabolic marker discovery, a candidate classification model of GP takes a set of input metabolite concentration values and outputs a prediction score of the disease risk.

GP can be a powerful addition to the metabolic marker discovery toolbox. Using arithmetic and branching operators to construct predictive models allows GP to approximate highly non-linear relationships that map the input metabolite concentrations to the disease risk assessment. Moreover, given the stochastic nature of evolution, metabolic feature selection is embedded and is coevolved as the model construction in the GP algorithm. Due to the symbolic forms, the evolved predictive models are also more amenable for interpretation, in comparison to "black-box" models trained by many machine-learning algorithms.

5. Conclusions

Metabolomics has an incredible amount of potential in human disease studies since today's most prominent diseases, including arthritis, diabetes, cardiovascular disease, obesity and Alzheimer's, have clear metabolic causes [93–96]. With rapidly-developing biological, analytical and computational technologies, the concentrations of hundreds of metabolites in a biological sample can be detected within minutes [97]. The comparison of their concentration levels in phenotypically-distinguished populations (e.g., diseased and healthy subjects) can help identify pathways and biological processes associated with a certain disease. This review fulfils three primary aims. First, a delineation of the general workflow of a metabolomics study from a biological question to model validation was provided. Following this, an overview of the historically- and currently-utilized computational methods for metabolic marker discovery across prominent complex diseases, such as AD, breast cancer and osteoarthritis, were discussed for the purpose of identifying notable trends in computational

technique use. Lastly, three rising areas of machine-learning methods, including ensemble learning, ANN and GP, were provided an introductory description and discussion regarding various advantages and disadvantages of usage.

Overall, there has been a clear shift in computational methodologies used by metabolomics researchers across complex disease areas. From univariate to multivariate analyses and linear to non-linear relationship modelling, the field of metabolomics is rapidly adopting the use of up-to-date machine learning algorithms to more appropriately match the intricate interplay of genetic, environmental and lifestyle factors, which converge on an estimated 100,000 downstream metabolites in the human body, in an attempt to better understand existing signalling theories and reduce significant pathway knowledge gaps that may be contributing to our lack of curative drugs for nearly all complex diseases. In the field of machine learning, improvements and innovations to these computational methods are published frequently. The importance of the choice of computational method should not be understated, as it can lead to dramatic improvements in biomarker panel performance in the clinic [6,7]. Lower performance can mean higher rates of false positives and negatives, leading to burdensome costs against the healthcare system and ultimately resulting in a reluctant phasing out of the test [40,41]. Having an understanding of existing statistical techniques, as well as new and upcoming computational methods to optimize the formation of accurate metabolic marker panels will be critical for knowledge translation efforts down the road. Therefore, this review hopes to provide researchers with an introduction to various methods for metabolomics research to advance the use of newer, potentially rich computational methods.

Funding: This research was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grant RGPIN-2016-04699 to T.H.

Acknowledgments: The authors are grateful to the Machine Intelligence and Biocomputing team at Memorial University of Newfoundland for helpful discussions on the covered topics, as well as Metabolites' reviewers for providing useful, constructive feedback on previous drafts.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

NMR	nuclear magnetic resonance
MS	mass spectrometry
LC-MS	liquid chromatography-mass spectrometry
PLS-DA	partial least squares-discriminant analysis
K-OPLS	kernel-based orthogonal projections to latent structures
PCA	principal component analysis
SVM	support vector machine
RF	random forest
GBM	gradient boosting machine
ANN	artificial neural networks
AUC	area under the curve
ANOVA	analysis of variance
Aβ	amyloid beta
AD	Alzheimer's disease
MCI	mild cognitive impairment
ADNI	Alzheimer's Disease Neuroimaging Initiative
BLSA	Baltimore Longitudinal Study of Aging
CSF	cerebrospinal fluid
GA	genetic algorithm
ES	evolution strategy
EA	evolutionary algorithm
GP	genetic programming
ATP	adenosine 5′-triphosphate

References

1. Patti, G.J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: The apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269. [CrossRef] [PubMed]
2. Kaddurah-Daouk, R.; Kristal, B.S.; Weinshilboum, R.M. Metabolomics: A Global Biochemical Approach to Drug Response and Disease. *Annu. Rev. Pharmacol. Toxicol.* **2008**, *48*, 653–683. [CrossRef]
3. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: The Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35*, D521–D526. [CrossRef] [PubMed]
4. Craig, J. Complex diseases: Research and applications. *Nat. Educ.* **2008**, *1*, 184.
5. Varma, V.R.; Oommen, A.M.; Varma, S.; Casanova, R.; An, Y.; Andrews, R.M.; O'Brien, R.; Pletnikova, O.; Troncoso, J.C.; Toledo, J.; et al. Brain and blood metabolite signatures of pathology and progression in Alzheimer disease: A targeted metabolomics study. *PLoS Med.* **2018**, *15*, e1002482. [CrossRef] [PubMed]
6. Alakwaa, F.M.; Chaudhary, K.; Garmire, L.X. Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data. *J. Proteome Res.* **2018**, *17*, 337–347. [CrossRef]
7. Hu, T.; Oksanen, K.; Zhang, W.; Randell, E.; Furey, A.; Sun, G.; Zhai, G. An evolutioanry learning and network approach to identifying key metabolites for osteoarthritis. *PLoS Comput. Biol.* **2018**, *14*, e1005986. [CrossRef]
8. Alonso, A.; Marsal, S.; Julià, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3*, 23. [CrossRef] [PubMed]
9. Xia, J.; Broadhurst, D.I.; Wilson, M.; Wishart, D.S. Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics* **2013**, *9*, 280–299. [CrossRef] [PubMed]
10. Fiehn, O.; Robertson, D.; Griffin, J.; van der Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L.W.; Goodacre, R.; Hardy, N.W.; Taylor, C.; et al. The metabolomics standards initiative (MSI). *Metabolomics* **2007**, *3*, 175–178. [CrossRef]
11. Spicer, R.A.; Salek, R.; Steinbeck, C. A decade after the metabolomics standards initiative it's time for a revision. *Sci. Data* **2017**, *4*, 170138. [CrossRef] [PubMed]
12. Broadhurst, D.I.; Kell, D.B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2006**, *2*, 171–196. [CrossRef]
13. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef] [PubMed]
14. Tan, S.Z.; Begley, P.; Mullard, G.; Hollywood, K.A.; Bishop, P.N. Introduction to metabolomics and its applications in ophthalmology. *Eye (Lond. Engl.)* **2016**, *30*, 773–783. [CrossRef] [PubMed]
15. Rockel, J.S.; Zhang, W.; Shestopaloff, K.; Likhodii, S.; Sun, G.; Furey, A.; Randell, E.; Sundararajan, K.; Gandhi, R.; Zhai, G.; et al. A classification modelling approach for determining metabolite signatures in osteoarthritis. *PLoS ONE* **2018**, *13*, e0199618. [CrossRef] [PubMed]
16. Liu, J.; Semiz, S.; van der Lee, S.J.; van der Spek, A.; Verhoeven, A.; van Klinken, J.B.; Sijbrands, E.; Harms, A.C.; Hankemeier, T.; van Dijk, K.W.; et al. Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study. *Metab. Off. J. Metab. Soc.* **2017**, *13*, 104. [CrossRef] [PubMed]
17. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Anal. Chem.* **2006**, *78*, 4281–4290. [CrossRef] [PubMed]
18. Bolstad, B.M.; Irizarry, R.A.; Åstrand, M.; Speed, T.P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics* **2003**, *19*, 185–193. [CrossRef] [PubMed]
19. Hoefer, I.E.; Steffens, S.; Ala-Korpela, M.; Back, M.; Badimon, L.; Bochaton-Piallat, M.L.; Boulanger, C.M.; Caligiuri, G.; Dimmeler, S.; Egido, J.; et al. Novel methodologies for biomarker discovery in atherosclerosis. *Eur. Heart J.* **2015**, *36*, 2635–2642. [CrossRef] [PubMed]
20. Kaddurah-Daouk, R.; Weinshilboum, R. Metabolomic signatures for drug response phenotypes: Pharmacometabolomics enables precision medicine. *Clin. Pharmacol. Ther.* **2015**, *98*, 71–75. [CrossRef] [PubMed]

21. Zhang, A.; Sun, H.; Yan, G.; Wang, P.; Wang, X. Mass spectrometry-based metabolomics: Applications to biomarker and metabolic pathway research. *Biomed. Chromatogr.* **2016**, *30*, 7–12. [CrossRef] [PubMed]
22. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
23. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
24. Cavill, R.; Keun, H.C.; Holmes, E.; Lindon, J.C.; Nicholson, J.K.; Ebbels, T.M.D. Genetic algorithms for simultaneous variable and sample selection in metabolomics. *Bioinformatics* **2009**, *25*, 112–118. [CrossRef] [PubMed]
25. Grissa, D.; Petera, M.; Brandolini, M.; Napoli, A.; Comte, B.; Pujos-Guillot, E. Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. *Front. Mol. Biosci.* **2016**, *8*, 30. [CrossRef]
26. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
27. Yu, M.K.; Ma, J.; Fisher, J.; Kreisberg, J.F.; Raphael, B.J.; Ideker, T. Visible machine learning for biomedicine. *Cell* **2018**, *173*, 1562–1565. [CrossRef]
28. Choromanska, A.; Henaff, M.; Mathieu, M.; Ben Arous, G.; LeCun, Y. The Loss Surfaces of Multilayer Networks. *arXiv* **2014**, arXiv:1412.0233.
29. Bird, T.D. Alzheimer Disease Overview. GeneReviews®. Available online: https://www.ncbi.nlm.nih.gov/books/NBK1161 (accessed on 4 April 2019).
30. Hebert, L.E.; Weuve, J.; Scherr, P.A.; Evans, D.A. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology* **2013**, *80*, 1778–1783. [CrossRef]
31. Masters, C.L.; Bateman, R.; Blennow, K.; Rowe, C.C.; Sperling, R.A.; Cummings, J.L. Alzheimer's disease. *Nat. Rev. Dis. Prim.* **2015**, *1*, 15056. [CrossRef]
32. Mehta, D.; Jackson, R.; Paul, G.; Shi, J.; Sabbagh, M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010–2015. *Expert Opin. Investig. Drugs* **2017**, *26*, 735–739. [CrossRef]
33. Cummings, J. Lessons Learned from Alzheimer Disease: Clinical Trials with Negative Outcomes. *Clin. Transl. Sci.* **2018**, *11*, 147–152. [CrossRef]
34. Han, X.; Holtzman, D.M.; McKeel, D.W., Jr.; Kelley, J.; Morris, J.C. Substantial sulfatide deficiency and ceramide elevation in very early Alzheimer's disease: Potential role in disease pathogenesis. *J. Neurochem.* **2002**, *82*, 809–818. [CrossRef] [PubMed]
35. Han, X.; Rozen, S.; Boyle, S.H.; Hellegers, C.; Cheng, H.; Burke, J.R.; Welsh-Bohmer, K.A.; Doraiswamy, P.M.; Kaddurah-Daouk, R. Metabolomics in early Alzheimer's disease: Identification of altered plasma sphingolipidome using shotgun lipidomics. *PLoS ONE* **2011**, *6*, e21643. [CrossRef] [PubMed]
36. Wang, G.; Zhou, Y.; Huang, F.J.; Tang, H.D.; Xu, X.H.; Liu, J.J.; Wang, Y.; Deng, Y.L.; Ren, R.J.; Xu, W.; et al. Plasma Metabolite Profiles of Alzheimer's Disease and Mild Cognitive Impairment. *J. Proteome Res.* **2014**, *13*, 2649–2658. [CrossRef]
37. Sabbagh, M.N.; Lue, L.F.; Fayard, D.; Shi, J. Increasing Precision of Clinical Diagnosis of Alzheimer's Disease Using a Combined Algorithm Incorporating Clinical and Novel Biomarker Data. *Neurol. Ther.* **2017**, *6*, 83–95. [CrossRef]
38. Siegel, R.; Miller, K.; Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **2018**, *68*, 7–30. [CrossRef]
39. Schifmann, J.; Fisher, P.; Gibbs, P. Early detection of cancer: Past, present, and future. *Am. Soc. Clin. Oncol. Educ. Book* **2015**, 57–65. [CrossRef] [PubMed]
40. Catalona, W.; Richie, J.; Ahmann, F.; Hudson, M.; Scardino, P.; Flanigan, R.; DeKernion, J.; Ratliff, T.; Kavoussi, L.; Dalkin, B.; et al. Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: Results of a multicenter clinical trial of 6630 men. *J. Urol.* **1994**, *151*, 1283–1290. [CrossRef]
41. Harris, L.; Fritsche, H.; Mennel, R.; Norton, L.; Ravdin, P.; Taube, S.; Somerfield, M.R.; Hayes, D.F.; Bast, R.C. American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer. *J. Clin. Oncol.* **2007**, *25*, 5287–5312. [CrossRef]

42. Vander Heiden, M.G.; Cantley, L.C.; Thompson, C.B. Understanding the Warburg effect: The metabolic requirements of cell proliferation. *Science (N. Y.)* **2009**, *324*, 1029–1033. [CrossRef] [PubMed]
43. Fouad, Y.A.; Aanei, C. Revisiting the hallmarks of cancer. *Am. J. Cancer Res.* **2017**, *7*, 1016–1036. [PubMed]
44. Torre, L.A.; Bray, F.; Siegel, R.L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. Global cancer statistics, 2012. *CA Cancer J. Clin.* **2015**, *65*, 87–108. [CrossRef]
45. Althuis, M.D.; Dozier, J.M.; Anderson, W.F.; Devesa, S.S.; Brinton, L.A. Global trends in breast cancer incidence and mortality 1973–1997. *Int. J. Epidemiol.* **2005**, *34*, 405–412. [CrossRef] [PubMed]
46. Chan, C.H.F.; Coopey, S.B.; Freer, P.E.; Hughes, K.S. False-negative rate of combined mammography and ultrasound for women with palpable breast masses. *Breast Cancer Res. Treat.* **2015**, *153*, 699–702. [CrossRef]
47. Polanski, M.; Anderson, N.L. A list of candidate cancer biomarkers for targeted proteomics. *Biomark. Insights* **2007**, *1*, 1–48. [CrossRef] [PubMed]
48. Hilvo, M.; Denkert, C.; Lehtinen, L.; Müller, B.; Brockmöller, S.; Seppänen-Laakso, T.; Budczies, J.; Bucher, E.; Yetukuri, L.; Castillo, S.; et al. Novel Theranostic Opportunities Offered by Characterization of Altered Membrane Lipid Metabolism in Breast Cancer Progression. *Cancer Res.* **2011**, *71*, 3236. [CrossRef] [PubMed]
49. Huang, J.H.; Fu, L.; Li, B.; Xie, H.L.; Zhang, X.; Chen, Y.; Qin, Y.; Wang, Y.; Zhang, S.; Huang, H.; et al. Distinguishing the serum metabolite profiles differences in breast cancer by gas chromatography mass spectrometry and random forest method. *RSC Adv.* **2015**, *5*, 58952–58958. [CrossRef]
50. WHO Scientific Group on the Burden of Musculoskeletal Conditions at the Start of the New Millennium. *The Burden of Musculoskeletal Conditions at the Start of the New Millennium*; World Health Organization Technical Report Series; World Health Organization: Geneva, Switzerland, 2003; Volume 919, pp. 1–218.
51. Peace, C.; Carr, A.; Loughlin, J. Recent advances in the genetic investigation of osteoarthritis. *Trends Mol. Med.* **2005**, *11*, 186–191.
52. Wallace, I.J.; Worthington, S.; Felson, D.T.; Jurmain, R.D.; Wren, K.T.; Maijanen, H.; Woods, R.J.; Lieberman, D.E. Knee osteoarthritis has doubled in prevalence since the mid-20th century. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 9332–9336. [CrossRef]
53. Loeser, R.F.; Collins, J.A.; Diekman, B.O. Ageing and the pathogenesis of osteoarthritis. *Nat. Rev. Rheumatol.* **2016**, *12*, 412–420. [CrossRef]
54. Zhang, Y.; Jordan, J.M. Epidemiology of osteoarthritis. *Clin. Geriatr. Med.* **2010**, *26*, 355–369. [CrossRef]
55. Valdes, A.M.; Meulenbelt, I.; Chassaing, E.; Arden, N.K.; Bierma-Zeinstra, S.; Hart, D.; Hofman, A.; Karsdal, M.; Kloppenburg, M.; Kroon, H.M.; et al. Large scale meta-analysis of urinary C-terminal telopeptide, serum cartilage oligomeric protein and matrix metalloprotease degraded type II collagen and their role in prevalence, incidence and progression of osteoarthritis. *Osteoarthr. Cartil.* **2014**, *22*, 683–689. [CrossRef] [PubMed]
56. Zhang, W.; Likhodii, S.; Zhang, Y.; Aref-Eshghi, E.; Harper, P.E.; Randell, E.; Green, R.; Martin, G.; Furey, A.; Sun, G.; et al. Classification of osteoarthritis phenotypes by metabolomics analysis. *BMJ Open* **2014**, *4*, e006286. [CrossRef] [PubMed]
57. Zhang, Q.; Li, H.; Zhang, Z.; Yang, F.; Chen, J. Serum metabolites as potential biomarkers for diagnosis of knee osteoarthritis. *Dis. Mark.* **2015**, *2015*, 684794. [CrossRef]
58. Hu, T.; Oksanen, K.; Zhang, W.; Randell, E.; Furey, A.; Zhai, G. Analyzing feature importance for metabolomics using genetic programming. In Proceedings of the 21st European Conference (EuroGP 2018), Parma, Italy, 4–6 April 2018; Springer: Cham, Switzerland, 2018.
59. Peddinti, G.; Cobb, J.; Yengo, L.; Froguel, P.; Kravić, J.; Balkau, B.; Tuomi, T.; Aittokallio, T.; Groop, L. Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia* **2017**, *60*, 1740–1750. [CrossRef] [PubMed]
60. Gromski, P.S.; Muhamadali, H.; Ellis, D.I.; Xu, Y.; Correa, E.; Turner, M.L.; Goodacre, R. A tutorial review: Metabolomics and partial least squares-discriminant analysis—A marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* **2015**, *879*, 10–23. [CrossRef] [PubMed]
61. Mahadevan, S.; Shah, S.L.; Marrie, T.J.; Slupsky, C.M. Analysis of metabolomic data using support vector machines. *Anal. Chem.* **2008**, *80*, 7562–7570. [CrossRef] [PubMed]
62. Determan, C.E., Jr. Optimal algorithm for metabolomics classification and feature selection varies by dataset. *Int. J. Biol.* **2015**, *7*, 100–115.

63. Brougham, D.F.; Ivanova, G.; Gottschalk, M.; Collins, D.M.; Eustace, A.J.; O'Connor, R.; Havel, J. Artificial neural networks for classification in metabolomic studies of whole cells using 1H nuclear magnetic resonance. *J. Biomed. Biotechnol.* **2001**, *2011*, I58094.
64. Hall, L.M.; Hill, D.W.; Menikarachchi, L.C.; Chen, M.H.; Hall, L.H.; Grant, D.F. Optimizing artificial neural network models for metabolomics and systems biology: An example using HPLC retention index data. *Bioanalysis* **2015**, *7*, 939–955. [CrossRef]
65. Kenny, L.C.; Dunn, W.B.; Ellis, D.I.; Myers, J.; Baker, P.N.; The GOPEC Consortium; Kell, D.B. Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics* **2005**, *1*, 227–234. [CrossRef]
66. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **2015**, *521*, 452–459. [CrossRef] [PubMed]
67. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
68. Breiman, L. Statistical modelling: The two cultures. *Stat. Sci.* **2001**, *16*, 199–215. [CrossRef]
69. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
70. Dietterich, T.G. Ensemble methods in machine learning. *Mult. Classif. Syst.* **2000**, *1857*, 1–15.
71. Schapire, R.E. The strength of weak learnability. *Mach. Learn.* **1990**, *5*, 197–227. [CrossRef]
72. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
73. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
74. Anderson, J.A.; Rosenfeld, E. *Neurocomputing: Foundations of Research*; MIT Press: Cambridge, MA, USA, 1988.
75. Wasserman, P.D. *Neural Computing: Theory and Practice*; Van Nostrand Reinhold: New York, NY, USA, 1989.
76. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
77. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
78. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*; MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
79. Ackley, D.H.; Hinton, G.E.; Sejnowski, T.J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **1985**, *9*, 147–169. [CrossRef]
80. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
81. Holland, J. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, USA, 1975.
82. Koza, J.R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*; MIT Press: Cambridge, MA, USA, 1992.
83. Back, T.; Fogel, D.B.; Michalewicz, Z. *Handbook of Evolutionary Computation*; Oxford University Press: Oxford, UK, 1997.
84. Hu, T.; Banzhaf, W. Evolvability and speed of evolutionary algorithms in light of recent developments in Biology. *J. Artif. Evol. Appl.* **2010**, *2010*. [CrossRef]
85. Goodacre, R.; Kell, D.B. Evolutionary computation for the interpretation of metabolomic data. In *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*; Harrigan, G.G., Goodacre, H., Eds.; Kluwer Academic PUblishers: Dordrecht, The Netherlands, 2003; Chapter 13, pp. 239–256.
86. Goodacre, R. Making sense of the metabolome using evolutionary computation: Seeing the wood with the trees. *J. Exp. Bot.* **2005**, *56*, 245–254. [CrossRef] [PubMed]
87. Pal, S.K.; Bandyopadhyay, S.; Ray, S.S. Evolutionary computation in bioinformatics: A review. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2006**, *36*, 601–615. [CrossRef]
88. Muni, D.P.; Pal, N.R.; Das, J. Genetic programming for simultaneous feature selection and classifier design. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2006**, *36*, 106–117. [CrossRef]
89. Hageman, J.A.; van den Berg, R.A.; Westerhuis, J.A.; van der Werf, M.J.; Smilde, A.K. Genetic algorithm based two-mode clustering of metabolomics data. *Metabolomics* **2008**, *4*, 141–149. [CrossRef]

90. Devi, R.V.; Sathya, S.S.; Coumar, M.S. Evolutionary algorithms for de novo drug design—A survey. *Appl. Soft Comput.* **2015**, *27*, 543–552. [CrossRef]
91. Banzhaf, W.; Nordin, P.; Keller, R.E.; Francone, F.D. *Genetic Programming: An Introduction*; Morgan Kaufmann: San Francisco, CA, USA, 1998.
92. Brameier, M.F.; Banzhaf, W. *Linear Genetic Programming*; Springer: Berlin, Germany, 2007.
93. Fiehn, O. Metabolomics—The link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–171. [CrossRef]
94. Zhang, W.; Sun, G.; Likhodii, S.; Liu, M.; Aref-Eshghi, E.; Harper, P.E.; Martin, G.; Furey, A.; Green, R.; Randell, E.; et al. Metabolomic analysis of human plasma reveals that arginine is depleted in knee osteoarthritis patients. *Osteoarthr. Cartil.* **2016**, *24*, 827–834. [CrossRef]
95. Wishart, D.S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* **2016**, *15*, 473–484. [CrossRef]
96. Wilkins, J.M.; Trushina, E. Application of metabolomics in Alzheimer's disease. *Front. Neurol.* **2018**, *8*, 719. [CrossRef]
97. Nandania, J.; Peddinti, G.; Pessia, A.; Kokkonen, M.; Velagapudi, V. Validation and Automation of a High-Throughput Multitargeted Method for Semiquantification of Endogenous Metabolites from Different Biological Matrices Using Tandem Mass Spectrometry. *Metabolites* **2018**, *8*, 44. [CrossRef] [PubMed]