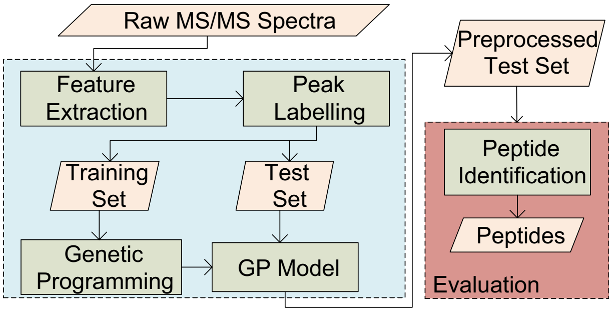
J. Am. Soc. Mass Spectrom. (2019) 30:1294–1307
DOI: 10.1007/s13361-019-02196-5

RESEARCH ARTICLE

Preprocessing Tandem Mass Spectra Using Genetic Programming for Peptide Identification

Samaneh Azari,[1,2] Bing Xue,[1] Mengjie Zhang,[1] Lifeng Peng[3]

[1]School of Engineering and Computer Science, Victoria University of Wellington, Wellington, Kelburn 6012, New Zealand
[2]School of Engineering and Computer Science, Victoria University of Wellington, PO Box 600, Wellington, 6140, New Zealand
[3]Centre for Biodiscovery and School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand



Abstract. One of the major challenges in proteomics is peptide identification from mass spectra containing high noise ratio and small number of signal (b-/y-ions) peaks. However, the accuracy and reliability of peptide identification in such highly imbalanced MS/MS data can be improved by applying a preprocessing step prior to peptide identification aiming at discriminating b-/y-ions from noise peaks in the spectra. In this study, we report a genetic programming (GP)–based preprocessing method for de-noising highly imbalanced and noisy CID MS/MS spectra. GP now becomes a popular machine learning method via automatic programming. GP preprocesses the highly noisy MS/MS spectra by classifying peaks as noise peaks or signal peaks in a binary classification manner. Meanwhile, a set of spectral fragment features based on the MS/MS fragmentation rules is extracted from the dataset to investigate their discriminating abilities by GP. A MS/MS spectral dataset containing thousands of spectra are used to train the GP model. As the GP tree-based representation has the capability for implicit feature selection during the evolutionary process, the evolved GP model with the selected features is compared with the best threshold-based method. The results show that the GP method improved the reliability of peptide identification and increased the identification rate of a de novo sequencing tool, PEAKS, to 99.4% from 80.1% achieved by the best threshold-based method. Moreover, the result of peptide identification by a database search tool, SEQUEST, using the data preprocessed by the GP method was statistically significant compared to the other methods.

Keywords: Mass spectrometry, Tandem mass spectrometry, Genetic programming, Classification, Preprocessing

Received: 25 September 2018/Revised: 15 January 2019/Accepted: 11 March 2019/Published Online: 25 April 2019

Introduction

Proteomics is the large-scale analysis of proteins [1]. The common method for identifying proteins' amino acid sequences and posttranslational modifications of their amino acids in proteomics is to digest the proteins into peptides, analyze the peptides using mass spectrometry, assign the resulting mass spectra to peptides, and match the assigned peptides to proteins using software such as Mascot [2] and SEQUEST [3]. Most of today's proteomics analyzes are done with tandem mass spectrometry (MS/MS) [4]. Peptide and protein identification can reveal potential genetic diseases in an organism, which is the reason that makes this task very crucial. However, the identification task is very challenging if the MS/MS spectra are highly noisy [5]. The large number of noise peaks compared to the small number of signal peaks makes the data extremely imbalanced. The more noise in MS/MS spectra, the more false peptides in the identification. Moreover, having noise in MS/MS spectra results in identifying peptides with low confidence scores, which causes to infer low confidence proteins with a risk of losing true identifications. To overcome the problems imposed by the noise, a

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13361-019-02196-5) contains supplementary material, which is available to authorized users.

Correspondence to: Samaneh Azari; *e-mail:* samaneh.azari@ecs.vuw.ac.nz preprocessing step to de-noise the MS/MS spectra and find signal peaks for more reliable peptide identification is usually performed.

Generally, there are three types of methods to de-noise MS/MS spectra: intensity-based thresholding [3, 6, 7], peak detection methods inspired from digital signal processing [8, 9], and machine learning (ML) algorithms [10–14]. Threshold methods are normally used in database search engines prior to searching in order to simplify the spectra by discarding peaks with intensity below a specific threshold. However, an optimal threshold value is hard to determine and varies from one dataset to another. Moreover, these methods, by only considering the intensity information of peaks and assuming independence of peaks, neglect the hidden interrelationship between them. In an MS/MS spectrum, signal peaks are related to each other. For example, the mass difference between two consecutive signal peaks may equal to the mass of one of the 20 amino acids. Peak detection methods such as Fourier analysis and wavelet analysis usually rely on the shape of the signals and assume stationary signals, which are not the characteristic of signals in mass spectrometry spectra. For low-quality MS/MS data where the peaks are not in well-defined shapes, these methods are considerably less effective [13]. Recently, there is a growing trend to apply ML techniques on MS data in order to discover peptide fragmentation patterns. An ML algorithm in a supervised learning tries to build a model which can predict the intensity pattern of the MS/MS spectra. Unlike the threshold-based methods that only consider one single intensity feature, ML methods typically consider more features and attempt to discover the hidden relationship between the peaks.

A classification method using artificial neural networks (ANNs) to distinguish the signal peaks from noise peaks using a comprehensive full factorial liquid chromatography (LC)-MS/MS benchmark dataset [15] has been developed in [10]. Although ANNs build a non-linear model which is able to detect the hidden interrelationship between the peaks, they are usually black-box or uninterpretable models and still tend to be over-fitted. The combination of machine learning–based preprocessing methods with intensity-based method has also been developed [11, 12]. It is worth investigating the effectiveness of the ML algorithms by themselves on improving the peptide identification.

Genetic programming (GP) belongs to evolutionary algorithms (EAs) which are a family of population-based problem solving techniques who employ Darwinian principle of natural selection and gene theory such as recombination, mutation, natural selection, and survival of the fittest in order to evolve a population of individuals. GP showed promising results when previously it was used in many symbolic regression [16], optimization [17], and classification problems [18]. One of the main advantages of GP in classification tasks is its flexible representation which can be adapted for domain-dependent problems. Moreover, GP using a tree-based representation has the capability for implicit feature selection during the evolutionary process. GP has the ability to automatically evolve a model that fits the training data without any prior knowledge or assumption. GP has the potential to cope with complex problems and has good learning capability even from imbalanced data. GP can adapt its fitness function to evolve an individual that is capable of dealing with the class imbalance problem [19]. Unlike other machine learning algorithms, GP has the ability to combine several advantages: GP can integrate various types of data and generate effective models; such models are not black-box models, instead they are highly interpretable and readable by human.

Recently, GP has been successfully applied on mass spectrometry (MS data) on various problems such as biomarker detection, metabolite quantification, and metabolic pathway modeling [20]. GP proved to be a promising tool in MS/MS analysis, but its potential for further improvement in more reliable peptide identification has not been systematically investigated.

In an MS spectrum, each precursor ion, which indicates the *m/z* value of a peptide, can be selected and fragmented into hundreds of fragment ions that construct an MS/MS spectrum. During fragmentation by the collision-induced dissociation (CID) technique, different fragment ion types are generated. In the CID fragmentation technique, we are only interested in b-/y-ions because the amino acid sequence of an MS/MS spectrum can be determined by the mass differences between b-/y-ions. However, during the fragmentation, different ion types such as isotopologues, neutral losses, and doubly charged ions are produced. The presence of different types of ions along with the background noise, resulted for example from low sensitive mass spectrometer instruments, can produce a large and complex search space for the peptide identification tool to explore, leading to a high false discovery rate. Therefore, prior to peptide identification, it is worth investigating which ion types should be considered as signals and noise peaks. This investigation helps to have a clear definition of background noise in the data. As the background noise does not necessarily mean noise from low instrument accuracy, anything which makes the peptide identification tool having a big search space and decreases the performance of the identification tool should be removed from data prior to peptide identification. Also, for the supervised classification methods, each peak in the training dataset is required to be labeled as either signal (b-/y-ions) or noise; it is worth investigating which ion types should be considered as signals in the training set of the golden standard dataset. An accurate and reliable MS/MS golden standard dataset should contain the fragment ions labeled as signal peaks that can help both peptide identification tools to identify high confident peptides and the machine learning–based preprocessing method to distinguish noise peaks from signal peaks. Therefore, an effective labeling can directly influence the performance of both classification method and the peptide identification rate.

In our previous work [14], genetic programming (GP) showed a great potential for imbalanced classification particularly in preprocessing tandem spectra aiming at improving the reliability of peptide identification. As the MS/MS data is highly imbalanced, GP proved to be more stable than the six investigated different types of classification algorithms k-nearest neighbor (K-NN), support vector machine (SVM), naive Bayes (NB), decision tree (DT), random forest (RF), and multilayer perceptron (MLP). Using a training set and a test set containing thousands of MS/MS spectra from the comprehensive full factorial LC-MS/MS benchmark dataset the G, GP got the highest average accuracy and sensitivity on both datasets compared to the aforementioned classification algorithms. However, the current GP method achieved 72% of retention of signal peaks and 86% of reduction of noise peaks as tested on a dataset containing 1674 MS/MS spectra. The GP-based preprocessing method only used 4 spectral features, while other potential ways to improve the performance have not been investigated.

The goal of this study is to investigate the capability of GP for preprocessing tandem spectra, in order to improve the capability of retention of signal peaks and reduce the noise peaks. Based on our previous work [14], we will improve the approach in the following ways:

1. Developing a GP-based preprocessing method to de-noise the MS/MS spectra and investigating extraction of effective spectral features aiming at providing more evidence for signal peaks to be distinguishable from noise peaks
2. As there is no golden standard highly imbalanced dataset containing already labeled signal and noise peaks, investigating creation of a suitable golden standard dataset aiming at increasing the peptide identification reliability and evaluating the effectiveness of the golden standard dataset using GP, and
3. Comparing the effectiveness of the proposed GP-based preprocessing method with un-preprocessed data and the best threshold-based preprocessing method in terms of reliability of the peptide identification

Methods

The Proposed GP-Based MS/MS Preprocessing Method

Figure 1 presents the full workflow of the GP-based MS/MS preprocessing method followed by an evaluation step. The GP-based preprocessing method includes three steps which are feature extraction, peak labeling, and binary classification. The feature extraction step is composed of extracting intensity-based features for each peak in the spectrum. The peak-labeling step determines the class label of each peak as either signal or noise. The training set is used by GP for building the model to classify the peaks while the test set is used to evaluate the model. The output of the GP model is the preprocessed data (test set) which is submitted to an evaluation step where a peptide identification tool is used to identify the peptides. This study uses PEAKS, a benchmark de novo sequencing software [21], as the peptide identification tool in the workflow. The results of the identification is analyzed to evaluate the effectiveness of the GP method in terms of improving the peptide identification reliability. The following section explains the three steps of the GP-based preprocessing method in detail.

Feature Extraction

The intensity value of each peak in an MS/MS spectrum can be used to extract a set of spectra features that explain the CID fragmentation properties of peptides. Table 1 presents a total number of 7 groups of spectral features extracted from the MS/MS data. These spectral features indicate peak characteristics which can be good discriminators between the signal and noise peaks. All groups include only one feature except for groups 3, 4, and 7, which contain parametric features where changing the parameter values results in extracting new features.

Given spectrum S with n peaks and precursor mass of m_{prec}, let $S = (mz_{(1)}, mz_{(2)}, mz_{(3)}, \ldots, mz_{(n)})$ denotes a spectrum with an intensity vector of $I = (I_1, I_2, I_3, \ldots, I_n)$. The ith peak in the spectrum corresponds to the mass-to-charge value of $mz_{(i)}$ with intenstity value of I_{i}. More details about how to extract the features of each group from the spectrum are explained as follows:

Group (1): the "normalized *m/z*" feature [13] normalizes the *m/z* value of each peak to an integer value between 0 and 100.

$$f_{\text{norm}_{mz}}\left(mz_{(\text{i})}\right) = \left\lfloor \frac{mz(i) \times 100}{m_{\text{prec}}} \right\rfloor$$

Group (2): The "normalized and discretized intensity" feature [11] divides the intensity value of the current peak to the highest intensity value in the whole spectrum. Since the values of intensities within the whole spectrum normally are very fluctuated from a very small intensity value of less than 100.00 to 10,000.00, to have a better scaled values, discretization is then applied on the normalized values to map them into n discrete bins. For example, for $n = 5$, the normalized intensities are rounded up to either 0.05, 0.10, 0.20, 0.40, 0.80, or 1.00 to be discretized [11].

$$f_{\text{norm}_{\text{intensity}}}(I_{\text{i}}) = \left\lfloor n\left(\frac{I_{(\text{i})}}{I_{(\max)}}\right) \right\rfloor / n$$

where $I_{\max}$ is the most abundant peak in the spectrum, I_{i} is the intensity of the current peak, and n indicates the number of discrete intervals.

Group (3): The "top X intensities in Win ± Z" features [7] determine whether or not the current peak is among the top X most intense peaks in the window size of "± Z" around itself. These features have binary values of either 0 or 1. The values of X and Z can be determined empirically or based on the literature [7].

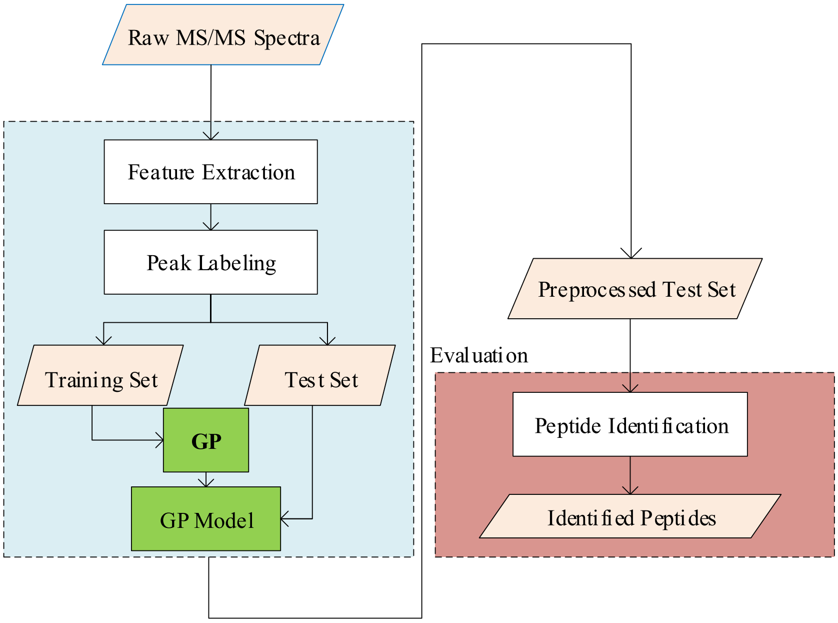


Figure 1. The MS/MS analysis workflow composed of the proposed GP method for preprocessing the spectra and an evaluating step

Group (4): The "local intensity rank in Win ± Z" features count the number of peaks that are the same or are less abundant than the current peak within a local window of ± Z. A big rank indicates a significant peak. The ranks are then normalized by dividing by the total number of peaks in the window of ± Z.
Group (5): The "global rank" feature normalizes the intensity rank of the current peak compared to the complete spectrum.

$$f_{\mathrm{norm}_{\mathrm{rank}}}(I_{\mathrm{i}}) = \frac{\mathrm{rank}(I_{\mathrm{i}})}{n}$$

Group (6): The "complementary ion" feature investigates if the complementary ion of the current peak exists in the whole spectrum. In CID fragmentation, a complete peptide fragmentation gives a contiguous series of ions. However, sometimes, due to the low ion fragmentation efficiency of the mass spectrometer, some ions are not available in the spectrum. By finding the complementary ion peaks, undetected ions can be added to the spectrum. The sum of the two complementary ions' masses should be equal to the precursor mass of the spectrum. Based on the CID fragmentation parameters and the dataset, a mass tolerance is considered to estimate the existence of the complementary ion of the current peak.

$$f_{\mathrm{c}}\left(mz_{(\mathrm{i})}\right) = \begin{cases} 1, & \text{if } mz(i) + mz(j) + \delta = m_{\mathrm{prec}} \\ 0, & \text{otherwise} \end{cases}$$

where $1 \le j \le n$ and n is the total number of peaks in the spectrum. δ is the mass tolerance of the mass spectrometry device used to ionize the spectra in the dataset and varies from one dataset to another.

Group (7): The "sister ion" features check the existence of the sister ions of the current peak. A sister ion is a peak that can be found at the fixed $\Delta m/z$ value away from the current m/z value. Based on the literature, a list of 10 common sister ions including Δ values of $\Delta = \{-2, -1, 1, 2, 17, 18, 28, 34, 35, 36\}$ are considered in this study. These numbers are related to the loss mass of H_2O, NH_3, H_2O–H_2O, H_2O–NH_3, or isotopic ions. Later on, in the experiment section, a set of experiments where a larger range of all possible values from −2 to 143 will be conducted. The purpose of running those experiments is to investigate if considering the entire 145 sister ions can improve

Table 1. List of Spectral Features. *N* Denotes a Normalized Value; *D* Specifies a Discretized Value; *B* Denotes a Binary Value

Group	Features	Feature name	Value
(1)	{f1}	Normalized *m/z*	*N*
(2)	{f2}	Normalized and discretized intensity	*N,D*
(3)	{f3,...,f15}	Is top X intensities in Win ± Z	*B*
(4)	{f16,...,f28}	Local intensity rank in Win ± Z	*N*
(5)	{f29}	Global rank	*N*
(6)	{f30}	Complementary ion	*B*
(7)	{f31,...,f40}	Sister ions	*B*

the classification performance comparing to the case when only 10 sister ions are used.

Labeling Peaks/Instances

Based on the CID fragmentation rules of doubly charged peptides [22], the theoretical spectrum of each experimental spectrum is constructed. It is worth mentioning that the theoretical spectrum does not contain noise peaks, whereas it only includes signal peaks. After constructing the theoretical spectrum, the theoretical and experimental spectra are matched against each other. Each peak in the theoretical spectrum within a mass tolerance of 0.2 Da (unified atomic mass unit or Dalton) is matched against that from the experimental spectrum. The 0.2-Da error tolerance was previously used for producing the full factorial benchmark dataset [15].

Creation of the Training Set and Test Set

After applying feature extraction and peak labeling on experimental spectra, each peak contains a set of features followed by a label which is either signal or noise. Then, the data is divided into two sets: training set and test set. The training set is used during the GP learning process to build the model and the test set is used to evaluate the GP model.

GP Program Representation, Classification Strategy, and Fitness Function

GP uses a population of computer programs as individuals to build a model, searching to find a good solution for the problem during the evolutionary process. The goodness of individuals, which determines their potential to survive and represents their ability to solve the problem, is measured by using the fitness function. The individuals can be modified by genetic operators to breed new individuals [23]. GP simulates evolution by employing fitness-based selection where the fitter programs are expected to be chosen for producing new individuals. The computer structures used in GP can be in the form of tree, linear, or graph-based structures. The most popular GP structure is tree-based which is composed of structural units namely terminal and function sets. The terminal set, which forms the leaves of the GP tree, provides the inputs to the individuals and may contain variables/features and constants. The function set represents the internal nodes and may consist of arithmetic operators, conditional operations, or user-defined operators.

The overall structure of a GP algorithm is illustrated in Figure 2 composing of the following steps:

1. Initialization: GP employs the function and terminal sets to generate a population of initial/candidate solutions.
2. Fitness evaluation: GP executes each individual (program) and evaluates the goodness of the individual using a user-defined fitness function.
3. Individual selection: Using a specific selection procedure (e.g., tournament selection is used in this study), GP selects the individuals with better fitness values for the reproduction process.
4. Genetic operators: GP transforms the (initial) population by applying genetic operators (crossover, mutation, and reproduction) to create new individual programs which are more likely to contain better fitness values.
5. Repeat steps 2–4 until the stopping criterion is met. A stopping criterion determines when to stop the evolutionary process. The process can be stopped when an ideal individual with a specified fitness value has been found or when a maximum number of generations has been reached.

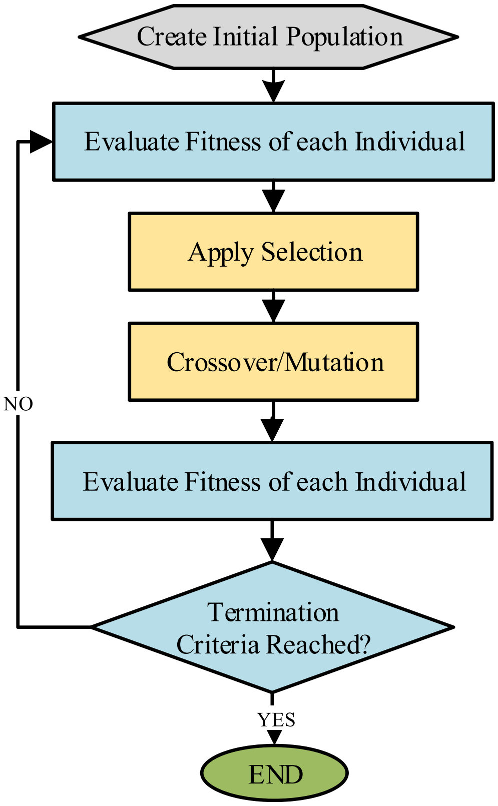

Create Initial Population
Evaluate Fitness of each Individual
Apply Selection
Crossover/Mutation
Evaluate Fitness of each Individual
Termination Criteria Reached?
NO
YES
END

Figure 2. The overall flowchart of a GP algorithm

A tree-based GP classification system is designed to classify highly imbalanced MS/MS data. Table 2 shows the list of parameters used in the GP system. A set of arithmetic operations including summation, multiplication, differentiation, protected division, and the trigonometric sin function are considered as the function set of the GP system. While all the arithmetic operators take two arguments, the sine function gets one argument. Each of the operator in function set returns one single argument. Moreover, the set of spectral features extracted from the MS/MS data along with randomly generated floating point numbers are used as the terminal set of the GP system. The output of the GP tree is a single floating point value. Since it is a binary classification, zero is considered as the threshold value of the GP-tree output to determine the class label of each peak as either noise or signal. A peak with a negative value of

Table 2. Genetic Programming Parameters

Parameter	Value
Function set	(+, −, ×, /, sin)
Terminal set	(Features from dataset, random constant)
Initial population	Ramped Half and Half
Population size	1024
Generations	50
Mutation rate	0.19
Elitism rate	0.01

the GP-output is considered to belong to the noise class (majority class), whereas a positive value indicates that the peak belongs to the signal class (minority class). To implement the GP system, the evolutionary computation Java-based (ECJ) package is used [24].

The fitness function is a key component in GP as it measures the goodness of the candidate solutions/individuals. As the MS/MS data is highly imbalanced, the traditional accuracy which is the average ratio of correctly classified instances might cause the evolved GP programs to be biased towards the majority class. Therefore, to avoid this issue, a weighted fitness function including true positive rate (TPR) and true negative rate (TNR) with coefficient of α is considered:

$$\begin{aligned} A{-}acc &= \alpha \times \left(\frac{TP}{TP+FN}\right) + (1{-}\alpha) \times \left(\frac{TN}{TN+FP}\right) \\ &= \alpha \times TPR + (1{-}\alpha) \times TNR \end{aligned} \tag{1}$$

where TP/TN indicates the correctly classified signal/noise peaks, whereas FP/FN represents the incorrectly classified signal/noise peaks and is α a coefficient that needs to be determined empirically.

Design of Experiments

MS/MS Datasets In this study, the spectra from the comprehensive full factorial LC-MS/MS benchmark dataset, which is designed for evaluating data analysis tools, are used [15]. The dataset is obtained from the linear ion trap Fourier-transform (LTQ-FT, Thermo Fisher Scientific) with the collision-induced dissociation (CID) technique. The full factorial dataset contains 50 protein samples extracted from *Escherichia coli* K12. For a 2 × 3 full factorial design, the samples are spiked with different concentrations of bovine carbonic anhydrase II and/or chicken ovalbumin. The data was acquired using multiple reversed-phase columns and instrument calibrations over a period of 2 months. The ground truth is composed of a comprehensive collection of validated identified peptides by the Mascot v2.2 which is searched against a curated Refseq [25] release 33 *E. coli* database. The set of peptide spectrum matches has a minimum Mascot score of 30. The identified peptides have a minimum length of six amino acids. The datasets used in this study are shown in Table 3. The following are more details about these datasets:

Table 3. Datasets

Datasets		No. of spectra	No. of peaks
Original dataset	Train	2630	1,730,190
	Test	253,732	185,224,471
Sampled dataset	Train	10	9958
	Test	5	4475

Original dataset: including 2630 MS/MS spectra in the training set and a test set of 253,732 MS/MS spectra. This is the base for creating the golden standard dataset. Also, this dataset is used for evaluating the methods in terms of peptide identification.

Sampled dataset: including 10 MS/MS spectra in the training set and five spectra for test set. All these 15 MS/MS spectra are selected from the training set of the original dataset. This dataset is used for the purpose of tuning the GP fitness function and finding appropriate feature parameters.

Experiment 1: Investigating Important Ion Types in Peptide Identification In order to have an accurate and effective golden standard MS/MS dataset to be used by any machine learning–based preprocessing method to increase the peptide reliability, important ion types in MS/MS spectra should be investigated. Figure 3 illustrates the workflow of investigating the important ion types in peptide identification. The workflow starts with an experimental MS/MS spectrum with known peptide. The spectrum is submitted to the Mascot, a database search engine, for labeling each peak in the spectrum. Different peaks/ions are extracted from the spectrum to create different scenarios from 1 to 7. Table 4 shows the Mascot parameter settings. Each scenario containing a spectrum with different ions is submitted to both PEAKS, a benchmark de novo sequencing software [21], and SPIDER, a benchmark database search tool [27], to re-identify the spectrum. The best common scenario with the highest confident identified peptide is chosen to be evaluated by the GP-based preprocessing method.

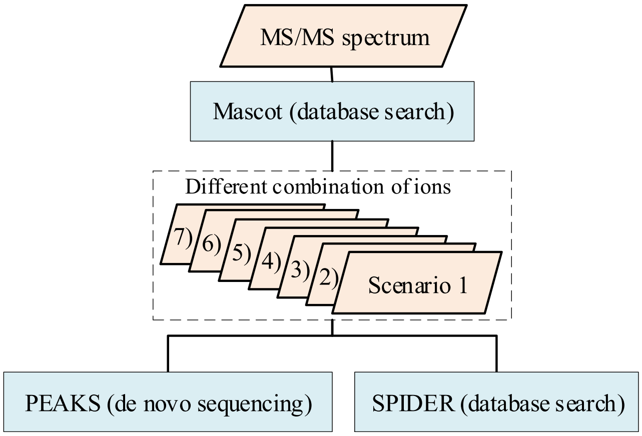


Figure 3. The workflow of investigating important ion types for more reliable peptide identification

Table 4. Mascot Database Search Parameter Setting

Protein database		Spectrum	
Database name	*Escherichia coli* (strain K12, containing 4313 proteins) [26]	Min. precursor mass	350 Da
Enzyme name	Trypsin	Max. precursor mass	5000 Da
Max. missed cleavage	1	Fragment match options (export search results)	
MS tolerance	10 ppm	Charge details	+ 1 and/or + 2
MS/MS tolerance	0.8 Da	Matched tolerance	0.2 Da
Percolator setting		Selected ions	Different combinations of b, y, -H_2O, -NH_3, immonium
Validation based on	*q*-value		
Cutoff *q*-value	0.01		

The single experimental spectrum is chosen from the ground truth provided by LC-MS/MS benchmark dataset [15]. The spectrum corresponds to doubly charged peptide sequence “SEQGMSLLQPGK.” This spectrum is submitted to the Mascot database search tool [2] to be searched against the *Escherichia coli* K12 [26] protein database with Fragment match tolerance of 0.8 Da. Table 4 shows more details of the database search parameter setting.

The Mascot search result returns peptide “SEQGMSLLQPGK” which is the same as the ground truth. The result of the Mascot database is exported as an annotated spectrum where each peak is labeled as: y(1+), b(1+), y(2+), b(2+), b(1+)-H_2O, b(1+)-NH3, y(1+)-H_2O, y(1+)-NH_3, and no label which is considered noise. To find out which ion types should be labeled in the golden standard dataset, different scenarios are provided to test the peptide identification rate using different combinations of ions. The scenarios are as follows:

1) Raw spectrum containing all ion types and noise peaks.
2) Labeling only matched doubly charges b-/y-ions as signal peaks. This includes {b(2+), y(2+)}.
3) Labeling only matched singly charged b-/y-ions as signal peaks. This includes {b(1+), y(1+)}.
4) Labeling both matched singly and doubly charges b-/y-ions as signal peaks. This includes {b(1+), y(1+), b(2+), y(2+)}.
5) Labeling matched singly charged and neutral losses as signal peaks. This includes {b(1+), y(1+), b(1+)-H_2O, y(1+)-H_2O, b(1+)-NH_3, y(1+)-NH3}.
6) Labeling CID-simulated fragments singly charged ions as signal peaks.
7) Labeling CID-simulated fragments singly charged ions and neutral losses as signal peaks.

In each scenario, only those peaks which are mentioned in the description of each scenario are submitted to a de novo sequencing tool and a database search engine, and other peaks are removed from the spectrum. The results of these experiments will indicate which scenario results in a higher score identified peptide. Two different commonly used peptide identification tools are used to make a stronger conclusion. Based on the results of this experiment, the peaks in the MS/MS data are labeled to create the golden standard dataset containing a training set and a test set.

Experiment 2: Tuning the GP-Based Preprocessing Method and Evaluating Its Effectiveness for Improving the Peptide Identification Reliability A set of experiments including tuning the GP fitness function to find appropriate coefficients, investigating the effectiveness of each individual features using different parameters and combining features, is performed to figure out the best GP model. The model is then applied to the golden standard dataset. The preprocessed MS/MS spectra (test set) is submitted to a peptide identification tool to evaluate the effectiveness of the GP method. In the flowchart of Figure 1, PEAKS is used for automating accurate sequencing MS/MS spectra. The result of PEAKS is a set of identified peptides with different confidence scores. It is worth mentioning that, in peptide identification in PEAKS, an average local confidence score (ALC) indicates how reliable the result is. An ALC score reflects the average correct ratio of the predicted amino acids in a peptide sequence. The higher the confidence score, the more reliable the peptide identification. In PEAKS, ALC scores range from 0 to 99% and a score at 55% or above, as suggested in the PEAKS website [28], is considered a confident match. The entire sequence of a peptide is not necessarily to be mapped due to incomplete fragmentation and difficulty in detecting the signal peaks of the fragments from the beginning and the end of the peptide sequences in MS/MS. The results of peptide identification are grouped into five intervals which are {[90, 99), [80, 90), [70, 80), [60, 70), [55, 60)}. For each interval, the number of peptides identified by PEAKS is counted. These results are then compared to those of unpreprocessed data and to the results of MS/MS spectra preprocessed by an intensity-based threshold method.

Results and Discussion

Results of Experiment 1: Important Ion Types in Peptide Identification

This section presents the results of submitting all scenarios to PEAKS and SPIDER illustrated in the flowchart of Figure 3. According to the flowchart, each scenario containing different set of peaks from the considered spectrum is submitted to these two tools. These tools identify the peptide sequence corresponds to each scenario and reports the identified peptide sequence along with a confidence score for the identification.

Table 5 presents the results of peptide identification on each scenario using the PEAKS de novo sequencing and the SPIDER database search. As previously mentioned, a single experimental spectrum is chosen from the ground truth and its corresponding doubly charged peptide sequence is "SEQGMSLLQPGK."

Any letter of the identified sequence which is in bold indicates an exact match against the corresponding letter of the ground truth sequence. PEAKS score is a confidence score reported by PEAKS and indicates the average correct ratio of the predicted amino acids in a peptide sequence. The ALC scores range from 0 to 99%. In this table, the scenario that obtains a high PEAKS and SPIDER confidence score with more similar sequence (more number of bold AA letters) to the ground truth is selected as the best case and determines the most important ion types to peptide identification. This indicates the best decision on the ion types to be selected for labeling the training data of the machine learning method. The following sections analyze the results of each tool separately.

The Scenarios Submitted to PEAKS and the Peptide Identification Results

Table 5 presents the results of each scenario submitted to the de novo sequencing tool, PEAKS, for automated and accurate sequencing of the spectrum. The results in Table 5 show that submitting the raw spectrum (scenario 1), which contains all the ion types, including noise peaks that are not satisfying, results in sequence with a low ALC score of 34. Moreover, if only matched doubly charged ions (scenario 2) are selected out of the all peaks in the spectrum and submitted to the de novo sequencing tool, a worse result, ALC = 10, is obtained. However, the matched singly charged of b-/y-ions (scenario 3) gives a better result of ALC = 61. In the next experiment, both singly and doubly charged ions are combined (scenario 4), but the ALC score decreases to 60. The next scenario combines singly charged and neutral losses ions (scenario 5) to investigate whether the presence of neutral ions can improve the identification rate. However, the results deteriorate (ALC = 52) due to the fact that the presence of neutral losses ions and doubly charged ions makes the ladder complicated and increases potential false positive sequences. Therefore, so far, only using the matched singly charged ions (scenario 3) are the best choice (ALC = 61). However, only using the matched ions may make the ladder incomplete and deteriorate the performance of peptide identification. In the next scenario, the ions are constructed virtually based on the known CID fragmentation rules of doubly charged peptides [22] using only b-/y-(1+) (scenario 6). This results in a complete ladder and presents a higher ALC score and closer to the exact match compared to the previous scenarios. The last experiment combines CID b-/y-(1+) and neutral losses (scenario 7). Although the ALC score is the highest among all experiments, the sequence is far from the exact match. Therefore, it can be seen that scenario 6 where only CID ions are used, gives the best results. Furthermore, another set of experiments using the database search tool, SPIDER, is conducted in the following to make a stronger conclusion.

The Scenarios Submitted to SPIDER and the Peptide Identification Results

Here, the same set of scenarios are submitted to SPIDER. The purpose of running this experiment is to see how a database search tool interact with different sets of ions/peaks in the spectrum and to check whether the previous results from the de novo sequencing is consistent with the results from a database search tool.

From Table 5, it can be seen that similar to the PEAKS results, using doubly charged ions and neutral losses ions (scenario 2) does not help the peptide identification. However, since there is a protein database to be searched again, it can be seen that in three scenarios, 3, 4, and 6, the results are the same as with each other. However, both tools on scenario 6 show good results. Therefore, only extracting CID singly charged b-/y-ions can "guarantee" to obtain a reasonable peptide identification rate. Shao et al. [13] has also reported that complementary signal peaks are more likely to be found at a charge state of + 1 than at other charge states.

So based on the results in this section, both datasets introduced in Table 3 are labeled by considering only CID singly charges ions as signal peaks and the rest of the peaks as noise peaks. Table 6 presents more details related to the number of signal and noise peaks in both datasets. SNR represents signal-to-noise ratios. It can be seen that both datasets are highly imbalanced.

Table 5. Comparisons of Various Scenarios Containing Different Set of Ion Types Submitted to PEAKS and SPIDER for Peptide Identification. The Ground Truth Sequence Is SEQGMSLLQPGK. The Scenario with High Scores for Both Identification Tool Indicates Containing the Most Important Ion Types

Scenario no.	Sequence identified by PEAKS	PEAKS score (%)	Sequence identified by SPIDER	SPIDER score (%)
1)	WNKVELASAE**K**	34	DLGFLPGDLAE**K**	17.42
2)	FLLLKEYGY**K**	10	FLLLDEPTRGL	19.34
3)	LCK**GMSLLQPGK**	61	**SEQGMSLLQPGK**	*42.02*
4)	CLK**GMSLLQPGK**	60	**SEQGMSLLQPGK**	*42.02*
5)	YHQ**LL**SMT**PGK**	52	LTTLLLSQGTPM	21.63
6)	LCK**GMSLLQPGK**	**70**	**SEQGMSLLQPGK**	**42.02**
7)	CLLV**LL**SMT**PGK**	72	GQDQLLSLAGGDT	25.02

Results of Experiment 2: Effectiveness of the Proposed GP Method on Peptide Identification

Tuning the Fitness Function of GP In binary classification of imbalanced datasets, it is highly important to identify instances belonging to the minority class correctly. Therefore, as previously shown in Eq. (1), a weighted average of the true positive rate (TPR) or sensitivity (SE) and true negative rate (TNR) or specificity (SP) is used to evaluate the evolved GP classifiers. In this section, different α coefficients are experimentally checked to find a suitable α value. The labeled sampled dataset containing 10 spectra in the training set and 5 spectra in the test set from Table 6 is used. Figure 4 shows the classification results of GP using different coefficients for SE and SP on training and test sets of the labeled sampled dataset, respectively. It can be seen that by increasing the coefficient of SE, the specificity in both the training and test sets drops, On the other hand, giving a high coefficient to SP can decrease the sensitivity and this is not desired. Therefore, it seems that the sets of coefficients ($0.5 \times SE + 0.5 \times SP$) or ($0.6 \times SE + 0.4 \times SP$) work better compared to other sets. However, ($0.6 \times SE + 0.4 \times SP$) results in higher sensitivity value decreases the precision due to the increase in false positives. Therefore, the rest of the experiments are done by using $\alpha = 0.5$, i.e., ($0.5 \times SE + 0.5 \times SP$) as the fitness function of GP.

Tuning the Parameters in "Top X Intensities in Win ± Z" Feature In this section, a set of experiments is conducted where GP is used to perform binary classification on labeled sampled dataset using one single feature of "top *X* intensities in Win ± *Z*." A range of 1–10 is considered for *X* while *Z* ranges from 27 to 100 with an increment of 10. The value 27 is suggested by [29] where a top 1 in Win 27 approach is applied on the MS/MS spectrum to remove the potential noise peaks. Here, the aim of running these experiments is finding appropriate parameter values for *X* and *Z* that keep the classification performance reasonably high.

The average accuracy graphs in Figure 5 show that for *X* values from 3 to 8, and for window size *Z* values less than 60 (the first four graphs starting from top left), the results of classification in terms of average accuracy is reasonably high (more than 85%). By increasing the window size *Z* to more than 70, the average accuracy graphs keep increasing for all *X* values in range the 1–10. It means that for window sizes more than 70, the *X* range of 1–10 is not sufficient to give a downward trend to the average accuracy graphs. The reason is that for a big window size *Z*, increasing the value of *X* gives more chances to keep the potential signal peaks. Because of retention of more signal peaks, the classification accuracy increases. As *Z* indicates a neighborhood around each peak, we are not interested in big neighborhoods which can turn the local "top *X* intensities in Win ± *Z*" feature to a global feature. Therefore, one solution could be limiting the range of *Z* to the mass of the smallest amino acid which is 57 Da (as listed in Table 1, 13 sets of *X*,*Z*). Also, another alternative would be considering all *X* and *Z* values in the graphs of Figure 6 where the average accuracy is more than 85% in both train and test sets. As there are 45 cases with that condition, so, the next experiments investigate the classification results of using 45 sets of *X*,*Z* as features. Moreover, 13 sets of *X*,*Z* reported in [7] including: *X*,*Z* = (1,27), (3,56), (4,40), (4,50), (4,60), (6,25), (6,30), (6,40), (6,50), (6,60), (8,40), (8,50), (8,60), were used for the purpose of noise thresholding of MS/MS spectra. However, they were not used as the features for the machine learning–based preprocessing method. In this study, these sets of features are used with the GP method and the results will be compared with other groups of features including the top 45 sets of *X*,*Z* values obtained in this experiment.

Classification Results of each Group of Features This experiment investigates the effectiveness of each group of features in improving the classification performance of the GP system. For "top *X* intensities in Win ± *Z*" and "sister ions" features, various parameter values are considered to investigate all the possibilities. Figure 6 a and b show the classification results of each group of features from Table 1 on the labeled sampled dataset from Table 6.

The results show that among the seven groups of features, "top *X* intensities in Win ± *Z*" and "local rank in Win ± *Z*" groups get the highest classification results on both the training and test sets compared to other feature groups. The reason is that these two groups try to identify possible noise peaks within a local window around the current peak and keep signal peaks. The second best group of features is "global rank," where each peak is compared to the all peaks in the spectrum.

"Complementary ion" and "sister ions" groups are the third best sets of features. These features are based on the CID fragmentation rules and try to find the hidden relationship between the peaks in the whole spectrum without considering the intensity of each peak. The last two best features are "normalized *m/z*" and "normalized intensity" groups.

As mentioned before, for the two groups of features, "top *X* intensities in Win ± *Z*" and "sister ions," two sets of

Table 6. Datasets with Different Ion Types Labeled as Signal Peaks

Datasets		No. of spectra	No. of signal peaks	No. of noise peaks	SNR
Golden standard dataset	Train	2630	42,960	1,687,230	40
	Test	253,732	4,095,873	181,128,598	44
Labeled sampled dataset	Train	10	278	9680	35
	Test	5	115	4360	38

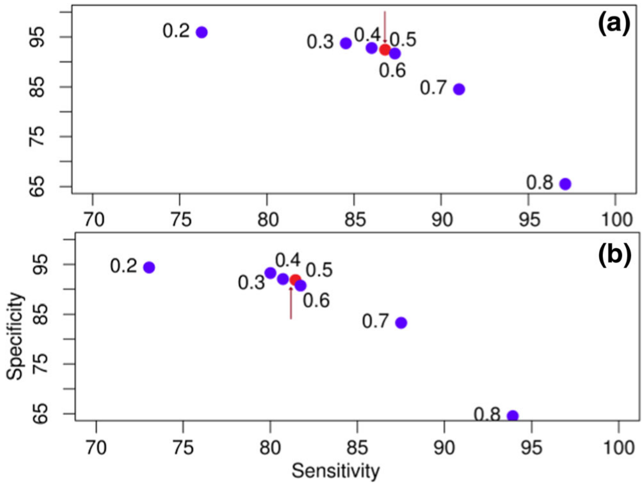


Figure 4. The classification results of GP on labeled sampled dataset for different coefficients of sensitivity and specificity in the GP fitness function: $\alpha \times \mathrm{SE} + (1 - \alpha) \times \mathrm{SP}$, where $\alpha = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. (**a**) The classification results on the training set. (**b**) The classification results on the test set. As there are overlaps in the scatter plots of both figures for more clarification, the arrows are pointing to $\alpha = 0.5$ for more clarification

experiments including different parameter values are run to investigate appropriate parameters for these groups. For the group "top X intensities in Win ± Z," two sets of 13 and 45 features are used. The bar charts (Figure 6 a and b) show that the classification results in terms of average accuracy, sensitivity, and specificity on both the training and test sets are relatively the same for both sets of 13 and 45 features. As using more features during the learning process requires more processing time, for the group "top X intensities in Win ± Z," 13 features are considered afterwards. Also, the results of the bar charts show that for "sister ions" group, considering only 10 common features will be sufficient to get a reasonable classification result on both train and test sets compared to having all 145 possible features, so far, the results of each individual groups of features are obtained. It is worth to investigate the classification results of combining all the features together.

Classification Results of Combining All the Features The existing GP-based preprocessing method [14], only considered 4 features. In this work, a different set of features are extracted from the MS/MS data. Based on the results of the classification of individual groups, a total number of 40 features including 1, 1, 13, 13, 1, 1, and 10 features from group (1) to (7) are extracted, respectively. For more investigation, 145 possible sister ions from group (7) are used together with another 30 features from group (1) to (6). Therefore, a total number of 175 features will be compared to 4 and 40 features. Shao et al. [13] show that a total number of 20 delta values including common neutral losses with Δ = {17, 18, 28, 34, 35, 36, 44, 45, 46, 64} and isotopic ions with Δ = {− 1, − 2, + 1, + 2} and delta values separated by masses of amino acids including Δ = {57, 63, 71, 87, 97, 99} are meaningful delta values and contribute to better signal and noise peak discrimination. Therefore, the 20 features above from group (7) along with 30 features from other groups

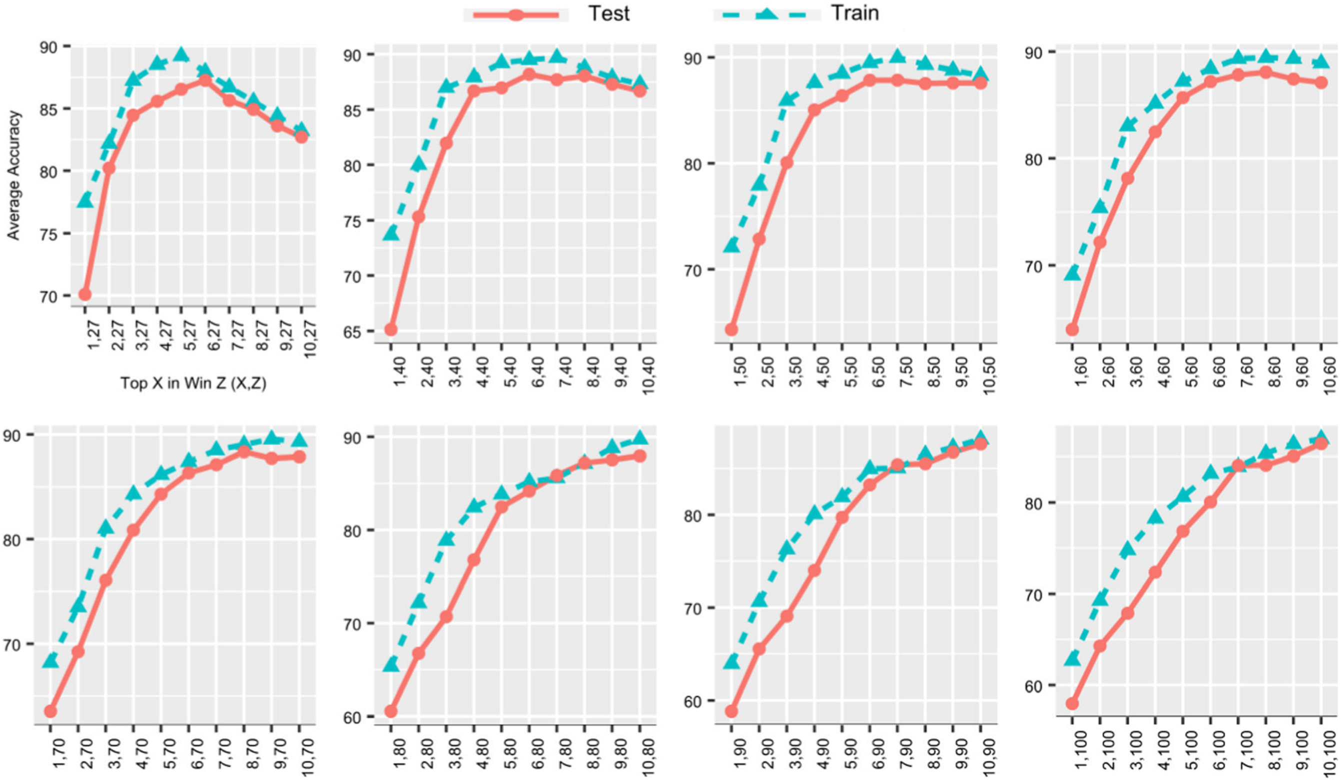


Figure 5. The classification results of GP on labeled sampled dataset in terms of average accuracy using different parameter values for the feature top X intensities in window size Z

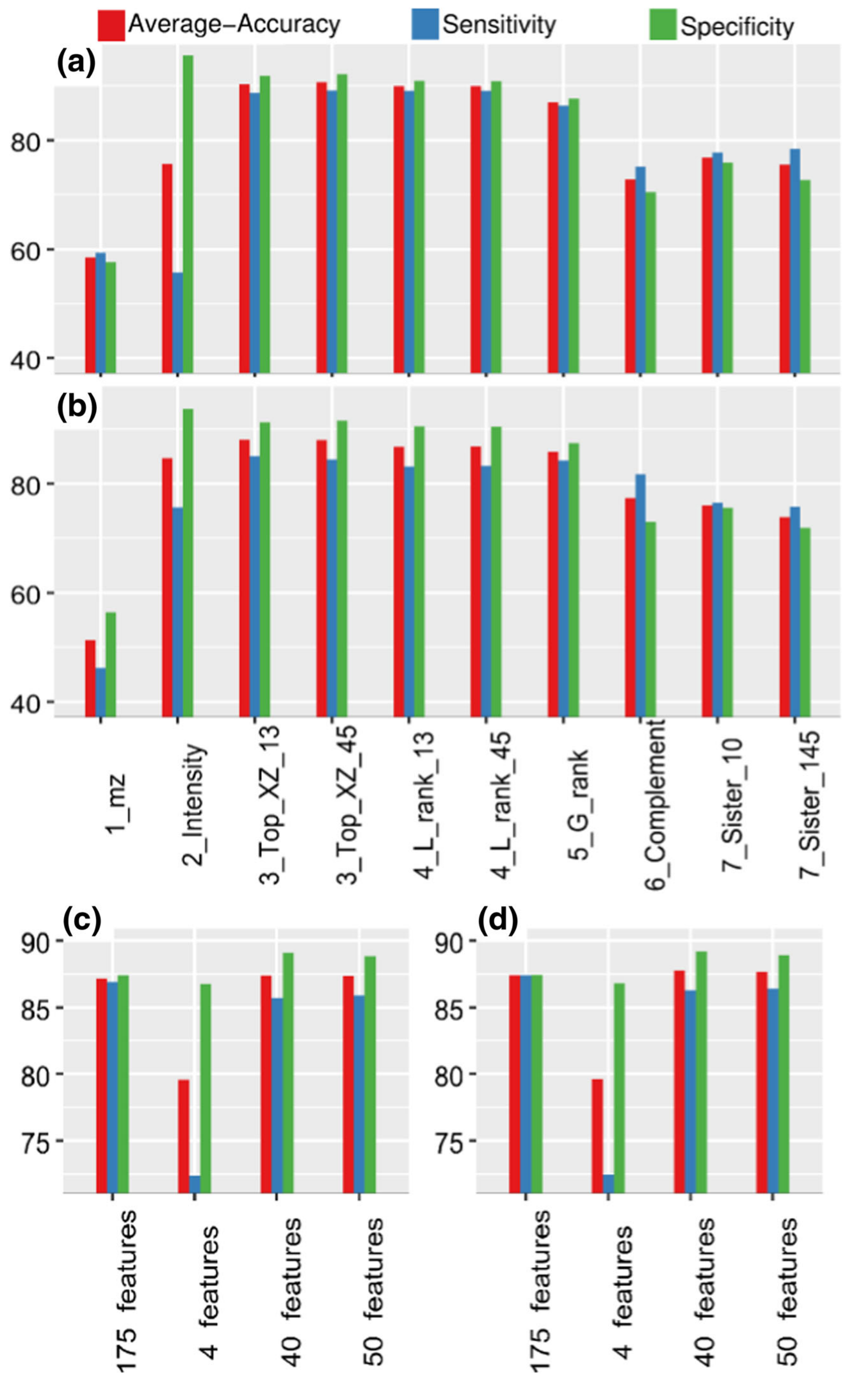


Figure 6. The classification results of GP on labeled sampled dataset. (**a**) Using each group of features individually on the training set. (**b**) The results on test set. (**c**) Using different sets of sister ions on training set. (**d**) The results on test set

(totally 50 features) will be also compared to 4, 40, and 175 features to find out the best set of features aiming at increasing the classification performance using the golden standard dataset (Table 6), which is a large dataset containing thousands of spectra.

Figure 6 c and d show the classification results of GP using different number of features on the training and the test sets of the labeled sampled dataset. It can be seen that there is a huge difference between using only 4 features and using more than 4 on both train and test sets. This is a good indication to motivate using more features. Among the other cases when using 40, 50, and 175 features, it can be seen that using 40 features gives higher classification result on test set compared to using 50 and 175 features. Also, training process takes shorter time when using 40 features compared to 50 and 175 features. In summary, the results show that choosing 40 spectral features are good discriminators to help GP to identify signal and noise peaks. As the main purpose of having a preprocessing method is improving the peptide identification reliability, the next section evaluates the effectiveness of the new GP-based preprocessing method using 40 features and compares the results with the GP system in [14], and the best threshold-based method [14] for the golden standard dataset and un-preprocessed data.

Evaluating the Effectiveness of GP-Based Preprocessing Method Using a De Novo Sequencing Tool for Peptide Identification It was investigated experimentally on full factorial LC-MS/MS benchmark dataset [15] that the threshold value of 100 gets the highest peptide identification rate among the other thresholds [14]. Therefore, the threshold value of 100 is used as a benchmark method. The preprocessed data is then submitted to PEAKS for peptide identification. Also, the test set without applying any preprocessing method is submitted to PEAKS to be a baseline of all the comparisons. Meanwhile, the preprocessed test set by the best GP-method (using 40 features) is also submitted to the peptide identification tool. The results are compared with the existing GP method as well.

Figure 7 shows the results of peptide identification done by PEAKS using different methods to preprocess the test set of the golden standard dataset. All experiments are tested on the same 253,732 MS/MS spectra in the test set of the golden standard dataset. The results are presented in five different ranges of ALC scores. For each ALC range, the rate of identified peptides by PEAKS has been calculated.

Overall, the proposed GP method achieved the highest number of identified peptides by PEAKS compared to the other methods. The proposed GP method helped PEAKS to identify more highly confident peptides with scores $70 < ALC < 99$. Since the method has already identified a large number of peptides in range $70 < ALC < 99$, there are fewer peptides to be identified with low confidence scores in range $55 < ALC < 70$.

For $55 < ALC < 99$, the results of the summation of identification rate for each ALC range, the results show that the new GP method could help PEAKS to find more highly confident peptides rather than the other methods. This method could improve the reliability of peptide identification by 26.6% compared to un-preprocessed data. Comparing the new GP method with the threshold method, GP had 19.3% improvement over the threshold method. Also, the new GP method had 7.1% improvement compared to the existing GP method [14].

In terms of the identification rate, the new GP method could help the peptide identification tool to identify 99.4% of the highly confident peptides, whereas the threshold method only achieved an identification rate of 80.1%.

In summary, the proposed GP method helps PEAKS find more highly confident identified peptides than threshold-based preprocessing method. The reason is that the threshold method ignores those peaks with intensities less than the threshold, resulting in loosing many low intensity signal peaks and keeping a number of high intensity noise peaks. That is one of the disadvantage of threshold method as it ignores the hidden relationship between the peaks and only filters them based on only the intensity feature. The new GP method achieved A-acc of 87.72% and SE of 86.26% on test set of labeled sampled dataset (see the 40 features bar charts in Figure 6 c and d, which means GP could keep a reasonable amount of signal peaks while removing a significant number of noise peaks and this allows GP to improve the results of PEAKS.

Evaluating the Effectiveness of GP-Based Preprocessing Method Using a Database Search Tool for Peptide Identification The same experiments explained in previous section are done using a database search engine, SEQUEST [3], to check the effectiveness of the GP-based preprocessing method. SEQUEST is a dominant benchmark database search tool and reports a confidence score for each peptide spectrum match. A cross-correlation (Xcorr) as a confidence score measures the goodness of fit of experimental spectra to theoretical spectra

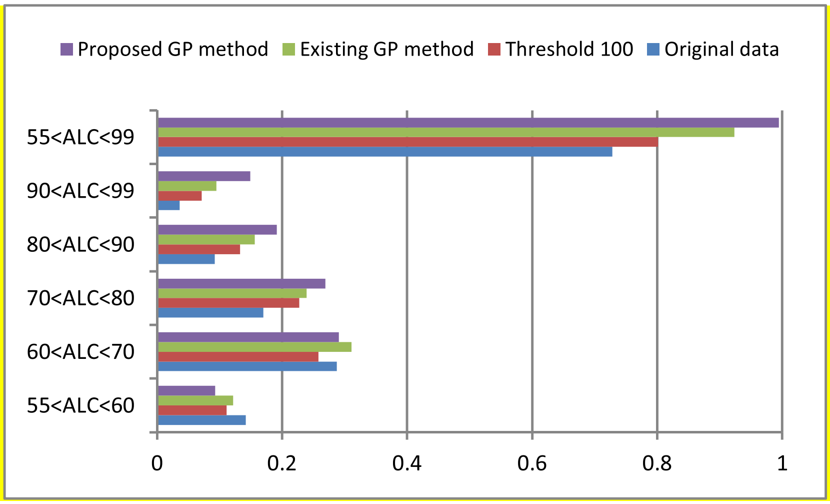


Figure 7. The results of peptide identification by PEAKS using the existing GP method, the new proposed GP method, an intensity-based thresholding method, and golden standard data (un-preprocessed/original data) in different ALC ranges. The ALC score is a confidence score given by PEAKS to each identified peptide and indicates the average ratio of correctly predicted amino acids of a peptide sequence

Table 7. The Statistical Results of Peptide Identification by SEQUEST Using the Existing GP Method, the Newly Proposed GP Method, an Intensity-Based Thresholding Method, and Golden Standard Data (Un-preprocessed/Original Data). The Average and Standard Deviation of Xcorr-Scores for Each Method Is Shown

	Proposed GP method	Existing GP method	Threshold 100	Original data
Xcorr-score	*3.07 ± 0.74*	2.93 ± 0.78	2.91 ± 0.77	2.54 ± 0.63

created from the sequence b- and y-ions. For each spectrum, the peptide candidate with the highest Xcorr-score is known to be a better match.

To compare the results of the proposed GP method with other methods, a statistical unpaired t test with 95 confidence interval is used. Table 7 shows that the result of proposed GP method compared to the results of the other methods is statistically significant (shown in italics in the table). The proposed GP method outperformed the existing GP method and increased the mean of the Xcorr score by 0.53 and 0.16 compared to the best threshold value and the un-preprocessed data, respectively.

In summary, the proposed GP method was also helpful for increasing the reliability of peptide identification done by SEQUEST as a database search engine. By filtering more noise peaks and retaining sufficient signal peaks, it increased the average of confidence scores of identified peptides and reduced the standard deviation of these scores.

Analysis on the Evolved GP Solution Figure 8 shows the best GP-evolved program using 40 features. It can be seen that the GP tree has used features {f_2}, {f_7, f_7, f_{12}, f_{12}, f_{14}}, {f_{16}, f_{16}, f_{19}, f_{21}}, {f_{29}, f_{29}}, {f_{31}, f_{34}} which correspond to the groups (2), (3), (4), (5), and (7) of Table 1. GP revealed that the features "normalized intensity," "top X intensities in Win ± Z," "local intensity rank in Win ± Z," "global rank," and "sister ions" are good discriminators which helped GP to distinguish signal peaks from noise peaks. This is the evidence of why the proposed GP method gets better results compared with the existing method which only used 2 groups of the features above (normalized intensity and global rank). Also, as "sister ions" features have appeared in the evolved GP program, which confirms the result of experiment 1 where it was expected that later, other ion types such as neutral losses will help GP to identify the signal peaks from noise peaks. So it can be seen that sister ion features have been found by GP and helped GP to distinguish the signal peaks from noise peaks. In addition, GP can automatically select 9 features from the 40 features in the 5 groups and achieve the best performance.

Conclusions and Future Work

The goal of this paper was to develop an affective preprocessing method to filter noise peaks and identify the signal peaks for improving the reliability of peptide identification using highly noisy CID spectra. The goal has been successfully achieved by proposing a classification-based preprocessing method using GP to classify peaks to signal or noise peaks. As the MS/MS data is highly imbalanced, average accuracy of true positive rate and true negative rate was used as the fitness function of GP, and this helped GP not be biased towards the accuracy of the majority class containing noise peaks. Meanwhile, a set of suitable spectral features based on the CID fragmentation rules was extracted from the data. With its tree-based representation, feature selection was implicitly applied during the evolutionary process to GP and the analysis of a GP model revealed the important spectral features that have better discrimination ability. A suitable golden standard dataset containing thousands of MS/MS spectra was created and used as the training set of the GP system. The experiments showed that

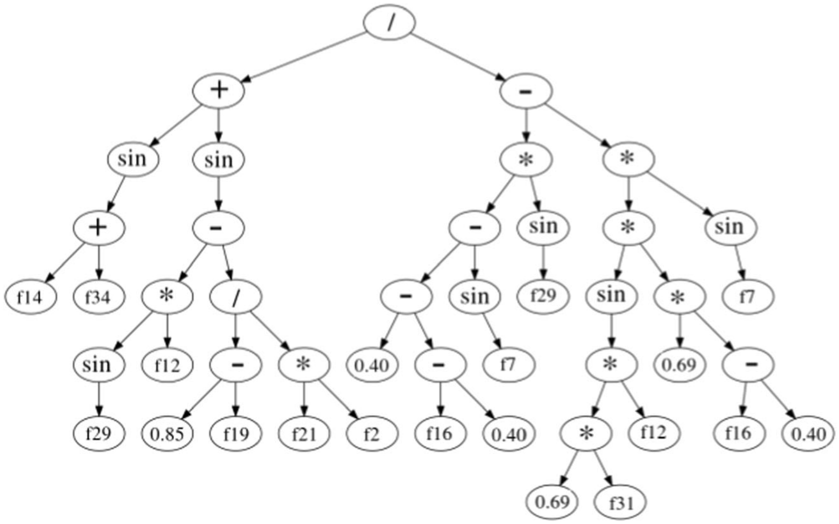


Figure 8. The best GP evolved program using 40 features

the GP-based preprocessing method improved the reliability of peptide identification and increased the identification rate of the de novo peptide identification tool by 26.6% compared to the un-preprocessed data and 19.3% over the threshold-based method. Moreover, the results of the database search tool using the data preprocessed by GP were statistically significant compared to the un-preprocessed data and the best threshold-based method.

As the GP preprocessing method showed the promising results in improving the reliability of peptide identification, in our next work, we will apply this method prior to our new de novo sequencing method on different MS/MS spectra datasets to check how GP can help the de novo sequencing method find the most likely peptide sequence for the given spectrum.

Acknowledgements

This work was supported in part by the Marsden Fund of New Zealand Government under contracts VUW1209, VUW1509, and VUW1615; Huawei Industry Fund E2880/3663; and the University Research Fund at Victoria University of Wellington 209862/3580 and 213150/3662.

References

1. Pandey, A., Mann, M.: Proteomics to study genes and genomes. Nature. **405**(6788), 837 (2000)
2. Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. **20**(18), 3551–3567 (1999)
3. Eng, J.K., McCormack, A.L., Yates, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. **5**(11), 976–989 (1994)
4. Ma, B.: Challenges in computational analysis of mass spectrometry data for proteomics. J. Comput. Sci. Technol. **25**(1), 107–123 (2010)
5. Mann, M., Jensen, O.N.: Proteomic analysis of post-translational modifications. Nat. Biotechnol. **21**(3), 255 (2003)
6. Yu, F., Li, N., Yu, W.: PIPI: PTM-invariant peptide identification using coding method. J. Proteome Res. **15**(12), 4423–4435 (2016)
7. Renard, B.Y., Kirchner, M., Monigatti, F., Ivanov, A.R., Rappsilber, J., Winter, D., Steen, H.: When less can yield more–computational preprocessing of MS/MS spectra for peptide identification. J. Proteome. **9**(21), 4978–4984 (2009)
8. Hoopmann, M.R., Finney, G.L., MacCoss, M.J.: High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Anal. Chem. **79**(15), 5620–5632 (2007)
9. Kwon, D., Vannucci, M., Song, J.J., Jeong, J., Pfeiffer, R.M.: A novel wavelet-based thresholding method for the pre-processing of mass spectrometry data that accounts for heterogeneous noise. J. Proteome. **8**(15), 3019–3029 (2008)
10. Zhou, C., Bowler, L.D., Feng, J.: A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. BMC Bioinform. **9**(1), 325 (2008)
11. Cleveland, J.P., Rose, J.R.: Identification of b−/y-ions in MS/MS spectra using a two stage neural network. Proteome Sci. **11**(1), S4 (2013)
12. Ma, B.: Novor: real-time peptide de novo sequencing software. J. Am. Soc. Mass Spectrom. **26**(11), 1885–1894 (2015)
13. Shao, W., Lam, H.: Denoising peptide tandem mass spectra for spectral libraries: a Bayesian approach. J. Proteome Res. **12**(7), 3223–3232 (2013)
14. Azari, S., Zhang, M., Xue, B., Peng, L.: Genetic programming for preprocessing tandem mass spectra to improve the reliability of peptide identification. IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, Brazil, 1-8 (2018). doi: 10.1109/CEC.2018.8477810
15. Wessels, H.J., Bloemberg, T.G., van Dael, M., Wehrens, R., Buydens, L.M., van den Heuvel, L.P., Gloerich, J.: A comprehensive full factorial LC-MS/MS proteomics benchmark data set. Proteomics. **12**(14), 2276–2281 (2012)
16. Smits, G.F., Kotanchek: Genetic programming theory and practice II. Springer, Boston (2005)
17. Simon, D.: Evolutionary Optimization Algorithms: Biologically-Inspired and Population-Based Approaches to Computer Intelligence. John Wiley & Sons. 141-179 (2013)
18. Espejo, P.G., Ventura, S., Herrera, F.: A survey on the application of genetic programming to classification. IEEE Trans. Syst. Man Cybern. C. Cybern. **40**(2), 121–144 (2010)
19. Bhowan, U., Johnston, M., Zhang, M., Yao, X.: Evolving diverse ensembles using genetic programming for classification with unbalanced data. IEEE Trans. Evol. Comput. **17**(3), 368–386 (2013)
20. Ahmed, S., Zhang, M., Peng, L.: Improving feature ranking for biomarker discovery in proteomics mass spectrometry data using genetic programming. Connect. Sci. **26**(3), 215–243 (2014)
21. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G.: PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. **17**(20), 2337–2342 (2003)
22. Wysocki, V.H., Cheng, G., Zhang, Q., Herrmann, K.A., Beardsley, R.L., Hilderbrand, A.E.: Peptide Fragmentation Overview. In Principles of Mass Spectrometry Applied to Biomolecules. Wiley. Hoboken, New Jersey. 277–300 (2006)
23. Koza, J. R.: Genetic programming as a means for programming computers by natural selection. Stat Comput. **4,** 87-112 (1994). https://doi.org/10.1007/BF00175355
24. White, D.R.: Software review: the ECJ toolkit. Genet. Program Evolvable Mach. **13**(1), 65–67 (2012)
25. Maglott, D.R., Katz, K.S., Sicotte, H., Pruitt, K.D.: NCBI's LocusLink and RefSeq. Nucleic Acids Res. **28**(1), 126–128 (2000)
26. Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J.: The complete genome sequence of Escherichia coli K-12. Science. **277**(5331), 1453–1462 (1997)
27. Han, Y., Ma, B., Zhang, K.: Spider: software for protein identification from sequence tags with de novo sequencing error. J. Bioinforma. Comput. Biol. **3**(03), 697–716 (2005)
28. Bioinformatics Solutions Inc. http://www.bioinfor.com. Accessed: 2018-09-12
29. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., et al.: Open mass spectrometry search algorithm. J. Proteome Res. **3**(5), 958–964 (2004)