Mining Mass Spectrometry Database Search Results—A Rough Set Approach

Jianwen Fang[1,⋆] and Jerzy W. Grzymala-Busse[2]

[1] Bioinformatics Core Facility
and
Information and Telecommunication Technology Center
University of Kansas, Lawrence, KS 66045, USA
jwfang@ku.edu
[2] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
and
Institute of Computer Science Polish Academy of Sciences, 01-237 Warsaw, Poland
jerzy@ku.edu
http://lightning.eecs.ku.edu/index.html

Abstract. This paper reports results of experiments on mass spectrometry database search results produced by Keller *et al.* This data set describes human proteins. Data mining was conducted using the LERS system. First, the data set was discretized by a cluster analysis algorithm based on agglomerative approach. Then the basic rule set was induced by the LEM2 algorithm. Finally, the rule set was refined using changing rule strength methodology and truncation of the rule set. Our results reach the level of sensitivity and specificity of competing methods. However, our results are explainable since they are in a form of rules and, additionally, we can interpret the role of important features.

1 Introduction

With the advance of soft ionization technologies of electrospray (ES) and matrix-assisted laser desorption ionization (MALDI), tandem mass spectrometry (MS/-MS) with database search has emerged as the method of choice for the identification of proteins in high-throughput proteomics studies. Such an approach usually starts with protein separation using 2D-gel or other technologies. The isolated proteins are then digested to peptides using proteases such as trypsin. The resulting peptides are fragmented and ionized using either ES or MALDI technology. The recorded mass spectra are compared to theoretical ones computed from all possible peptides obtained from a protein sequence database using database search software such as SEQUEST [16], Mascot [14], ProteinProspector [3] and X!Tandem [4]. The spectra are then assigned to peptides that best match theoretical spectra. Most of these programs use scores to rank the candidate peptides

⋆ This research has been partially supported by the K-INBRE Bioinformatics Core, NIH grant P20 RR016475.

M. Kryszkiewicz et al. (Eds.): RSEISP 2007, LNAI 4585, pp. 340–349, 2007.

that indicate the degree of agreement between spectra and assigned peptides. A validation procedure is generally required to discriminate false positives in the assigned peptides due to the imperfect nature of these search algorithms. This can be done by manual inspection of an expert or by applying empirical filtering criteria based on database search scores and properties of the assigned peptides, such as the number of tryptic termini. However, the manual validation is prohibitively time-consuming when the database is large and the filtering criteria are not reliable and may miss a large number of true positives. We have found that it is common to miss 30–50% of true positives in the tests as presented in Table 1.

Table 1. The performance of conventional filtering approaches, where *charge* denotes peptide charge

Filtering method	Sensitivity	Specificity
XCorr $\geq$ 2, ΔCn $\geq$ 0.1, SpRank $\leq$ 50, NTT = 2	0.567	0.99844
XCorr $\geq$ 2, ΔCn $\geq$ 0.1, SpRank $\leq$ 50, NTT $\geq$ 1	0.732	0.99290
charge = +1, XCorr $\geq$ 1.5, NTT = 2 OR charge = +2 OR charge = +3, XCorr $\geq$ 2.0, NTT = 2	0.572	0.99796
ΔCn > 0.1 AND (charge = +1, XCorr $\geq$ 1.9, NTT = 2 OR (charge = +2 AND (XCorr $\geq$ 3 OR 2.2 $\leq$ XCorr $\leq$ 3.0, NTT $\geq$ 1)) OR charge = +3: XCorr $\geq$ 3.75, NTT $\geq$1)	0.641	0.99514
ΔCn $\geq$ 0.08 AND (charge = +1, XCorr $\geq$ 1.8 OR charge = +2, XCorr $\geq$ 2.5 OR charge = +3, XCorr $\geq$ 3.5)	0.555	0.99718
ΔCn $\geq$ 0.1 AND (charge = +1, XCorr $\geq$ 1.9, NTT = 2 OR charge = +2, XCorr $\geq$ 2.2, NTT = 1 OR charge = +3, XCorr $\geq$ 3.75, NTT = 1)	0.567	0.99825
ΔCn $\geq$ 0.1, SpRank $\leq$ 50, NTT $\geq$ 1, AND (charge = +1 not included OR charge = +2, XCorr $\geq$ 2.0 OR charge = +3, XCorr $\geq$ 2.5)	0.712	0.99494

In the past several years there have been several attempts to develop software tools using statistical and machine learning algorithms to validate database search hits and consequently improve the results [1,11,15]. Keller *et al.* were among the first to use these approaches to classify the results of SEQUEST searches [11]. They formulated a new metric based on SEQUEST scores that

takes into consideration the length of peptide and penalizes lower ranker and poor mass accuracy. Anderson *et al.* used Support Vector Machine (SVM), a powerful machine learning algorithm, to classify SEQUEST peptide assignment as correct and incorrect, also based on SEQUEST scores [1]. They found that SVM yielded fewer false positives and false negatives comparing to conventional cutoff approaches. Very recently, Ulintz *et al.* used SVM, boosting and Random Forest (RF) to classify MS/MS database search results using SEQUEST and Spectrum Mill, a search engine based on ProteinProspector algorithms [15]. All three algorithms improved sensitivity and specificity considerably over conventional cutoff approaches. While all these approaches delivered better performance than conventional filtering approaches, they failed to provide details how the improvements were achieved, as all methods used in previous studies belong to "black-box" approaches. In this study, we sought to develop interpretable classifiers based on rough set theory. The classifiers resulted in rules that can be readily examined by biomedical researchers to further improve database search engines.

2 Data Set

The original experimental dataset was generated by Keller *et al.* as described in [11]. This dataset was also used by Ulintz *et al.* in their data validation studies [15]. In brief, these data were generated in a ThermoFinnigan ion trap mass spectrometer from twenty-two different LC/MS/MS runs on mixtures of eighteen proteins mixed in varying concentrations. Overall 37044 spectra were generated in the experiments. These spectra were then searched by SEQUEST against a protein database that was composed from human protein database with eighteen additional known proteins. Only top-scoring peptides were retained in the database search. Peptides matching the known eighteen proteins were considered as true positives and the remaining top hits were negatives. For direct comparison, we retained the same division of the dataset into training and test datasets as in [15]. We also used the fifteen descriptive features as in [15], see Table 2.

Usually, in the medical field, the problem is to diagnose a specific disease, where all cases affected by the disease are defined as elements of the primary class. Any subset of the set of all cases, defined by the same value of the decision is called a *class* (or *concept*). All remaining cases are defined as elements of a secondary class (healthy patients). Diagnosis is characterized by *sensitivity* (the conditional probability of the set of correctly diagnosed cases from the primary class given the primary class) and by *specificity* (the conditional probability of the set of correctly diagnosed cases from the secondary class given the secondary class). Thus the sensitivity is the ratio of the number of true positives to the sum of the numbers of true positives and false negatives, while specificity is the ratio of the number of true negatives to the sum of the numbers of true negatives and false positives.

Table 2. Descriptive features used in the study

Feature name	Description
Delta	Parent ion mass error
Charge	Parent ion charge
Intensity	Normalized intensity of the peaks
Length	Length of the peptide
Matching peptide	The number of peptides matching the parent ion mass within the mass tolerance
Sp	Preliminary score
SpRank	Rank based on Sp
ΔCn	Difference in normalized correlation scores between next-best and best hits
XCorr	Cross-correlation score
ratio	Fraction of experimental ions matched with the theoretical ions
N_pro	Number of prolines
N_arg	Number of arginines
C_term	C-terminal residue
NTT	Number of tryptic termini
PMF	Proton mobility factor

Our training data set contained 25931 cases, with 1930 cases being the primary class and remaining 24001 cases being the secondary class. The testing data set contained 11113 cases, distributed into 827 cases from the primary class and 10286 cases from the secondary class.

3 Discretization, Rule Induction and Classification

All numerical attributes were discretized before rule induction, i.e., numerical values of these attributes were converted into symbolic. For our experiments we selected a discretization based on cluster analysis. First clusters were formed, using bottom-up (agglomerative) approach. The process was continued until each elementary set, defined by all attributes, was contained in some concept or all attributes defined the same indiscernibility relation as for the original data set. Both ideas, of the *elementary set* and *indiscernibility relation*, are taken from rough set theory [12, 13]. Then the clusters were projected on numerical attributes and initial intervals were created. Finally, these intervals were merged together using the same criterion to stop as in the process of forming clusters.

For rule induction, classification, and validation we used the data mining system LERS (Learning from Examples based on Rough Sets) [5, 6]. After discretization, in the next step of processing the input data file, LERS checks if the input data file is consistent. If the input data file is inconsistent, LERS computes lower and upper approximations of all classes. The ideas of *lower* and *upper approximations* are fundamental for rough set theory [12,13].

In general, LERS uses two different approaches to rule induction: one is used in machine learning, the other in knowledge acquisition. In machine learning, or more specifically, in learning from cases (examples), the usual task is to learn the smallest set of minimal rules, describing the class. To accomplish this goal LERS uses two algorithms: LEM1 and LEM2 (LEM1 and LEM2 stand for Learning from Examples Module, version 1 and 2, respectively). In our experiments we used only LEM2 algorithm since, in general, LEM2 induces simpler and more accurate rule sets.

The classification system of LERS is a modification of the *bucket brigade algorithm* [2,10]. The decision to which concept a case belongs is made on the basis of two factors: strength and support. They are defined as follows: *strength* is the total number of cases correctly classified by the rule during training. The second factor, *support*, is defined as the sum of strengths for all matching rules from the concept. The concept C for which the support, i.e., the following expression

$$\sum_{matching\ rules\ R\ describing\ C} Strength(R)$$

is the largest is the winner and the case is classified as being a member of C.

In the classification system of LERS, if complete matching is impossible, all partially matching rules are identified. These are rules with at least one attribute-value pair matching the corresponding attribute-value pair of a case. For any partially matching rule R, the additional factor, called $Matching\_factor\ (R)$, is computed. Matching_factor (R) is defined as the ratio of the number of matched attribute-value pairs of R with a case to the total number of attribute-value pairs of R. In partial matching, the concept C for which the following expression is the largest

$$\sum_{\substack{partially\ matching \\ rules\ R\ describing\ C}} Matching\_factor(R) * Strength(R)$$

is the winner and the case is classified as being a member of C.

Every rule induced by LERS is preceded by three numbers: the total number of attribute-value pairs on the left-hand side of the rule, strength, and the rule domain size, i.e., the total number of training cases matching the left-hand side of the rule.

4 Postprocessing of Rules

Once rule sets were induced we used two different postprocessing techniques applied to these rule sets. The first technique was called *increasing rule strengths* [7, 8]. This technique is used for imbalanced data sets, that is, data sets with different class sizes. Our data set was imbalanced, the total size of primary class was much smaller than the total size of secondary class. In such data, during

classification of unseen cases, rules matching a case and voting for the primary classes are outvoted by rules voting for the bigger, secondary classes. Thus the diagnosis of a primary classes is poor and the resulting classification system would be rejected by diagnosticians.

Therefore it is necessary to decrease the error rates for the primary class. Since the data set is imbalanced, the simplest idea is to add cases to the primary class in the data set, e.g., by adding duplicates of the available cases. The total number of training cases will increase, hence the total running time of the rule induction system will also increase. Adding duplicates will not change the knowledge hidden in the original data set, but it may create a balanced data set so that the average rule set strength for both classes will be approximately equal. The same effect may be accomplished by increasing the average rule strength for the primary class. In our research we selected the optimal rule set by multiplying the rule strength for all rules describing the primary class by the same real number called a *rule strength multiplier*. In general, the error rates for the primary classes decrease with the increase of the rule strength multiplier. At the same time, the error rates for the secondary classes increase.

The second mechanism to increase the conditional probabilities for primary class was *rule truncation*, a method of reducing the rule set by deleting weak rules, describing a few training cases, by removing rules with strengths not exceeding some cutoff. The truncation algorithm was already used for diagnosis of melanoma, see, e.g., [8]. By removing weak rules the total number of rules describing the class is reduced. This may result in rules that may not match the cases completely as they would have before the truncation process. However, the LERS classification system is equipped with partial matching. A case may still be very closely related to the correct class and thus may be correctly recognized.

5 Experiments

Our experiments were performed on the training data set (with 25931 cases) discretized by the agglomerative cluster analysis algorithm. A basic rule set was induced from the discretized data set by the LEM2 algorithm. Then we incrementally increased the rule strength multiplier for all rules describing the primary classes, see Table 3. Sensitivity and specificity presented in Table 3 were computed using the testing data set (with 11113 cases). During these experiments the truncation cutoff was not used (all rules participated in classification). Then, with the rule strength multipler equal to 1000, we gradually increased the truncation cutoff, up to 100, for the rule set describing the secondary class. The size of the rule set describing the secondary class decreased from 282 (the original rule set) to 141 (the rule set corresponding to the truncation cutoff equal to 100). During all of our experiments the size of the rule set describing the primary class was always equal to 244. The ROC (Receiver Operating Characteristic) graph, illustrating our experiments, is presented in Figure 1.

Table 3. Performance of rough set models

Strength multiplier	Truncation cutoff	Sensitivity	Specificity
1	0	0.8440	0.99543
20	0	0.8839	0.99217
100	0	0.9117	0.98493
500	0	0.9178	0.98260
1000	0	0.9190	0.98085
1000	5	0.9202	0.97968
1000	20	0.9287	0.97832
1000	50	0.9323	0.97579
1000	100	0.9383	0.96850

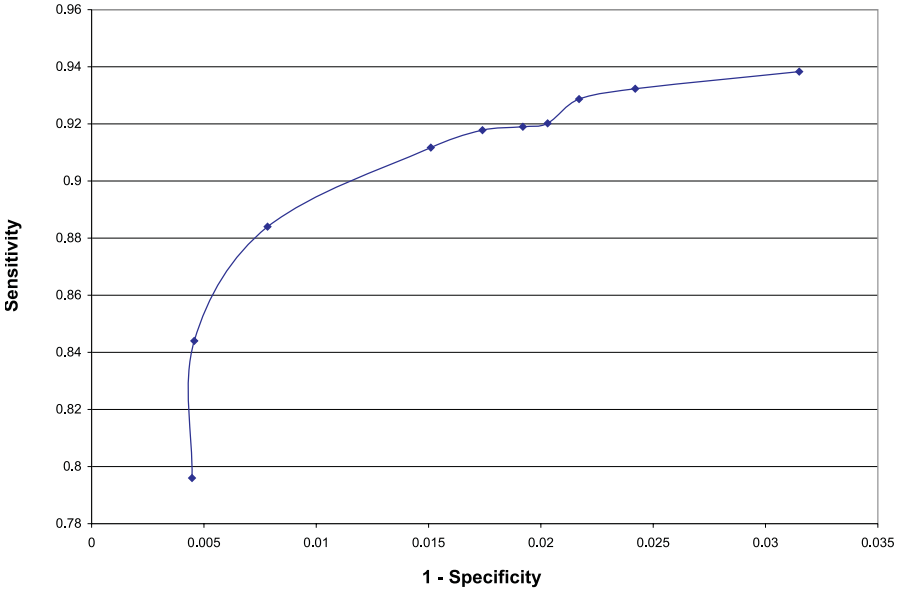

Fig. 1. ROC graph

6 Results and Comparison with Other Approaches

The dataset that was the subject of our experiments was previously analyzed in other studies using various machine learning algorithms [11,15]. For example, Ulintz *et al.* reported that approaches using boosting and random forest achieved a sensitivity of 0.99, PeptideProphet and SVM delivered 97 – 98% sensitivity at a false positive rate of roughly 0.05 [15]. Thus the performance of our approaches is comparable to Ulintz's results as we achieved better false positive rates but poorer sensitivities. Keller *et al.* reported a sensitivity of 89% with an error of

2.5% [11]. Although a direct comparison to this study is not applicable because Keller *et al.* used different division of training and test datasets, it appears that our model is competitive.

7 Interpretation of the Decision Rules

An advantage of white-box approaches such as rough set theory over "black-box" methods is that the detailed knowledge of the classification process is available for better understanding the problem under study. In the present study, the classification rules discovered by our classifiers reveal several important observations leading to better understanding the chemistry underlying the molecule fragmentation and ionization. For example, the mobile proton factor (MPF) was discovered as a very useful indicator. A single rule involving only two features can eliminate approximately 40% of true negatives without error:

(PMF, 0.699..5.5) & (C_term, others) –> (label, -1)

The PMF is calculated as:

$$\frac{R + 0.8 * K + 0.5 * H}{charge}$$

where R is the number of arginine, K is the number of lysine, and H stands for the number of histidine. Charge means the charge on the parent peptide. Although it was known that a smaller value of PMF indicates higher protein mobility [15], it was unclear the degree that PMF would affect the peptide detection using MS/MS technologies. From our results, it seems that PMF is particularly useful to eliminate peptides with a terminal residue other than arginine and lysine. It is worth to note that the rule does not use any SEQUEST score.

NTT (the number of tryptic terminals) is important since the peptides are the products from tryptic digestions. It measures whether the peptide is fully tryptic (NTT = 2), partially tryptic (NTT = 1), or non-tryptic (NTT = 0). However, the NTT of a fully tryptic terminal peptide can be equal to one. NTT was found as the most important attribute in Ulintz's study [15]. A higher NTT is a strong indication of a true positive; however, the NTT of a small portion of true positives is either 0 or 1. For example, this type of peptides accounts for about one quarter of the true positives in our dataset. Thus improvement in this type of peptide identification will significantly increase the sensitivity and specificity. We found that the following single rule correctly classifies approximately 40% of these partially tryptic and non-tryptic peptides. Thus peptides with lower NTT but higher XCorr and ΔCn are likely true positives.

(XCorr, 3.4218..7.2792) & (ΔCn, 0.2362..0.5565) & (NTT, 0..1.5) –> (label, 1)

Most of the rules discovered in our study involve one or more features that are not SEQUEST scores. These features are either peptide physicochemical properties (e.g., MPF, Length, etc.) or protein sequence environment (e.g., NTT). The results further confirm the conclusion in our recent study that these properties can be used to improve data validation models [5].

8 Conclusions

We have proposed a rough set based approach to validate MS/MS database search results. The performance of our approach is comparable to competing methods. However, some important rules discovered in this study may lead to better understanding of the chemistry underlying the molecule fragmenttion and ionization. In addition, these rules may be used in the development of novel mass spectrometry database search engines.

References

1. Anderson, D.C., Li, W., Payan, D.G., W.S., N.: A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J. Proteome. Res. 2, 137–146 (2003)
2. Booker, L.B., Goldberg, D.E., Holland, J.F.: Classifier systems and genetic algorithms. In: Carbonell, J.G. (ed.) Machine Learning. Paradigms and Methods, pp. 235–282. MIT Press, Menlo Park, CA (1990)
3. Clauser, K.R., Baker, P.R., Burlingame, A.L.: Role of accurate mass measurement (+/– 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal. Chem. 71, 2871–2882 (1999)
4. Craig, R., Ronald, C., Beavis, R.C.: TANDEM: matching proteins with mass spectra. Bioinformatics 20, 1466–1467 (2004)
5. Fang, J.W., Dong, Y.H., Williams, T.D., Lushington, G.H.: Classification of MS/MS Peptide Identifications and Its Application in Data Validation (Submitted)
6. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundamenta Informaticae 31, 27–39 (1997)
7. Grzymala-Busse, J.W.: MLEM2: A new algorithm for rule induction from imperfect data. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2002, Annecy, France, July 1–5, 2002, 243–250 (2002)
8. Grzymala-Busse, J.W., Hippe, Z.S.: Postprocessing of rule sets induced from a melanoma data set. In: Proceedings of the COMPSAC 2002, 26th Annual International Conference on Computer Software and Applications, Oxford, England, August 26–29, 2002, pp. 1146–1151 (2002)
9. Grzymala-Busse, J.W., Goodwin, L.K., Grzymala-Busse, W.J., Zheng, X.: An approach to imbalanced data sets based on changing rule strength. In: Learning from Imblanced Data Sets, AAI Workshop at the 17th Conference on AI, AAAI-2000, Austin, TX, July 30–31, 2000, pp. 69–74 (2000)
10. Holland, J.H., Holyoak, K.J., Nisbett, R.E.: Induction. Processes of Inference, Learning, and Discovery. MIT Press, Menlo Park, CA (1986)
11. Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R.: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 74, 5383–5392 (2002)
12. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
13. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)

14. 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, 3551–3567 (1999)
15. Ulintz, P.J., Zhu, J., Qin, Z.S., Andrews, P.C.: Improved classification of mass spectrometry database Search results using newer machine learning approaches. Mol. Cell. Proteomics 5, 497–509 (2006)
16. Yates III, J.R., Eng, J.K., McCormack, A.L.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in the protein database. J. Am. Soc. Mass. Spectrom. 5, 976–989 (1994)