Using a Probable Time Window for Efficient Pattern Mining in a Receptor Database[⋆]

Edgar H. de Graaf and Walter A. Kosters

Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
{edegraaf,kosters}@liacs.nl

Abstract. The analysis of sequences is one of the major research areas of bio-informatics. Inspired by this research, we investigate the discovery of sequential patterns for use in classification. We will define variations of a fit function that enables us to tell if one pattern is "better" than another. Furthermore we will show how domain knowledge can be used for faster discovery of better sequential patterns in specific types of databases, in our case a receptor database.

1 Introduction

Sequence analysis has many application areas, e.g., protein sequence analysis and customer behavior analysis. We investigate extraction of features for protein sequence classification where features are *sequential patterns*: ordered lists of items (for proteins the items are amino acids). As a motivating example, we would like to know if a protein sequence, an ordered list of amino acids, belongs to the Olfactory family or not, where the Olfactory family is a group of proteins than deals with smell. We focus on a special group of proteins called GPCRs. These G-protein-coupled receptors (GPCRs) play fundamental roles in regulating the activity of virtually every body cell [17]. Usually classification is done unsupervised using alignment, however in the case of GPCRs this turned out to be difficult. Fortunately, we know for some protein sequences whether they are of the Olfactory family or not. These sequences can thus be divided into two disjoint classes: Olfactory and No-olfactory, and from these classes we can extract sequential patterns to be used as attributes in a classification algorithm (as is being proposed in [12]). The question we try to answer in this paper is which sequential patterns are *best* used as features/patterns? And how can domain knowledge be used to improve the search for such patterns?

Classification based on sequential patterns is also applicable in many other areas. For example, in the case of customer behavior analysis, we might want to characterize groups of clients based on sequential patterns in their behavior.

The "best" sequential patterns are discovered through a function that judges patterns. In Section 2 we will discuss different instances of this function and

⋆ This research is carried out within the Netherlands Organization for Scientific Research (NWO) MISTA Project (grant no. 612.066.304).

select one for our purposes. Section 3 adapts the PrefixSpan algorithm of [15] to deal with this function. In addition, a pruning strategy is introduced in Section 4, increasing efficiency by first searching in a certain area of the sequence, the *probable time window.* Section 4 also describes how preferring small patterns can further increase classification performance. The effectiveness of these improvements will be shown in Section 5. Earlier work on this subject has been presented at BNAIC2005 [7].

Related work. Our algorithms will be based on the pattern growth approach called PrefixSpan proposed in [15]. Classification by means of patterns has been done before but not so much in the sequence domain. We now mention related work in the non-sequence domain. Apriori-C [9] constructs classification rules by extending the Apriori algorithm [2, 3]. Apriori-C discovers a large number of rules from which a fixed number of rules with the highest support are selected. Apriori-SD [10] solves the problem of selecting the right rules with *subgroup discovery.* This algorithm selects a subgroup of rules by calculating their *weighted relative accuracy.* This means that the probability of a pattern occurring in a class is compared with the probability of its occurrence outside the class. This is weighted with the probability of a class. Most class association rule mining algorithms work with unordered sets of items frequently occurring together in *item sets.* Classification with association rules is presented in [13] and [14]. Furthermore CorClass [18] describes an algorithm that also works with item sets. It introduces a new method of pruning. Specialized rules are only added if the upper bound of its correlation is higher than the minimal correlation of k rules. In our work we use a similar method of pruning. Much work has been done in the field of molecular feature mining, e.g., the MolFea algorithm described in [11]. MolFea employs a level-wise version space algorithm to discover those molecule fragments often occurring in one data set and less often in another. Finally some other researchers try to use domain knowledge to speed up the search for frequent patterns, e.g., the Carpenter algorithm presented in [5]. In this work the authors perform row enumeration instead of the standard column enumeration done in Apriori-like algorithms. This is done because biological data sets often have many columns/items and only a few rows.

2 The Maximal Discriminating Patterns

One would like to select the best patterns for use as attributes in a classification algorithm. But how can we tell if one pattern is better than the other? In this section we will first explain the notion of support and why it is less useful for selecting the best pattern. Next we introduce the notion of confidence which will give more useful patterns, but it also has disadvantages. Finally we will discuss and motivate so-called maximal discriminating patterns, enabling us to have patterns specific to one class, but without the disadvantages of confidence.

Assume given a database D with $D = D_1 \cup D_2 \cup \ldots \cup D_c$, with c classes. The D_i's $(1 \leq i \leq c)$ are mutually disjoint and not empty.

Each record in the database is a non-empty finite *sequence* (i.e., an ordered list) of items from the set $\Sigma = \{\texttt{A},\texttt{B},\texttt{C},\ldots\}$, e.g., $(\texttt{C},\texttt{B},\texttt{G},\texttt{A},\texttt{A},\texttt{A},\texttt{C},\texttt{B})$. Now fit_0 is defined as support (as used in association rule mining algorithms like Apriori [3]), because support can be seen as a measure of how well a pattern fits the data. Commonly a sequence d is said to support a *pattern* s if the pattern is contained (in the set sense) in the sequence:

$$\mathit{supp}_0(s,d) = \begin{cases} 1 & \text{if for all } i \ (1 \le i \le k) \text{ there is a } j \ (1 \le j \le \ell) \text{ with } s_i = d_j; \\ 0 & \text{otherwise,} \end{cases}$$

for $s = (s_1, s_2, \ldots, s_k)$ and $d = (d_1, d_2, \ldots, d_\ell)$. This means that s is a *subset* of d. We then can define fit_0:

$$\mathit{fit}_0(s, D_i) = \frac{1}{|D_i|} \sum_{d \in D_i} \mathit{supp}_0(s,d)$$

$(1 \le i \le c)$, where s is a pattern.

We now specialize support to sequences. A sequence $d = (d_1, d_2, \ldots, d_m)$ is called a *super-sequence* of a sequence $s = (s_1, s_2, \ldots, s_k)$ if $k \le m$ and for each s_i $(1 \le i \le k)$ there is a d_{j_i} $(1 \le j_i \le m)$ with $s_i = d_{j_i}$ and $j_{i-1} < j_i$ $(i > 1)$. We denote this with $s \prec d$. The sequence s is called a *sub-sequence* of d. This defines *sequential patterns* on sequences of items. (Another definition of sequential patterns was given by Agrawal et al. in [3], in which they define sequential patterns on sequences of item sets). We now let

$$\mathit{supp}_1(s,d) = \begin{cases} 1 & \text{if } s \prec d; \\ 0 & \text{otherwise,} \end{cases}$$

and define fit_1 in the same way as fit_0 was defined using supp_0.

Now fit_1 or fit_0 by itself is not useful for selection of features for classification. One of the patterns of size 1 will always have the highest fit and these small patterns are probably often present in more than one D_i. Thus the presence of such a pattern will not give a good distinction between classes.

The next most logical step is to use confidence to select the best patterns. The patterns x_r $(1 \le r \le c)$, one for each class, are then chosen to maximize *confidence* $(\mathit{fit}_1(x_r, D_r)|D_r|)/(\mathit{fit}_1(x_r, D)|D|)$. The class t of sequence s is the t $(1 \le t \le c)$ where $x_t \prec s$. If more than one t is possible we select based on the highest confidence. One is selected at random if more than one class t has a pattern with the highest confidence. If there is no t where $x_t \prec s$ then the sequence could be said to be "undecided".

A problem is that we only pick one pattern per class. This is plausible if a family of a sequence is only decided by one sequence of features. However, it is often the case that the class of a sequence is decided by multiple patterns. Moreover there can be constraints on the pattern. This means that the class deciding pattern x_t *with* the constraint is not necessarily equal to the x_t *without* the constraint. As a consequence it is usually possible to find a combination of patterns with a better classification performance. Finally it is possible that a single sequential pattern x_t is equal for two or more classes, and as a consequence a classification will be done at random. This problem will occur with a lower probability if we use multiple patterns for each class.

Another major drawback of the confidence method is that the size of the D_i's seriously influences the classification. E.g., assume we have databases D_1 and D_2. Furthermore assume D_1 contains 500 sequences and D_2 only 100. The pattern p_1 occurs 100 times in D_2 and 60 times in D_1, thus a confidence with respect to D_2 of 0.625. Another pattern p_2 occurs 70 times in D_2 and 10 times in D_1, giving a confidence of 0.875. The pattern p_2 will be used for classification if no other pattern has a higher confidence. However p_1 occurs in every sequence of D_2 and only in a small percentage of the sequences in D_1. One could argue that p_1 should be preferred over p_2.

Therefore we define fit_2, which we use in the sequel. For a pattern s and $1 \leq q, r \leq c$ we define $\delta(s, D_q, D_r) = \mathit{fit}_1(s, D_q) - \mathit{fit}_1(s, D_r)$, and we let $\mathit{fit}_2(s, D_r) = \min\{\delta(s, D_r, D_q) \mid 1 \leq q \leq c \,\wedge\, q \neq r\}$. We then choose patterns x_r $(1 \leq r \leq c)$ with maximal $\mathit{fit}_2(x_r, D_r)$. We can then use them to classify sequences as before, without the drawbacks mentioned above. We will usually find those patterns that are characteristic for one class. With *characteristic* we mean that fit_1 will have a high value in D_t and a lower value in the other $D_i{'}s$, $i \neq t$.

Our new fit has some similarities with the concept of *emerging patterns* presented in [4] and [6]. In order to discover emerging patterns patterns are preferred where the ratio $\mathit{fit}_1(s, D_1)/\mathit{fit}_1(s, D_2)$ is the highest, where D_1 and D_2 are two databases each containing one class of sequences. Bailey et al. [4] further investigate jumping emerging patterns. These are patterns that have a support of zero in D_2 and a non-zero support in D_1. Emerging patterns can also be defined in a way similar to fit_2, but now using $\mathit{fit}_1(s, D_q)/\mathit{fit}_1(s, D_r)$ instead of $\delta(s, D_q, D_r)$. Dong et al. [6] point out that the growth rate measure used by emerging patterns doesn't take into account the coverage, a problem they solve with a score function. However in the case of fit_2 coverage is less of a problem, a pattern with a low $\mathit{fit}_1(s, D_1)$ is less likely to have a high fit_2 value. Also the fit_2 measure allows us to more easily explain and implement the pruning rules that will be discussed in the remainder of this paper.

Classification algorithms usually need a limited number of attributes. In order to classify a sequence s we use a finite number of n sequential patterns $p^t_1, p^t_2, \ldots, p^t_n$ per class t, where $\mathit{fit}_2(p^t_1, D_t) \geq \mathit{fit}_2(p^t_2, D_t) \geq \ldots \geq \mathit{fit}_2(p^t_n, D_t)$ and p^t_n has the n-th highest fit_2 for all possible patterns. These patterns, the so-called *maximal discriminating patterns*, could be used by any classification algorithm when we first convert each sequence to a vector indicating for each pattern if it is contained in the sequence, see [12]. However it is possible that, e.g., p^t_1 is supported by all or most of the sequences supporting p^t_2. Thus p^t_2 might not improve classification. This problem could be solved by removing all sequences containing p^t_1 from D_t. The algorithm for searching the sequence with maximal fit is then again applied to this subset of D_t in order to find p^t_2. In this paper we do not further focus on the precise classification performance, but rather on the discovery of the discriminating patterns. Our algorithm aims at finding the set $P = P^t$ of maximal discriminating patterns.

3 Algorithm without Domain Knowledge

Our pattern search algorithm, coined PrefixTWEAC (Time Window Exploration And Cutting), is based on PrefixSpan. The algorithm does not generate candidates, but it grows patterns from smaller patterns. This principle makes it faster than most Apriori like algorithms [15]. PrefixSpan is a depth first algorithm, which will be explained in more detail in Section 4 when we adapt this algorithm to our current needs (see Algorithms 1 and 2). PrefixSpan as described in [15] searches for those patterns with *support* larger than or equal to a given support threshold *minsupp*, where support is defined as fit_1. The algorithm starts with all frequent sub-sequences of size one. For each sub-sequence a projected database is created. These frequent sub-sequences are extended to all frequent sub-sequences of size two by only looking in the projected database. This *projected database* is a database of pointers to the first item occurring after the current pattern, also called the *prefix*. A sequence is only in the projected database if it contains the prefix. Again for each frequent sub-sequence of size two a corresponding projected database is created. This process continues recursively until no extension is frequent anymore.

PrefixTWEAC (Algorithm 1) is different from PrefixSpan in that it searches for the maximal *fit2* instead of the maximal support *fit1*. The function *fit2* is by definition not anti-monotone (so $fit_2(s_1, D_t) > fit_2(s_2, D_t)$ might happen, where s_1 is a super-sequence of s_2). However the anti-monotone property for fit_1 can still be used in two ways, when looking for the one pattern with maximal fit_2. First of all in PrefixTWEAC we only examine an extended pattern p if $fit_1(p, D_t) \geq minsupp$ where *minsupp* is the support threshold. Secondly p is not further examined if $fit_1(p, D_t) <$ *current n-th maximal fit*, where *current n-th maximal fit* is the current n-th best fit of all patterns found while searching. The value of $fit_2(p, D_t)$ will never become larger than the current n-th maximal fit, because it can at most become $fit_1(p, D_t)$. Note that CorClass uses similar methods to prune [18].

4 Domain Specific Improvements

In the previous section we stated that fit_2 can be used to "prune": certain pattern extensions are not further examined because they can never lead to the maximal fit_2. The faster we get to a large fit_2 for the n-th pattern in $P = P^t$ the better, because all extensions with a lower $fit_1(p, D_t)$ can be pruned. The improved version of PrefixTWEAC will be explained in the sequel.

If we consider protein sequences then pattern discovery might be done faster and/or classification might improve when using certain knowledge about the sequences:

Algorithm 1 The PrefixTWEAC algorithm

```
PrefixTWEACCore(prefix, projected_database)
1. For all items $i$ that can extend the prefix
2.    new_prefix = prefix extended with item $i$
3.    Count $w_1 = \mathit{fit}_1$ in the projected_database$_t$ for new_prefix
4.    Calculate $f_2 = \mathit{fit}_2$ for new_prefix
5.    Create a projected database new_projected_database with new_prefix
6.    Get $\delta_{min}$, the lowest $\mathit{fit}_2$ in $P$
7.    Get $s_{min}$, $\mathit{fit}_1$ corresponding with the lowest $\mathit{fit}_2$ in $P$
8.    if $w_1 \geq \mathit{minsupp}$ and $|P| < n$ then
9.       Add new_prefix to $P$
10.      Call PrefixTWEACCore(new_prefix, new_projected_database)
11.   else if $w_1 \geq \mathit{minsupp}$ and $w_1 \geq \delta_{min}$ then
12.      if $f_2 > \delta_{min}$ or
13.         ($f_2 = \delta_{min}$ and $w_1 > s_{min}$) or
14.         ($f_2 = \delta_{min}$ and $w_1 = s_{min}$ and new_prefix $\prec p_n$) then
15.         Remove $p_n$ from $P$ and add new_prefix to $P$
16.      Call PrefixTWEACCore(new_prefix, new_projected_database)
```

– Protein sequences are sequences of amino acids. Certain parts of such a sequence are shaped like a helix in 3D space. These helices will probably contain most of the maximal fitting sequences since parts outside the helix have more variation in size and content. Patterns (partially) outside the helix are less likely to occur in most members of the protein family.
– Small patterns are preferred. Smaller patterns are less specific and biologists prefer smaller patterns in their analysis.

For certain problems we know the approximate area of important features, e.g., protein sequences should have most of the discriminating patterns in the helix. Also in other problems this might be the case, for example — in the case of customer relations — customers tend to behave differently during the night. These *probable time windows* can easily be defined with an *inclusion vector*. An inclusion vector is a vector $v = (v_1, v_2, \ldots, v_n)$, $v_i \in \{0, 1\}$ $(1 \leq i \leq n)$. This vector will indicate where to search in the first phase of the algorithm, see Algorithm 2. We then let

$$\mathit{supp}_1^{\mathrm{PTW}}(s, d) = \begin{cases} 1 & \text{if } s \prec d/v; \\ 0 & \text{otherwise,} \end{cases}$$

where $(d/v)_i = d_i$ if $v_i = 1$ and \$ otherwise (\$ $\notin \Sigma$), so only positions with nonzero v_i are considered.

First PrefixTWEACExt (Algorihtm 2) is applied to the databases D_t, one at a time, each time starting with an empty $P = P^t$. After using PrefixTWEACExt with the inclusion vector we apply PrefixTWEAC (Algorithm 1) without the vector to the remaining states stored in the state database S.

Algorithm 2 PrefixTWEAC Extended: extension using the probable time window

```
PrefixTWEACExt(prefix, projected_database)
1. For all items $i$ that can extend the prefix
2.    new_prefix = prefix extended with item $i$
3.    Count $fit_1$ for new_prefix:
4.       $w_1 = fit_1$ in the projected_database$_t$ without the inclusion vector, using $supp_1$
5.       $w_2 = fit_1$ in the projected_database$_t$ with the inclusion vector, using $supp_1^{\mathrm{PTW}}$
6.    Calculate $f_2 = fit_2$ for new_prefix (without the inclusion vector)
7.    Create new_projected_database (without using the inclusion vector)
8.    Get $\delta_{min}$, the lowest $fit_2$ in $P$
9.    Get $s_{min}$, $fit_1$ of the lowest $fit_2$ in $P$
10.   if $w_1 \geq minsupp$ and $|P| < n$ then
11.      Add new_prefix to $P$
12.      Call PrefixTWEACExt(new_prefix, new_projected_database)
13.   else if $(w_1 \geq minsupp$ and $w_2 < minsupp)$ or
14.            $(w_2 \geq minsupp$ and $w_1 \geq \delta_{min}$ and $w_2 < \delta_{min})$ then
15.      storeState($S$,new_prefix, new_projected_database)
16.   else if $w_2 \geq minsupp$ and $w_2 \geq \delta_{min}$ then
17.      if $f_2 > \delta_{min}$ or
18.         $(f_2 = \delta_{min}$ and $w_1 > s_{min})$ or
19.         $(f_2 = \delta_{min}$ and $w_1 = s_{min}$ and new_prefix $\prec p_n)$ then
20.         Replace $p_n$ with new_prefix
21.      Call PrefixTWEACExt(new_prefix, new_projected_database)
```

Figure 1 shows an example of the extensions made to a sequence `A`. The dotted lines are extensions that do not have a high enough *fit₁* and *fit₂* inside and outside the probable time window. These extensions and their extensions are pruned. The dashed lines indicate extensions that are currently good enough with regards to the entire sequence only. Finally the solid lines are already good enough when we only count patterns inside the probable time window.

If we prefer small patterns, then we can add a new rules, using so-called *smallest maximal discriminating patterns* to improve classification:

- $fit_1(s, D_r) = 0$ for all r $(1 \leq r \leq n, r \neq t)$. Then fit_2 of the extended patterns will never increase.
- $fit_1(s, D_t) \leq fit_2(p, D_t)$ where both p and s are sequences and s is created by extending p. Then fit_2 of the extended patterns will never be better than the fit_2 of p.

These rules sacrifice some completeness for classification performance; if extensions do not improve a smaller pattern then they are not always explored further. These pruning rules will not lower classification performance because they leave out only non-improving extensions. Rather the classification is expected to improve because the set of patterns will contain less small variations of the same

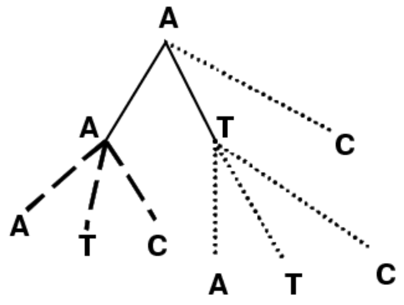

Fig. 1. Extending the single item sequence A

pattern. We will from now on abbreviate the use of these rules with SP or "small patterns".

Protein sequences usually are very long, about 300 amino acids. However these sequences are constructed out of only 20 types of amino acids. We need to use constraints to make the problem tractable. It was chosen to use the time window constraint, because the discovered patterns will be concentrated in one area. The *time window constraint* means that the distance between the first and last item of the pattern in the sequence is bounded by some constant. This is easily implemented in the algorithm used. It was also considered to use the *gap constraint* [1], that allows some gaps in the matches. However this constraint would have required more memory, e.g., if we count fit_1 of (A,C,G) and we want to know whether the sequence (A,C,C,C,G) contains it. Furthermore assume the maximal gap is 1, thus in the sequence one letter is allowed between two letters of the pattern. If the algorithm only looks at the first C then the gap constraint will be broken because the gap between the C and the G is 2. An algorithm has to check two C's to match (A,C,G). PrefixSpan will have to add both projections to the projected database for at least two C's. One other reason for not using the gap constraint is that it would allow patterns to be spread all over the sequence as long as it doesn't break the gap constraint.

5 Experimental Results

The experiments are aimed at showing the effectiveness of the pruning rules we described. The protein sequences used during our experiments where extracted from the GPCRDB website [8]. The effectiveness was also tested on a synthetic data set: the two classes consist of 1000 sequences of length 130, having 20 item types. First each item is chosen with a uniform probability and then we insert one of ten patterns at each starting position within the time window (position 20 to 60) of class one with 80% probability.

The results are shown in Figure 2 and Figure 3. All experiments were done on a Pentium 4 2.8 GHz with 512MB RAM. On the horizontal axis in the graphs we have the number of used sequences in the data set. As both synthetic and protein data set have two classes, we take one half of these sequences from the first class and the rest from the second class. In the case of the GPCRDB Olfactory data set the first class contain the Olfactory sequences and the second class the No-olfactory sequences. Furthermore the GPCRDB Amine data set contains Amine and Peptide sequences. With this data we want to show that some groups of sequences are harder to distinguish. On all the vertical axis we have the pruning effectiveness indicated by a real number between 0 and 1. This effectiveness is calculated by dividing the search time by the worst search time in the experimental results. During the experiments we searched for the 100 maximal discriminating patterns in the GPCRDB and 10 in the synthetic data set, each with a time window of 8 and a *minsupp* of 0. Note that time window and probable time window are different concepts. The experiments on the synthetic data are done to indicate that the probable time window can improve pruning efficiency. Other experiments will show the effectiveness of the method in the case of GPCRDB data.

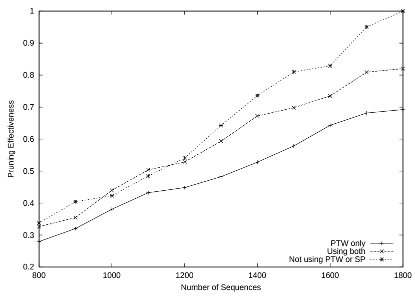

Fig. 2. Effectiveness on the GPCRDB Olfactory/No-olfactory data set

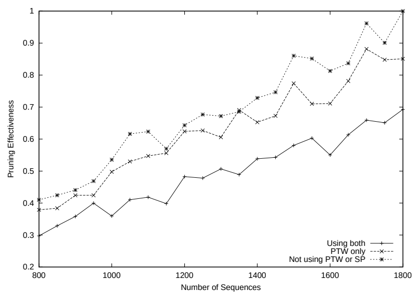

Fig. 3. Effectiveness on the synthetic data set

Figure 2 shows the effectiveness of using probable time windows (PTW) of Algorithm 2 and pruning when using "small patterns" (SP) on the GPCRDB Olfactory data. The algorithm not using PTW or SP is shown in Algorithm 1. Note that SP lowers pruning effectiveness with regards to the GPCRDB Olfactory data, because less variations of the same pattern fill up the set of patterns. Some of the patterns discovered with this data set were used for classification: these two protein families (Olfactory and No-olfactory) could be correctly distinguished in more than 90% of the cases, depending on the chosen time window size and the classification algorithm at hand.

In the synthetic data set we have most of the best patterns in the probable time window. The n-th pattern p will get a large fit_2 earlier in the search, thus more extensions can be ignored. Figure 3 shows the effectiveness as the number of sequences in the synthetic data set increases when searching for the 10 maximal discriminating patterns. The "small pattern" rules (SP) increase the effectiveness even further, because in the synthetic data set many patterns are quickly non-improving.

Table 1. Confusion matrices of Olfactory (GPCRDB) patterns without (left) and with (right) "small patterns" (SP)

	classified as no-olfactory	classified as olfactory
no-olfactory	2015	22
olfactory	16	1909

	classified as no-olfactory	classified as olfactory
no-olfactory	2024	13
olfactory	22	1903

Table 2. Confusion matrices of Amine/Peptide (GPCRDB) patterns

	classified as amine	classified as peptide
amine	489	16
peptide	3	1091

The confusion matrices of Table 1 and Table 2 were generated using the C4.5 implementation by Weka [16] with the 10 (Olfactory) and 20 (Amine) best patterns discovered in the GPCRDB data. In Table 1 we get a slightly better classification in 10-fold cross-validation when using SP: 99.12% instead of 99.04%. This is as expected because the set of 10 patterns used in Table 1 will contain less small variations of the same pattern. The results of Table 2 required 20 patterns instead of 10. The Amine/Peptide problem is more difficult than the Olfactory/No-olfactory problem and it requires more patterns. The effect of SP on classification is small, however to show that the difference in classification performance is significant a two-tailed unpaired t-test was performed. Ten-fold crossover with 1999 sequences was done 100 times with two groups of 50 Amine/Peptide patterns, with and without SP, and a time window of 4. The t-value of 6.420 with a probability of less than 0.001 of happening by chance shows that the patterns found with SP classify significantly better when using the C4.5 algorithm with these patterns as attributes.

Figure 4 shows less improvement of the pruning effectiveness. This is because the patterns in the probable time window of the Amine sequences are less discriminating compared to the patterns in the probable time window of the Olfactory sequences. We still need to evaluate many patterns if the δ_{min} stays low, even though we might find the maximal discriminating patterns quickly.

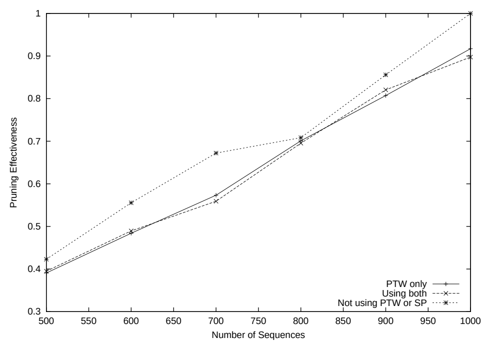

Fig. 4. Effectiveness on the GPCRDB Amine/Peptide data set

6 Conclusion

In this paper we introduced and compared two sequential pattern mining algorithms by using knowledge from the application area of protein sequence analysis. Given some assumptions, we can improve mining for the maximal discriminating patterns. The effectiveness depends on the quality of the assumptions, e.g., how probable a discriminating pattern is within a certain time window. Our method also depends on the discriminative power of the patterns. Pruning will be less effective if this is low, even though we might find the maximal discriminating patterns quickly. It is shown that using probable time windows in protein sequences can speed up the search. Protein sequences are long but contain only a few types of items; constraints are required to make the discovery of patterns in these sequences tractable.

In future research we will further investigate methods for automatically discovering the probable time window. Furthermore we plan to use maximal discriminating patterns in other application areas like workflow analysis.

References

1. Antunes, C., Oliveira, A.L.: Generalization of Pattern-Growth Methods for Sequential Pattern Mining with Gap Constraints. In Machine Learning and Data Mining in Pattern Recognition (MLDM 2003), Lecture Notes in Computer Science 2734, Springer, pp. 239–251.
2. Agrawal, R., Imielinski, T., Srikant, R.: Mining Association Rules between Sets of Items in Large Databases. In Proc. of ACM SIGMOD Conference on Management of Data (1993), pp. 207–216.
3. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In Proc. International Conference Data Engineering (ICDE 1995), pp. 3–14.

4. Bailey, J., Manoukian, T., Ramamohanarao, K.: Fast Algorithms for Mining Emerging Patterns. In Proc. 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002), Lecture Notes in Artificial Intelligence 2431, Springer, pp. 39–50.
5. Cong, G., Pan, F., Yang, J., Zaki, M.J.: CARPENTER: Finding Closed Patterns in Long Biological Datasets. In Proc. Conference on Knowledge Discovery in Data (SIGKDD 2003), pp. 637–642.
6. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by Aggregating Emerging Patterns. In Proc. International Conference on Discovery Science (DS-1999), pp. 30–42.
7. Graaf, E.H. de, Kosters, W.A.: Efficient Feature Detection for Sequence Classification in a Receptor Database, to appear in the Proceedings of the Seventeenth Belgium-Netherlands Conference on Artificial Intelligence, BNAIC2005.
8. GPCRDB: Information System for G Protein-Coupled Receptors (GPCRs), Website `http://www.gpcr.org/7tm/`.
9. Jovanoski, V., Lavrač̃, N.: Classification Rule Learning with APRIORI-C. In Proc. 10th Portuguese Conference on Artificial Intelligence (EPIA 2004), pp. 44–51.
10. Kavsẽk, B., Lavrac̃, N., Jovanoski, V.:APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery. In Proc. International Symposium on Intelligent Data Analysis (IDA 2003), Lecture Notes in Computer Science 2810, Springer, pp. 230–241.
11. Kramer, S., Raedt, L. De, Helma, C.: Molecular Feature Mining in HIV Data. In Proc. Conference on Knowledge Discovery in Data (SIGKDD 2001), pp. 136–143.
12. Lesh, N., Zaki, M.J., Ogihara, M.: Mining Features for Sequence Classification. In Proc. International Conference Knowledge Discovery and Data Mining (KDD 1999), pp. 342–346.
13. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In Proc. Conference on Knowledge Discovery in Data (SIGKDD 1998), pp. 80–86.
14. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In Proc. of the 2001 IEEE International Conference on Data Mining (ICDM 2001), pp. 369–376.
15. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. IEEE Trans. Knowl. Data Eng. 16(11) (2004), pp. 1424–1440.
16. Weka 3: Data Mining Software in Java, Website `http://www.cs.waikato.ac.nz/ml/weka/`.
17. Wess, J.: G-Protein-Coupled Receptors: Molecular Mechanisms Involved in Receptor Activation and Selectivity of G-Protein Recognition, FASEB Journal 11 (5) (1997), pp. 346–354.
18. Zimmermann, A., Raedt, L. De: CorClass: Correlated Association Rule Mining for Classification. In Proc. International Conference on Discovery Science (DS-2004), pp. 60–72.