




Analyzing Feature Importance for Metabolomics Using Genetic Programming

Ting Hu¹(✉) , Karoliina Oksanen¹, Weidong Zhang^{2,3}, Edward Randell²,
Andrew Furey², and Guangju Zhai²

¹ Department of Computer Science, Memorial University,
St. John's, NL A1B 3X5, Canada
ting.hu@mun.ca

² Faculty of Medicine, Memorial University, St. John's, NL A1B 3V6, Canada

³ School of Pharmaceutical Sciences, Jilin University, Jilin, Changchun, China

Abstract. The emerging and fast-developing field of metabolomics examines the abundance of small-molecule metabolites in body fluids to study the cellular processes related to how the human body responds to genetic and environmental perturbations. Considering the complexity of metabolism, metabolites and their represented cellular processes can correlate and synergistically contribute to a phenotypic status. Genetic programming (GP) provides advanced analytical instruments for the investigation of multifactorial causes of metabolic diseases. In this article, we analyzed a population-based metabolomics dataset on osteoarthritis (OA) and developed a Linear GP (LGP) algorithm to search classification models that can best predict the disease outcome, as well as to identify the most important metabolic markers associated with the disease. The LGP algorithm was able to evolve prediction models with high accuracies especially with a more focused search using a reduced feature set that only includes potentially relevant metabolites. We also identified a set of key metabolic markers that may improve our understanding of the biochemistry and pathogenesis of the disease.

Keywords: Metabolomics · Osteoarthritis · Biomarker discovery
Genetic programming · Classification

1 Introduction

Systems biology is an emerging research field that takes a holistic approach to modeling complex biological systems rather than examining different levels of biological systems separately [1–3]. It requires collaborative efforts from disciplines including biomedicine, statistics, and computer science. Systems biology approaches embrace the complexity of biological systems and focus on modeling the interactions among multiple components including genome, transcriptome,

proteome, and metabolome [4–6]. By integrating a variety of “omics” data, systems biology for human disease studies aims at better understanding the pathogenesis of common diseases, discovering biomarkers that can help predict early disease onset, progression, and severity, and identifying new drug targets [7,8].

Integrative data analysis and mining for systems biology often include hundreds to thousands of variables such as genes, proteins, and metabolites [9], in order to find the most relevant biomarkers that can explain a specific phenotype or disease. Most conventional tools adopt a univariate analysis strategy and examine one variable at a time on its individual association with the disease. This may overlook the intertwined relationships among multiple variables that contribute to the disease. Thus, retooling for systems biology is needed such that a large set of variables can be analyzed simultaneously on their synergistic effects [10,11]. However, the high dimensionality has imposed both methodological and computational challenges since learning algorithms that can model the complex non-linear relationships of multiple variables are yet to be explored, and searching combinations of variables becomes prohibitive as the search space grows exponentially with the number of variables.

Machine learning and heuristic search algorithms, including principal component analysis [12], artificial neural networks [13], and random forest [14], have seen increasing and successful applications in omics data mining for biomarker discovery. However, despite a few attempts [15,16], genetic programming, as a powerful learning and modeling algorithm, has not caught up with other comparable algorithms in wide applications.

Genetic programming (GP) holds great potentials for systems biology research. First, it can construct highly non-linear models of multiple variables (features) that can best predict a phenotypic or disease outcome using arithmetic functions, Boolean functions, and conditional statements. Second, the selection of relevant features in a model classifier is achieved automatically in GP. This feature selection process is embedded in model construction such that the inclusion of a feature is decided based on the classification performance of the model. Such an automatic and embedded feature selection mechanism distinguishes GP from many approaches that select features and construct classification models in separate steps. Third, the stochastic population-based search property of evolutionary algorithms allows to generate multiple best classification models. This provides a diverse set of classification models for subsequent interpretation and feature importance analysis.

In this study, we use a GP algorithm, specifically a Linear GP representation, to train classification models and to identify key biomarkers for metabolomics, in order to demonstrate the power of GP in the coming era of systems biology and big biomedical data research.

Recent developments in the field of metabolomics provide an array of new tools for the study of human diseases. A large number of small-molecule metabolites from body fluids or tissues can be quantitatively detected simultaneously, which promises an immense potential for early diagnosis, therapy monitoring and understanding the pathogenesis of complex diseases [17]. Metabolites are

intermediate and end products of various cellular processes and their levels of concentration serve as a good indicator of a sequence of biological systems in response to genetic and environmental influences. This can, in turn, help us better understand the diseases and develop new drug treatments.

We use population-based metabolomics data where two phenotypically distinguished individuals, i.e., diseased cases and healthy controls, are recruited and their blood samples are collected to measure the concentration levels of a variety of metabolites. Classification models are then evolved and trained using GP algorithm. We adopt a two-round design where GP uses the full set of metabolites in the initial round of model exploration and selects a subset of potentially more relevant metabolites for the second round of more focused search. The importance of metabolites in terms of their contribution to the disease is then assessed based on their occurrence frequencies in the final best classification models.

2 Methods

2.1 Metabolomics Data on Osteoarthritis

Osteoarthritis (OA) is a slowly progressive joint disease and is the most common form of arthritis. It occurs when the protective cartilage on the ends of bones breaks down often because of mechanical stress or biochemical alterations. It causes a substantial morbidity and disability in the elderly populations, and imposes a great economic burden on our society [18, 19]. Despite high prevalence and societal impact, there is no medication that can cure it, or reverse or halt the disease progression, partly because its pathogenesis is still unclear and there is no reliable method that can be used for early OA diagnosis.

In this study, we used a OA metabolomics dataset from the Newfoundland Osteoarthritis Study (NFOAS) [20, 21]. The goal of the NFOAS is to identify novel genetic, epigenetic, and biochemical markers for OA, in order to better understand the diseases and to develop new drug treatment. In the NFOAS, knee OA patients who underwent a total knee replacement surgery due to primary OA were recruited. Healthy controls were selected from volunteering participants.

Both cases and controls were from the same source population. Knee OA diagnosis was made based on the American College of Rheumatology clinical criteria for the classification of idiopathic OA of the knee [22] and the judgment of the attending orthopedic surgeons. Controls were individuals without self-reported family doctor diagnosed knee OA based on their medical information collected by a self-administered questionnaire. A total number of 153 OA cases and 236 healthy controls were collected.

Blood samples were collected after at least 8 hours of fasting and plasma was separated from blood using the standard protocol. Metabolic profiling was performed on plasma using the Waters XEVO TQ MS system (Waters Limited, Mississauga, Ontario, Canada) coupled with Biocrates AbsoluteIDQ p180 kit, which measures 186 metabolites including 90 glycerophospholipids, 40 acylcarnitines (1 free carnitine), 21 amino acids, 19 biogenic amines, 15 sphingolipids and 1 hexose (above 90 percent is glucose). The details of the 186 metabolites

and the metabolic profiling method were described in a previous publication [23]. Over 90% of the metabolites (167/186) were successfully determined in each sample.

The study protocol was approved by the Health Research Ethics Authority (HREA) of Newfoundland and Labrador with reference number 11.311 and a written consent was obtained from all the participants.

We followed a two-stage design and divided the samples randomly into *discovery* and *replication* datasets, such that our genetic programming algorithm can be applied separately to the two datasets and only the key features (metabolites) successfully replicated were reported. Since samples were collected and their metabolite concentrations were measured in various batches, certain biases can exist when samples from different batches were compared. We performed batch corrections to remove such biases by multiplying each metabolite concentration value by the ratio of the overall mean and the batch mean for that metabolite. In addition, age and BMI are known factors correlated with OA. Therefore, the residual of a linear regression using attributes age and BMI was applied to remove any partial correlations as a result of those two factors, and to adjust the data for subsequent analysis. Finally, each metabolite concentration value was normalized to zero mean and unit variance across the population.

2.2 Linear Genetic Programming Algorithm

Linear genetic programming (LGP) encodes evolutionary individuals as imperative programs that are executed sequentially [24]. Although LGP follows a linear instructional structure, it is very powerful and capable of modeling complex nonlinear relationships among multiple attributes. Comparing to the more traditional representation of trees, such an instructional structure of LGP enables fast execution and thus speedy fitness evaluation. Therefore, LGP has gained increasing popularity being applied to a variety of modeling and classification problems [25–27].

In the current study, an instruction of an LGP program can be either an assignment statement or a conditional statement. An assignment statement manipulates values stored in calculation registers by applying arithmetic operations such as addition, subtraction, multiplication, division, and the exponential function. We use `if-then` statements to change the flow of program execution by skipping one subsequent instruction when the condition in the `if` statement is false.

Feature registers contain input values of corresponding variables from data samples, and calculation registers are used to enhance the computational capacity of LGP programs. A feature register can only serve as an operand on the right-hand side of an assignment statement, while a calculation register can be used as an operand or a return on the left-hand side of an assignment statement. The calculation register `r[0]` is designated as the output register, and its final stored value is the outcome of the entire program. Since we consider a classification problem in the current study, the Sigmoid function will be applied to `r[0]`.

If $S(r[0])$ is greater than or equal to 0.5, the sample is predicted as diseased (class one), otherwise, the sample is predicted as healthy (class zero).

Therefore, an LGP program represents a classification model that takes a data sample with a set of feature values (metabolite concentration levels) as input, and outputs the predicted class status (diseased or healthy) of this sample. An example LGP program with eight instructions is given below.

```

if r[1] > r[5]
    then r[0] = r[7] + 5
r[4] = r[2] / r[0]
if r[0] > 4
    then if r[3] < 10
        then r[6] = r[3] - r[5]
r[2] = r[5] * r[5]
r[0] = r[2] + r[7]

```

At the initial generation, a population of LGP programs is generated randomly. The fitness of each program is evaluated using mean classification error (MCE), computed as the average number of incorrectly classified training samples. A set of programs are chosen as parents based on their fitness, and variation operators, including mutation and recombination, are applied to them. A micro mutation alters an element of a randomly picked instruction, i.e., replacing a return or an operand register by a randomly generated one or replacing the operator. A macro mutation deletes a randomly chosen instruction or inserts a randomly generated instruction. Recombination swaps segments of instructions of two parent programs. Survival selection picks fitter programs to form the population for the next generation. Such an evolution process iterates for a certain number of generations, and the program with the lowest MCE at the end is output as the final best model of a run.

In our study, the LGP algorithm is implemented using the Julia programming language [28]. The main parameters used in the implementation are shown in Table 1. A five-fold cross-validation strategy was used to prevent overfitting. That is, the data samples are randomly divided into five partitions, and each partition serves as the testing set once while the remaining four partitions are input to the LGP algorithm as the training set. Therefore, for each implementation, the algorithm produces five best classification models based on the five testing sets.

2.3 Full vs. Focused Feature Analysis

The goal of our metabolomics study is to identify key metabolites that can best explain the phenotypic class, i.e., diseased or healthy. The importance of a metabolite (feature), can be assessed by computing its occurrence frequency in the best classification models found by the LGP algorithm. Such an occurrence frequency measures how often a feature appears in the final outcome model of an LGP run, and thus reflects its contribution to the correct classification of the disease status.

Table 1. LGP parameter configurations for classification on metabolomics data.

Fitness function	Mean classification error (MCE)
Program length	[1, 500]
Number of calculation registers	150
Operator set	{+, -, ×, ÷, x^y , if <, if >}
Constant set	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
Population size	500
Mutation rate	0.1
Mutation operators	Micro and macro to effective instructions
Crossover rate	0.9
Parent selection	Tournament with size 16
Survival selection	Truncation
Number of generations	500
Number of runs	200

For the first round of analysis, the LGP algorithm is run using the full feature set of 167 metabolites on both the discovery and replication datasets using 200 distinct seed values for the random number generator. Each run gives five different best classification models as a result of the five-fold cross-validation. Therefore, our implementation produces a total of 1000 best classification models.

We investigate the resulting classification models by calculating various statistics of the fitness (MCE) values, sensitivity, specificity and area under the curve (AUC) as computed on the testing fold for each run. In addition, we inspect the models by counting how often each of the 167 metabolites appears as a predictive variable in the set of 1000 best models.

Note that although a total of 167 metabolites are measured in the OA metabolomics data, not all of them are relevant to the disease. In machine learning, removing irrelevant features can speed up the training process and improve the prediction accuracy of the models [29]. Therefore, we perform the second round of analysis by only using a focused subset of metabolites. The focused subset of metabolites is defined as the metabolites that have occurrence frequencies higher than the average among all 167 metabolites. We re-run the LGP algorithm using such focused feature sets on both the discovery and replication datasets, and investigate if reducing the number of features can improve the prediction performance.

3 Results

3.1 Best Models Found Using Full Feature Set

First, we investigate the 1000 best models found by the LGP algorithm on the discovery dataset using the full set of 167 metabolites. The statistics of the

Table 2. Statistics of the classification performance of the 1000 best models (discovery, full feature set).

	MCE	Sensitivity	Specificity	AUC
Mean	0.367	0.684	0.584	0.663
Median	0.367	0.667	0.600	0.667
Min	0.067	0.200	0.200	0.320
Max	0.667	1.000	0.933	0.947
Std dev	0.095	0.146	0.142	0.110
5% confidence	0.181	0.398	0.305	0.447
95% confidence	0.553	0.970	0.862	0.879

Table 3. Statistics of the classification performance of the 1000 best models (replication, full feature set).

	MCE	Sensitivity	Specificity	AUC
Mean	0.357	0.685	0.601	0.664
Median	0.367	0.667	0.600	0.664
Min	0.100	0.267	0.067	0.309
Max	0.667	1.000	1.000	0.960
Std dev	0.103	0.140	0.169	0.118
5% confidence	0.156	0.411	0.271	0.432
95% confidence	0.558	0.958	0.932	0.895

classification performance of those 1000 best models are shown in Table 2. The best classifier can achieve a mean classification error (MCE) as low as 0.067, and the area under the curve (AUC) as 0.947. This demonstrates the effectiveness of using the LGP algorithm to train a classifier for metabolomics studies.

We look at the distributions of the fitness (MCE) and the number of effective features of those 1000 best models (Fig. 1). The majority of those 1000 best models have an MCE in the range of [0.3, 0.5]. A feature is effective if it takes a role modifying the value stored in the output register when the LGP program, i.e., classification model, is executed to make a prediction. Although any subsets of those 167 metabolites can be chosen by a classification model, the LGP algorithm selects the most relevant features as the result of the evolutionary learning process. The majority of those 1000 best models have between 25 and 40 effective features. Figure 2 shows that the fitness and the number of effective features are not correlated (Spearman’s correlation test $\rho = 0.044$ with a significance level $p = 0.16$).

The same analysis is then repeated on the replication dataset, and the statistics of the classification performance of the 1000 best models found by LGP

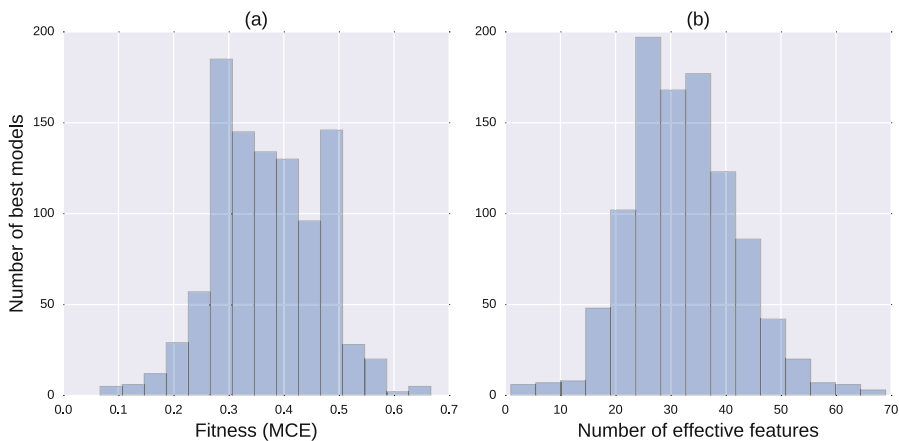


Fig. 1. Distributions of (a) the fitness and (b) the number of effective features for the 1000 best models (discovery, full feature set).

are shown in Table 3. We see that using the discovery and replication datasets achieve comparable classification performance.

3.2 Best Models Found Using Focused Feature Sets

For the second round of analysis, we reduce the feature set and only provide a more relevant subset of features to the LGP algorithm in order to perform a more focused classification model construction. In our study, the relevance, or importance, of a metabolite is assessed using its occurrence frequency in the 1000 best models, i.e., the number of times a metabolite appears in the 1000 best models as an effective feature. We follow the intuition that if a metabolite appears often in the evolved best models, it may play an important role explaining the disease.

Figure 3(a) shows the distribution of metabolite occurrence frequency in the 1000 best models using the discovery dataset. The majority of metabolites have occurrence frequencies between 170 and 220. The mean of the distribution is 193.562, and we use that as the threshold to select the focused feature set. That is, the focused feature set only includes 75 metabolites that have occurrence frequencies higher than or equal to the average value of 193.562. The distribution of metabolite occurrence frequency in the best models using the replication dataset is shown in Fig. 4(a). The mean of the distribution is 191.898, and similarly, we use it as the threshold to select the replication focused feature set with 60 metabolites for the second round of analysis.

The statistics of the classification performance using focused feature sets are shown in Tables 4 and 5 for the discovery and replication datasets respectively. Comparing to Tables 2 and 3, we can see that the classification performance is improved by examining all statistics. Specifically, the average MCE is reduced from 0.367 to 0.317 and the average AUC is improved from 0.663 to 0.714 for

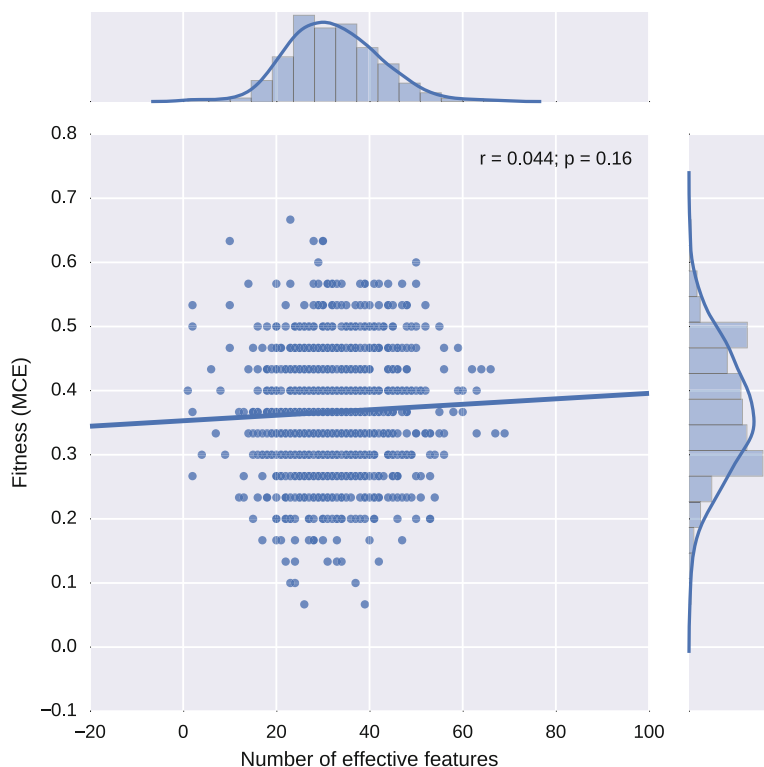


Fig. 2. Correlation of the fitness and the number of effective features in the best prediction models (discovery, full feature set). Each data point represents one of the 1000 best classification models found by LGP. The solid line provides a visual guide on the correlation between the fitness and the number of effective features.

discovery dataset, and from 0.357 to 0.286 and from 0.664 to 0.740 for replication dataset respectively. The improvement of the classification performance by reducing the feature set indicates that our LGP algorithm is able to identify important and relevant metabolites that can better explain the disease of OA.

Moreover, the best classifier among the 1000 evolved models can achieve an MCE as low as 0.067 and an AUC as high as 0.971 for the discovery dataset and 0.067 and 1 for the replication dataset respectively. Given the complexity of the disease, this suggests the effectiveness of using the LGP algorithm to infer the underlying highly non-linear interacting relationships of multiple metabolites that are associated with the disease.

3.3 Identification of Key Metabolic Markers

The goal of our informatics study is to provide a list of important metabolites for future biological validation, such that we can better understand the etiology of the disease and better design its drug treatments. To estimate the importance of each metabolite, we examine its occurrence frequency in both the discovery and

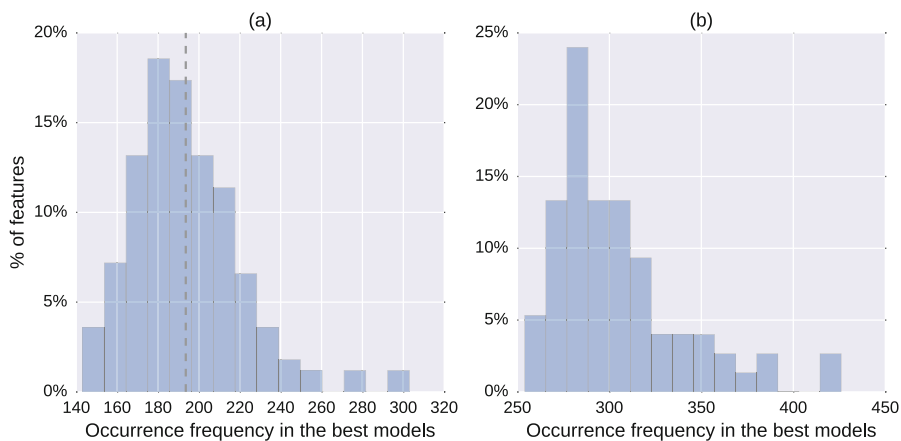


Fig. 3. Distributions of feature occurrence frequency in the 1000 best models on (a) the full set of 167 features and (b) the focused set of 75 features (discovery). In (a), the vertical dashed line represents the mean of the distribution.

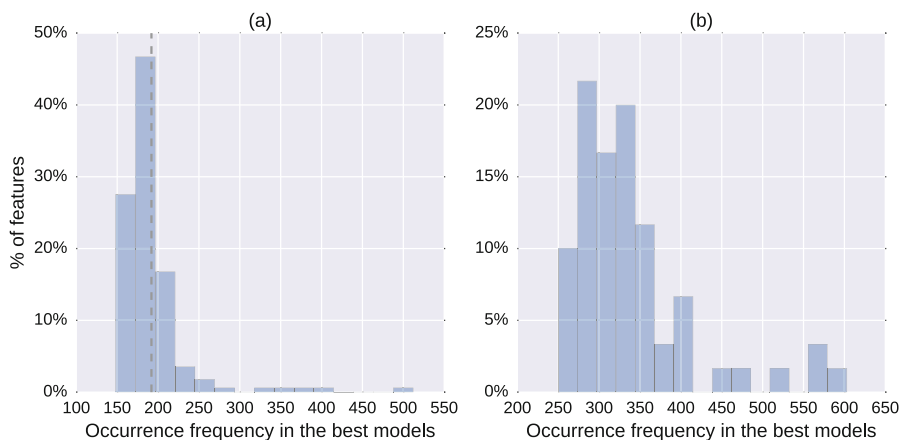


Fig. 4. Distributions of feature occurrence frequency in the 1000 best models on (a) the full set of 167 features and (b) the focused set of 60 features (replication). In (a), the vertical dashed line represents the mean of the distribution.

replication datasets. Figures 3(b) and 4(b) show the distributions of metabolite occurrence frequencies in both datasets in the second round of a more focused classification model construction using reduced feature sets. Comparing to using the full feature sets (Figs. 3(a) and 4(a)), there are more metabolites having much higher occurrence frequencies in the best models. The explanation could be that by removing irrelevant features, our LGP algorithm is able to pick up more important features through a more focused search.

Table 4. Statistics of the classification performance of the 1000 best models (discovery, focused feature set).

	MCE	Sensitivity	Specificity	AUC
Mean	0.317	0.732	0.635	0.714
Median	0.333	0.733	0.667	0.718
Min	0.067	0.267	0.200	0.353
Max	0.600	1.000	1.000	0.971
Std dev	0.088	0.137	0.135	0.103
5% confidence	0.144	0.464	0.370	0.512
95% confidence	0.490	0.999	0.899	0.917

Table 5. Statistics of the classification performance of the 1000 best models (replication, focused feature set).

	MCE	Sensitivity	Specificity	AUC
Mean	0.286	0.751	0.678	0.740
Median	0.267	0.733	0.667	0.744
Min	0.067	0.267	0.067	0.244
Max	0.600	1.000	1.000	1.000
Std dev	0.102	0.135	0.169	0.118
5% confidence	0.086	0.487	0.348	0.509
95% confidence	0.485	1.015	1.009	0.971

Recall that the discovery focused feature set has 75 metabolites, and the replication focused feature set has 60. We make the union set of those two (98 metabolites) and assign the occurrence frequency as zero for those metabolites that do not appear in the opposite set. That is, if a metabolite A only appears in the discovery focused feature set, we treat A 's occurrence frequency as zero in replication. We then show the occurrence frequencies of those metabolites in the union set of discovery and replication (Fig. 5) in order to identify key metabolites whose importance can be both discovered and replicated.

By using a threshold of 0.3 on both axes, we identify 17 key metabolites at the right-upper corner of the scatter plot (Fig. 5). Those 17 key metabolites include the ones that have been reported previously with a strong association with the disease of OA, as well as the ones that haven't been linked to the disease in the literature yet but hold great potentials improving our understanding of the disease. Those new discoveries are particularly interesting since with further biological validation, they could help identify metabolic processes that are potentially related to the disease. The biology of those 17 key metabolites will be explained in more detail in the Discussion section.

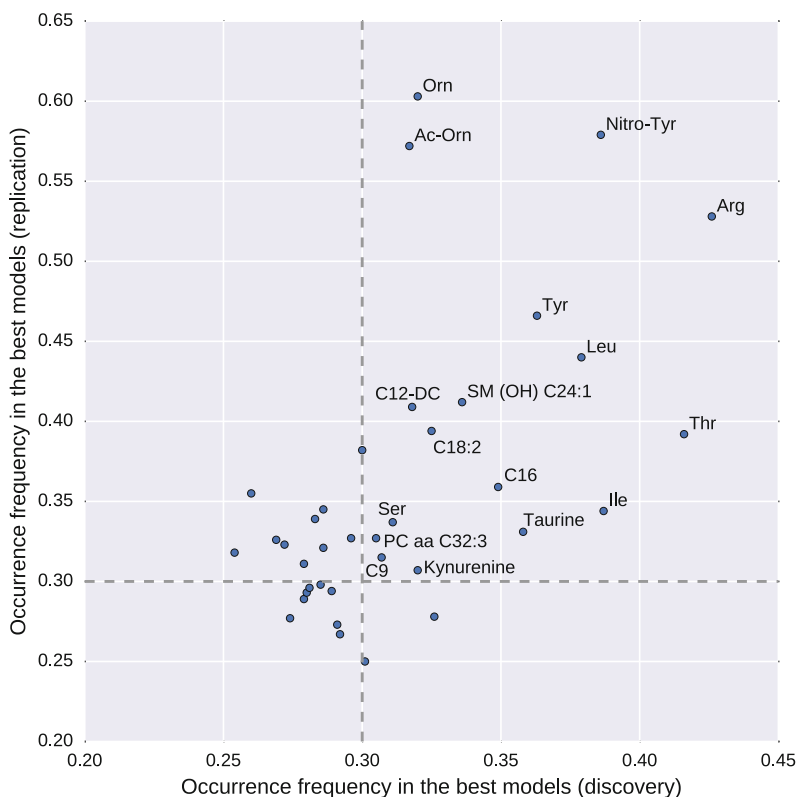


Fig. 5. Scatter plot of normalized metabolite occurrence frequencies in the best models using the focused feature sets. Each data point represents a metabolite. The x-axis is its occurrence frequency in the discovery dataset, and the y-axis is that in the replication dataset. Dashed lines define a set of 17 key metabolites that have higher occurrence frequencies in both datasets comparing to the rest of the features.

4 Discussion

The advancing of biomedical and computational technologies has brought about a new era for systems biology research, where abundant and various types of data become available for quantitative analysis for us to better understand the biology of living systems. The underlying causes of complex human diseases are often multifactorial such that intelligent learning algorithms are needed to identify the combinations of the most relevant biomarkers from hundreds to thousands of biological variables.

Machine learning techniques are often employed for modeling the complex non-linear relationships of combinations of biomarkers and the disease outcome, thanks to their robust heuristic search and learning abilities. However, genetic programming (GP), positioned at the intersection of machine learning and evolutionary computing, has not seen wide applications in systems biology.

In this study, we designed an informatics framework of using a Linear GP (LGP) algorithm to construct classification models and to identify key features for metabolomics studies on the disease of osteoarthritis (OA). Metabolomics is a newly emerging field that looks at the abundance of large sets of metabolites in the human body to study their represented biological processes that are associated with diseases or responses to drug treatment. Given the complexity of metabolism, we speculate that metabolites are associated with the disease in terms of high-dimensional interactions rather than individual effects. The LGP algorithm was able to infer such interactions by constructing highly non-linear symbolic models, as well as ranking features based on their occurrence frequencies in the classification models that can best predict the disease outcome.

We designed a two-round analysis scheme where the full feature set was used to train LGP models at first, and then the subset of more important features was used for a more focused model search. It was observed that the classification performance was significantly improved using the reduced feature set comparing with using the full feature set (Tables 2, 3, 4 and 5). Moreover, by ranking metabolites based on their occurrence frequencies in the best prediction models, we were able to identify 17 metabolites considered important in both of the independent discovery and replication datasets (Fig. 5). Those 17 metabolites include both known metabolic markers in the disease of OA and novel findings.

Arginine (Arg) and its pathway related metabolites, such as *ornithine* (Orn), have been identified to be associated with OA in a previous analysis using traditional methods including pairwise comparison and regression technique [30]. Similarly, branched chain amino acids such as *leucine* (Leu), several acylcarnitines and phosphatidylcholines identified in the current analysis were also reported previously to be associated with OA [31–33] or OA classification [34]. Importantly, the current analysis identified several novel metabolic markers that were otherwise missed by using traditional analytic methods. *Taurine* is the most abundant free amino acid in humans, and may play an important role in inflammation associated with oxidative stress [35], which has been implicated in the pathogenesis of OA [36]. *Taurine* has been reported to be associated with rheumatoid arthritis [37], suggesting taurine might be a novel marker to monitor disease progression of OA but not a diagnosis. *Nitrotyrosine* (Nitro-Tyr) is also associated with oxidative damage and has been found to be associated with aging and the development of OA in cartilage samples from both monkeys and humans [38]. *Kynurenine* pathway from tryptophan generates compounds which can act on glutamate receptors in peripheral tissues or modulate free radical activity and have been implicated in rheumatoid arthritis [39]. Together, these novel findings suggest the involvement of oxidative stress associated metabolic pathways in OA. Further investigations in independent cohorts are warranted to confirm these findings.

Our study demonstrates the power of a GP algorithm in complex classification model search and automatic feature selection for systems biology research. We have entered a golden era for bioinformatics research where large volumes of data that capture the different levels of biological systems are becoming

available and are in need of intelligent and powerful learning algorithms that embrace the complexity of biological systems. We hope this small step can encourage more interdisciplinary communications between evolutionary computing and biomedicine and more explorations on the research front of evolutionary algorithm applications.

Acknowledgments. This research was supported by Newfoundland and Labrador Research and Development Corporation (RDC) Ignite Grant 5404.1942.101 and the Natural Science and Engineering Research Council (NSERC) of Canada Discovery Grant RGPIN-2016-04699 to TH. GZ acknowledges grants from Canadian Institute of Health Research (CIHR), Newfoundland and Labrador Research and Development Corporation (RDC) and Memorial University. We thank all the study participants who made this study possible and all the Operation Room staff at Eastern Health General Hospital and St. Clare’s Hospital who helped for collecting samples.

References

1. Kitano, H.: Systems biology: a brief overview. *Science* **295**(5560), 1662–1664 (2002)
2. Kitano, H.: Computational systems biology. *Nature* **420**(6912), 206–210 (2002)
3. Ideker, T., Galitski, T., Hood, L.: A new approach to decoding life: systems biology. *Annu. Rev. Genom. Hum. Genet.* **2**(1), 343–372 (2001)
4. Cusick, M.E., Klitgord, N., Vidal, M., Hill, D.E.: Interactome: gateway into systems biology. *Hum. Mol. Genet.* **14**(suppl 2), R171–181 (2005)
5. Bruggeman, F.J., Westerhoff, H.V.: The nature of systems biology. *Trends Microbiol.* **15**(1), 45–50 (2007)
6. Shim, S.H.: Cell imaging: an intracellular dance visualized. *Nature* **546**, 39–40 (2017)
7. Wang, K., Lee, I., Carlson, G., Hood, L., Galas, D.: Systems biology and the discovery of diagnostic biomarkers. *Dis. Markers* **28**(4), 199–207 (2010)
8. Butcher, E.C., Berg, E.L., Kunkel, E.J.: Systems biology in drug discovery. *Nat. Biotechnol.* **22**(10), 1253–1259 (2004)
9. Li, Y., Chen, L.: Big biological data: challenges and opportunities. *Genom. Proteomics Bioinf.* **12**(5), 187–189 (2014)
10. Alfieri, R., Milanese, L.: Multi-level data integration and data mining in systems biology. In: *Handbook of Research on Systems Biology Applications in Medicine*, pp. 476–496. IGI Global (2009)
11. Sugimoto, M., Kawakami, M., Robert, M., Soga, T., Tomita, M.: Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr. Bioinf.* **7**(1), 96–108 (2012)
12. Hotelling, H.: Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**(6), 417 (1933)
13. Bishop, C.M.: *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford (1995)
14. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
15. Worzel, W.P., Yu, J., Almal, A.A., Chinnaiyan, A.M.: Applications of genetic programming in cancer research. *Int. J. Biochem. Cell Biol.* **41**(2), 405–413 (2009)
16. Kandpal, M., Kalyan, C.M., Samavedham, L.: Genetic programming-based approach to elucidate biochemical interaction networks from data. *IET Syst. Biol.* **7**(1), 18–25 (2013)

17. Gowda, G.N., Zhang, S., Gu, H., Asiago, V., Shanaiah, N., Raftery, D.: Metabolomics-based methods for early disease diagnostics. *Expert Rev. Mol. Diagn.* **8**(5), 617–633 (2008)
18. WHO Scientific Group: the burden of musculoskeletal conditions at the start of the new millennium. WHO Technical Report Series 919, 218 (2003)
19. Reginster, J.Y.: The prevalence and burden of arthritis. *Rheumatology* **41**, 3–6 (2004)
20. Zhai, G., Aref-Eshghi, E., Rahman, P., Zhang, H., Martin, G., Furey, A., Green, R.C., Sun, G.: Attempt to replicate the published osteoarthritis-associated genetic variants in the newfoundland & labrador population. *J. Orthop. Rheumatol.* **1**(3), 5 (2014)
21. Hu, T., Zhang, W., Fan, Z., Sun, G., Likhodi, S., Randell, E., Zhai, G.: Metabolomics differential correlation network analysis of osteoarthritis. *Pac. Symp. Biocomput.* **21**, 120–131 (2016)
22. Altman, R., Alarcon, G., Appelrouth, D., Bloch, D., Borenstein, D., Brandt, K., Brown, C., Cooke, T.D., et al.: The american college of rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum.* **34**(5), 505–514 (1991)
23. Zhang, W., Likhodii, S., Aref-Eshghi, E., Zhang, Y., Harper, P.E., Randell, E., Green, R., Martin, G., Furey, A., Sun, G., Rahman, P., Zhai, G.: Relationship between blood plasma and synovial fluid metabolite concentrations in patients with osteoarthritis. *J. Rheumatol.* **42**(5), 859–865 (2015)
24. Brameier, M.F., Banzhaf, W.: *Linear Genetic Programming*. Springer, New York (2007)
25. Brameier, M.F., Banzhaf, W.: A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Trans. Evol. Comput.* **5**(1), 17–26 (2001)
26. Guven, A.: Linear genetic programming for time-series modeling of daily flow rate. *J. Earth Syst. Sci.* **118**(2), 137–146 (2009)
27. Song, D., Heywood, M.I., Zincir-Heywood, A.N.: A linear genetic programming approach to intrusion detection. In: Cantú-Paz, E. (ed.) *GECCO 2003*. LNCS, vol. 2724, pp. 2325–2336. Springer, Heidelberg (2003). <https://doi.org/10.1007/3-540-45110-2.125>
28. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. *CoRR* abs/1411.1607 (2014). <http://arxiv.org/abs/1411.1607>
29. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003)
30. Zhang, W., Sun, G., Likhodii, S., Liu, M., Aref-Eshghi, E., Harper, P.E., Martin, G., Furey, A., Green, R., Randell, E., Rahman, P., Zhai, G.: Metabolomic analysis of human plasma reveals that arginine is depleted in knee osteoarthritis patients. *Osteoarthr. Cartil.* **24**, 827–834 (2016)
31. Zhai, G., Wang-Sattler, R., Hart, D.J., Arden, N.K., Hakim, A.J., Illig, T., Spector, T.D.: Serum branched-chain amino acid to histidine ratio: a novel metabolomic biomarker of knee osteoarthritis. *Ann. Rheum. Dis.* **69**(6), 1227–1231 (2010)
32. Zhang, W., Sun, G., Likhodii, S., Aref-Eshghi, E., Harper, P.E., Randell, E., Green, R., Martin, G., Furey, A., Rahman, P., Zhai, G.: Metabolomic analysis of human synovial fluid and plasma reveals that phosphatidylcholine metabolism is associated with both osteoarthritis and diabetes mellitus. *Metabolomics* **12**, 24 (2016)

33. Zhang, W., Sun, G., Aitken, D., Likhodii, S., Liu, M., Martin, G., Furey, A., Randell, E., Rahman, P., Jones, G., Zhai, G.: Lysophosphatidylcholines to phosphatidylcholines ratio predicts advanced knee osteoarthritis. *Rheumatology* **55**(9), 1566–1574 (2016)
34. Zhang, W., Likhodii, S., Zhang, Y., Aref-Eshghi, E., Harper, P.E., Randell, E., Green, R., Martin, G., Furey, A., Sun, G., Rahman, P., Zhai, G.: Classification of osteoarthritis phenotypes by metabolomics analysis. *BMJ Open* **4**, e006286 (2014)
35. Marcinkiewicz, J., Kontny, E.: Taurine and inflammatory diseases. *Amino Acids* **46**(1), 7–20 (2014)
36. Loeser, R.F.: Aging and osteoarthritis: the role of chondrocyte senescence and aging changes in the cartilage matrix. *Osteoarthr. Cartil.* **17**(8), 971–979 (2009)
37. Kontny, E., Wojtecka-Lukasik, E., Rell-Bakalarska, K., Dziewczopolski, W., Maśliński, W., Maśliński, S.: Impaired generation of taurine chloramine by synovial fluid neutrophils of rheumatoid arthritis patients. *Amino Acids* **23**(4), 415–418 (2002)
38. Loeser, R.F., Carlson, C.S., Carlo, M.D., Cole, A.: Detection of nitrotyrosine in aging and osteoarthritic cartilage: correlation of oxidative damage with the presence of interleukin-1 β and with chondrocyte resistance to insulin-like growth factor 1. *Arthritis Rheumatol.* **46**(9), 2349–2357 (2002)
39. Forrest, C.M., Kennedy, A., Stone, T.W., Stoy, N., Darlington, L.G.: Kynurenine and neopterin levels in patients with rheumatoid arthritis and osteoporosis during drug treatment. In: Allegri, G., Costa, C.V.L., Ragazzi, E., Steinhart, H., Varesio, L. (eds.) *Developments in Tryptophan and Serotonin Metabolism*. AEMB, vol. 527, pp. 287–295. Springer, Boston (2003). https://doi.org/10.1007/978-1-4615-0135-0_32