# Approaches to the automatic discovery of patterns in biosequences.

Alvis Brāzma
Institute of Mathematics and Computer Science, University of Latvia*

Inge Jonassen, Ingvar Eidhammer
Department of Informatics, University of Bergen, Norway

David Gilbert
Department of Computer Science, City University, London, UK

September 15, 1997

**Abstract**

This paper surveys approaches to the discovery of patterns in biosequences, and places these approaches within a novel framework. The framework is based on a systematisation of the types of patterns and discovery algorithms. Patterns with the expressive power in the class of regular languages are considered, and a classification of pattern languages in this class is developed, covering those patterns which are the most frequently used in molecular bioinformatics. A formulation is given of the problem of the automatic discovery of such patterns from a set of sequences, and an analysis presented of the ways in which an assessment can be made of the significance and usefulness of the discovered patterns. It is shown that the problem is related to problems studied in the field of machine learning. The major part of this paper comprises a review of a number of existing methods developed to solve the problem and how these relate to each other, focusing on the algorithms underlying the approaches. A comparison is given of the algorithms, and examples are given of patterns that have been discovered using the different methods.

**Keywords:** automatic discovery, bioinformatics, biosequences, machine learning, patterns.

## Introduction

Recently it has become relatively cheap and easy to determine nucleotide and protein sequences, and a considerable number of sequences has been amassed, with a total length of several hundreds of millions of symbols. Moreover there are many different databases containing such sequence

*Work done whilst at the Department of Computer Science, University of Helsinki, Finland

data. At present the emphasis in genome projects is moving from the acquisition of sequence data towards the analysis of this data. The aim of the analysis is the extraction of all sorts of biological "meaning" of these sequences, for example the evolutionary history of the respective macromolecules, and their three-dimensional structure and function. One way of analysing the sequences is to group them into *families*, each family being a set of sequences believed to be biologically (i.e. evolutionarily, structurally or functionally) related, and for each family to try to find common features or *patterns*. In this paper we survey methods for the *automatic* discovery of such patterns.

Different kinds of patterns can be used for characterising sequences and the corresponding macromolecules. For proteins and protein sequences we should distinguish between *sequence patterns* and *structure patterns*. Structure patterns describe features of the three-dimensional structures of the macromolecules, and may also include information about the corresponding sequences. Sequence patterns describe pure sequence (syntactic) properties. In this paper we consider only sequence patterns, and also regard *sequences* and *strings* to be synonymous terms. For instance `C-x(2,4)-[DE]` is a sequence pattern matching any sequence containing a substring starting with `C` followed by between two and four arbitrary symbols followed by either a `D` or an `E`. This pattern is an example of a *deterministic* pattern. Patterns may also be *probabilistic*, i.e., assigning a probability to the match between a sequence and the pattern. Examples of probabilistic patterns are profiles and Hidden Markov Models. Deterministic patterns are simple and pure mathematical concepts, and are easier to interpret than probabilistic patterns. On the other hand, probabilistic patterns have more modelling power.

If a pattern is discovered to be common to a set of biologically related sequences, it is possible that the presence of this particular pattern plays a part in determining the biological function of the corresponding macromolecules. Also if we detect in a new sequence the presence of a pattern known to be characteristic to a certain family, then we can hypothesise that the new sequence belongs to that family, even if we do not know its biological properties yet. In this way patterns may be used for the classification of bio-sequences and for predicting their properties. A pattern is said to be *diagnostic* for a family if it matches all the known sequences in the family, and no other known sequences.

Many of the known protein families have been collected in the PROSITE database [Bai92]. For most of the families a diagnostic pattern is given; for some families, the pattern given is not perfectly diagnostic — it may fail to match some sequences in the family, and/or it may match some known sequences outside the family. For example, accession number PS00028 in PROSITE (release 13, November 1995) gives the zinc finger c2h2 family containing 279 proteins in SWISS-PROT (release 32, November 1995). The pattern `C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H` matches all 279 known member sequences, but also 26 other sequences in SWISS-PROT.

Powerful automatic pattern discovery algorithms may enable us to look for patterns in a large variety of potentially related biosequences. For example, such methods may be used to automatically construct patterns for the PROSITE database. These patterns are currently constructed semi-manually. Apart from the fact that this is a tedious process, this method does not guarantee that all the possible patterns are explored and that the best patterns are found. Another example use of pattern discovery algorithms, could be the analysis of DNA sequences believed to be involved in gene regulation. More generally, one could possibly discover new and unexpected

relationships and regularities in biological sequences. Thus pattern discovery methods may prove to be an important tool for knowledge discovery (data mining) in biosequence databases.

In this paper we consider the problem of the automatic discovery of deterministic patterns in biosequences. This is a machine learning problem, namely that of extracting general rules from particular instances. In this context the pattern is the general rule and sequences are the instances of the rule. Given a set of positive examples (sequences in some family), and possibly a set of negative examples (sequences not in this family), the problem is to find patterns matching the positive examples and not the negative examples (if given). Many nontrivial and interesting methods have been developed for this problem and it would appear that the field would benefit from some systematisation. For this reason, we have designed a framework consisting of the following five dimensions.

The first is based on formulating the problem of pattern discovery in the framework of machine learning, e.g., the pattern discovery problem is related to learning from only positive examples, from both positive and negative examples, and to learning from noisy data. Second, we define a pattern language that effectively includes most of the deterministic pattern languages used in biocomputing, and by restricting this general language we can obtain pattern languages studied by particular authors. The third dimension is based on the formulation and study of the concept of a fitness measure describing how well a discovered pattern fits the training data as well as rating our *a priori* belief in the patterns. Fourth, the pattern discovery algorithms can be rated according to whether they are guaranteed to find the patterns of the highest fitness or not. Finally, along the fifth dimension, we divide the pattern discovery algorithms themselves depending on the algorithmic paradigms underlying them. We distinguish between algorithms that are based on the enumeration of possible patterns and choosing the fittest ones (*pattern-driven* or PD), algorithms based on looking for common parts in sequences (*sequence-driven* or SD), and algorithms based on a combination of these paradigms.

The structure of this paper is as follows. In section 2 we describe the first four dimensions of the framework. We also give a discussion of whether or not the pattern languages used gives sufficient expressive power to describe the crucial biological features of the sequences. In section 3 we define the PD and SD approaches to pattern discovery, and also describe ways of combining these approaches. A survey is given of many of the currently known automatic pattern discovery methods. We place each of the surveyed methods within the common framework, and furthermore we identify the main algorithmic ideas of each method and show how these ideas relate to each other.

In Appendix A we present some basic information about the specific existing algorithms, and in Appendix B we give examples of the results of computational experiments, and samples of the patterns which have been discovered, for some of the existing methods. These experimental data are taken from the papers describing each of the methods. We conclude with a discussion of the possibility of establishing some benchmarks in the area of pattern discovery in biosequences.

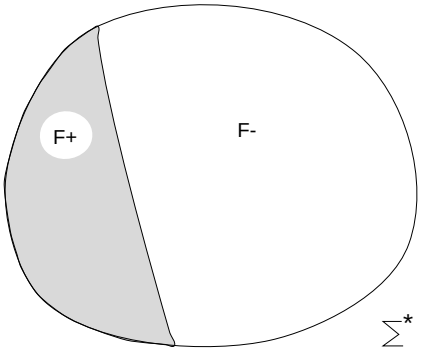


Figure 1: Schematic figure showing the relationships between the set $F_+$ of sequences in a family, $F_-$ the sequences outside the family, and $\Sigma^*$ the set of all possible sequences (strings).

## 2. Definition and discussion of the problem.

In this section we formulate the problems related to learning from biosequences. After giving semi-formal specification of the problems, we sketch a three step approach to solving these problems, and discuss each step separately.

### 2.1 The formulation of the problems

Let $F_+$ be a set of sequences corresponding to a set of proteins sharing definite functional or structural properties (e.g. containing the same type of domain). We say that $F_+$ is a *family*. The set $F_+$ is a subset of the total set $\Sigma^*$ of all possible strings over the amino-acid (or nucleotide) alphabet $\Sigma$. The sequences in $F_- = \Sigma^* - F_+$ is the set of sequences outside the family. This is sketched in figure 1. Note that while all sequences in $F_+$ correspond to biological macromolecules, there may be sequences in $F_-$ not corresponding to such molecules. For example, not all sequences over the amino-acid alphabet correspond to foldable proteins.

Let $g$ be a function $g : \Sigma^* \rightarrow \{\text{FALSE}, \text{TRUE}\}$ assigning boolean values to strings. Such a function will be called a *string function*. Let $g$ be a string function defined as follows:

$$g(s) = \begin{cases} \text{TRUE} & \text{if } s \in F_+ \\ \text{FALSE} & \text{if } s \in F_- \end{cases} \tag{1}$$

We will call $g$ a *characteristic function* for the family $F_+$. including

The basic problem studied in this paper is that of automatically finding string functions $f$ approximating the characteristic function $g$ for the family $F_+$. An algorithm for solving this problem takes as input a *training set* consisting of *positive examples*, which are sequences from $F_+$, and optionally *negative examples* which are sequences from $F_-$. This is a machine learning problem, namely that of learning a general rule from a set of examples. When both positive and negative examples are given, we choose to call it the *classification problem*, and when only positive examples are given, we call it the *conservation problem*. These problems are discussed below, and several more formal problem definitions given, each with different assumptions about

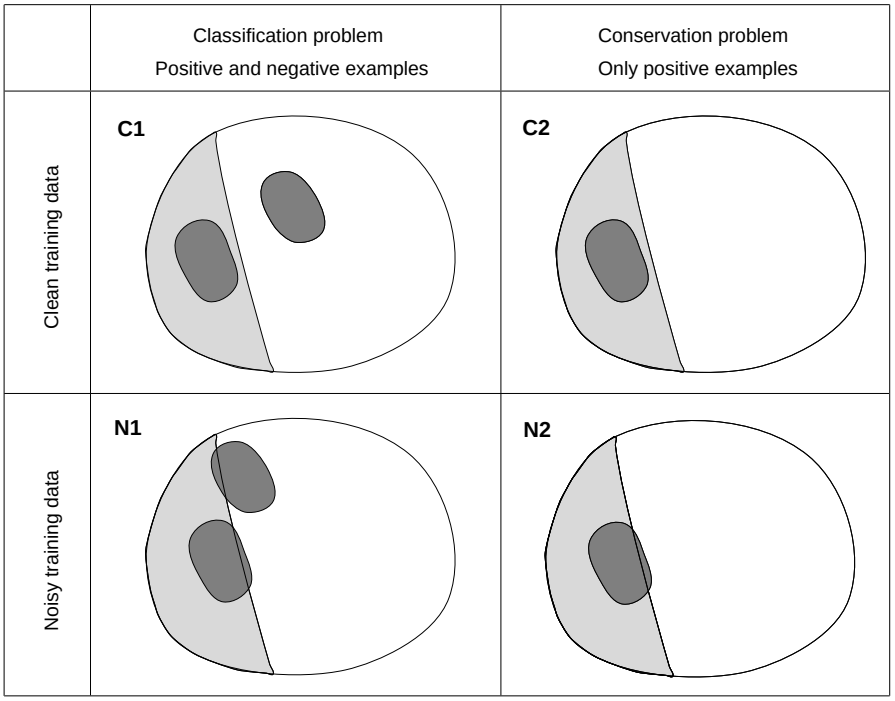


Figure 2: Schematic figure showing the training set for the different problems formulated in section 2.1. The training sets are in dark colour. In the left column, both positive and negative examples are given, while in the right column only positive examples are given. In the top row, clean training set can be assumed, while in the bottom row clean training set cannot be assumed.

the training set. We discriminate between the case when the training set is assumed to be correct (clean data), and the case when there may be errors in the training set (noisy data). These different situations and the corresponding problem definitions are illustrated in figure 2.

and

case training

### 2.1.1 Classification problem

Suppose we are given a set of sequences $S_+$ believed to be members of a family $F_+$, and a set $S_-$ of sequences believed not to be members, i.e. $S_+ \subset F_+$ and $S_- \subset F_-$. The goal is to find a string function approximating the characteristic function for this family. Let us call such a function a *classifier function*. Assuming clean data the problem can be stated as follows.

**C1:** Suppose there exist two disjoint sets of sequences $F_+$ and $F_-$ ($F_+ \cap F_- = \emptyset$) such that $F_+ \cup F_- = \Sigma^*$. Given two sets $S_+$ and $S_-$ such that $S_+ \subset F_+$, and $S_- \subset F_-$, find *compact* classifier functions that return TRUE for sequences in $S_+$, FALSE for sequences in $S_-$, and have high likelihood of returning TRUE for sequences in $F_+$, and FALSE for sequences in $F_-$.

By compact we mean having a short description. We do not define precisely here what we mean by a 'short description' and by 'high likelihood', and we discuss different ways of defining these notions later. As stated, **C1** consists of two parts:

**C1a:** find compact "explanations" of known sequences, and family.

**C1b:** try to predict the family membership of yet unknown sequences.

In reality sequences come from biological experiments and may contain errors [KLP92], as well as possibly having been wrongly included in the set of positive or negative examples. Therefore in general we cannot assume clean input and we should allow for some noise in the training set. In the noisy case it is difficult to formulate all the aspects of the problem precisely. One possible definition is the following.

**N1:** Suppose there exist two disjoint sets of sequences $F_+$ and $F_-$ ($F_+ \cap F_- = \emptyset$) such that $F_+ \cup F_- = \Sigma^*$. Given sets $S_+ \subset \Sigma^*$ and $S_- \subset \Sigma^*$ such that intersections $S_+ \cap F_-$ and $S_- \cap F_+$ are small, find compact classifier functions that return TRUE for most sequences in $S_+$, FALSE for most sequences in $S_-$, and have high likelihood of returning TRUE for sequences in $F_+$, and FALSE for sequences in $F_-$.

does not imply that most sequences in $S_+$ ($S_-$) are in $F_+$ ($F_-$).

Unlike the case **C1**, we cannot select the classifier functions only from those which return correct TRUE/FALSE values for the entire training set. We need to find a balance between how well the classifier function fits the training set (i.e. how much is meant by "most"), and our *a priori* belief in how likely is it that different functions are able to correctly classify new sequences. The ways of assessing the latter are discussed later.

### 2.1.2 Conservation problem

Sometimes it is useful to find features common to a family of sequences, even if they are not unique to the family. In this case we do not want to construct a classifier function, but rather a function showing what is characteristic of the family. We call such a function a *conservation function*.

Let us say that a function is conserved in a set of sequences $S$ if it returns TRUE for all sequences in $S$. Also, we say that a conservation function is *interesting* if it has a low probability of returning TRUE for random sequences[1], one function being more interesting than another if it has a lower probability of returning TRUE for random sequences.

**C2:** Suppose there exists a set of sequences $F_+$. Given a set $S_+$ such that $S_+ \subset F_+$, find interesting conservation functions (i.e. having low probability of returning TRUE for

[1] We assume some distribution for random sequences, for example assuming that the symbols in the sequences are independent and identically distributed (i.i.d.), i.e. $p_a = \frac{1}{|\Sigma|}$. Alternatively the frequencies of the symbols (in the training set, or in a database of nucleotide/protein sequences) can be used to define the symbol probabilities, i.e. $p_a = f_a$.

random sequences) that return TRUE for all sequences in $S_+$ and have high likelihood of returning TRUE for the sequences in $F_+$.

Note that an instance ($S_+$) of the conservation problem is equivalent to an instance ($S_+, S_-$) of the classification problem where $S_-$ consists of random sequences outside $F_+$. A method for solving the classification problem should be used if a classifier function is wanted in cases where there are sequences outside the family $F_+$ that are very similar to sequences in that family, and negative examples similar to sequences in $F_+$ should be included in $S_-$.

$F_+$. solving sequences

most of sequences in $F_+$.

**N2:** Suppose there exists a set of sequences $F_+ \subset \Sigma^*$. Given a set $S_+ \subset \Sigma^*$ such that $S_+ \cap \overline{F_+}$ is small[2], find interesting conservation functions (having low probability of returning TRUE for random sequences) such that they return TRUE for most sequences in S+ and have high likelihood of returning TRUE for the sequences in $F+$.

In order to solve the problems **C1-N2** we will split them into three subproblems.

1. Find a good class of string functions from which the approximating function $f$ will be chosen for a particular real-world problem. We call this class the *solution space*, *hypothesis space*, or *target class*.
2. Define a ranking of the solution space, evaluating how good each function is for the given training set (i.e. how likely it is to approximate $g$). We call it a *fitness measure*.
3. Develop an algorithm returning those classifier functions from the given solution space that rate high enough according to the fitness measure.

The success in solving the prediction part of the problems will depend on how successfully we will have chosen the solution space and fitness measure, even if the algorithm is perfect in the sense of finding all the fittest patterns. In the next section we discuss possible solution spaces.

## 2.2 Solution spaces.

We have defined string functions to be predicates which return TRUE or FALSE values. However, it is possible to generalise the definitions of string functions so that they return real values indicating the likelihood (or probability) that a given sequence belongs to a classes given family. A possible classification of different functions ranging from probabilistic to deterministic ones, has been given by Douglas Brutlag in a keynote address at the third international conference on Intelligent Systems for Molecular Biology (ISMB-95) as follows:

```
Deterministic                    Consensus patterns
     |                           Alignments
```

[2] $\overline{F_+}$ is the complement of $F_+$ with respect to $\Sigma^*$, i.e. $\overline{F_+} = \Sigma^* - F_+$

```
        |                       Blocks or Weight Matrices
        |                       Templates or Profiles
        V                       Bayesian Networks
  Statistical                   HMMs
```

The distinction between Hidden Markov Models (HMMs), Bayesian Networks, Templates and profiles is not strict, and depends on the detailed definitions of the models and the profiles used. Each of these models have application fields for which it is better suited. More on the statistical functions can be found in [KBM+94, BCHM94] (on Hidden Markov Models), [GME87, BB94] (profiles), and [CWC92] (alignments). Some authors have used maximum likelihood models for finding good blocks (ungapped local alignments) [**?**, LAB+93, **?**, **?**]. Here we focus on the deterministic end, and particularly on patterns with the expressive power within the class of regular languages.

### 2.2.1 Input alphabets

The ways of defining string functions will depend on the properties of the input alphabets. There are differences between nucleotide (DNA/RNA) and protein sequences that should be taken into account. Protein sequences are sequences over a 20-letter alphabet $\Sigma_p = \{$ `D, E, K, R, H, Q, S, T, I, L, V, F, W, Y, C, M, A, G, P`$\}$. L,V,F,W,Y,C,M,A,G,P }. Nucleotide sequences (DNA/RNA) are sequences over 4-letter alphabets $\Sigma_{DNA} = \{$`a,t,g,c`$\}$, and $\Sigma_{RNA} = \{$`a,u,g,c`$\}$. The set of amino acids $\Sigma_p$ may be grouped in classes $K_1, \ldots, K_n$ in different ways according to their physio-chemical properties, e.g. in AACC hierarchical groups taken from [SS90] ($K_1 = \{$`D,E`$\}$, $K_2 = \{$`K,R,H`$\}$, $K_3 = \{$`N,Q`$\}$, $K_4 = \{$`S,T`$\}$, $K_5 = \{$`I,L,V`$\}$, $K_6 = \{$`F,W,Y`$\}$, $K_7 = \{$`C,M`$\}$, $K_8 = \{$`A,G`$\}$, $K_9 = \{$`D,E,K,R,H,Q,S,T`$\}$, $K_{10} = \{$`I,L,V,F,W,Y,C,M`$\}$, and $K_{11} = \Sigma_p$), or using Venn-diagrams [Tay86], or in some other way. Also, *substitution matrices*, e.g. PAM [Day78], and the Blosum [HH92], can be defined giving statistics for each pair of amino acids for how often they are found in equivalent positions in proteins that are within a certain evolutionary distance. The nucleotides may also be divided into two groups: pyrimidines and purines, and scoring matrices can be defined, however, this seems to play much less important role than in protein sequence analysis.

Both protein sequences and nucleotide sequences can be translated into smaller alphabets. Such a translation is called an *indexing*, and is obtained by making a partition[3] of the basic alphabet $\Sigma$, and translating symbols in the same partition into the same symbol in the reduced alphabet. For example, amino acids are either hydrophobic, neutral, or hydrophilic, and can be mapped onto a three-symbol alphabet $\Sigma_{hydro} = \{+, 0, -\}$ (e.g. [AKM+92]). Similarly nucleotide sequences can be translated into a purine-pyrimidine alphabet $\Sigma_{n_{red}} = \{R, Y\}$.

### 2.2.2. Generalised regular patterns

Pattern definitions are usually given by examples in the biocomputing literature. Here we give a more formal definition, and define a class of what we call *generalised regular patterns* that will

[3] A partition of a set $A$ is a set $B$ of disjoint subsets of $A$ such that the union of the sets in $B$ is $A$.

unify most of the deterministic pattern classes used in biocomputing. Various more restrictive pattern classes will be obtained as appropriate subclasses.

Let $\Sigma = \{a_1, \ldots, a_m\}$ be an alphabet called the *basic alphabet*. I.e. for protein sequences $\Sigma = \Sigma_p$, for nucleotide sequences $\Sigma = \Sigma_{DNA}$, or $\Sigma = \Sigma_{RNA}$; if an indexing is applied, $\Sigma$ may be, for example, $\Sigma_{hydro}$ or $\Sigma_{n_{red}}$, or any other abstract alphabet. Let $K_1, \ldots, K_n$ be some subsets of $\Sigma$, such that each subset contains more than one element ($|K_i| \geq 2$). Let $\Pi = \{b_1, \ldots, b_n\}$ be another alphabet disjoint with $\Sigma$, and let us define $L(b_i) = K_i$. For convenience let us also assume that $L(a_i) = \{a_i\}$, for $a_i \in \Sigma$. In practice $K_1, \ldots, K_n$ are classes of amino-acids (e.g. the AACC classes) or nucleic acids, and $b_1, \ldots, b_n$ are the characters denoting these classes. The character denoting a class $K_i = \{a_{i_1}, \ldots, a_{i_l}\}$ is usually denoted by $[a_{i_1} \ldots a_{i_l}]$. For instance, in our representation of the AACC hierarchy, $b_1$ denoting $K_1 = \{\texttt{D,E}\}$, would be denoted by [DE]. This does not apply, however, to the character $b_i$ standing for the whole $\Sigma$, which is usually denoted by $x$ (or sometimes by $\cdot$), effectively meaning the *wildcard* (or *don't-care*) character. Here we will use $x$ for the wildcard.

Let $x(p, q)$, where $p$ and $q$ are non-negative integers and $p \leq q$, be a wildcard of the length from $p$ to $q$, and let $L(x(p, q))$ be defined as a set of all words over $\Sigma$ of length between $p$ and $q$, i.e. $L(x(p, q)) = \{\alpha \in \Sigma^* | p \leq |\alpha| \leq q\}$. Let $X$ be the set of all objects of the form $x(p, q)$. Finally, let $*$ be a character such that $* \notin \Sigma \cup \Pi$, and let $L(*) = \Sigma^*$, i.e. $*$ is the wildcard of an arbitrary length - $x(0, \infty)$.

A *generalised regular pattern* $\pi$ is a string over an alphabet $(\Sigma \cup \Pi \cup X \cup \{*\})$. We define the language $L(\pi)$ of the pattern $\pi$, where $\pi = c_1 \ldots c_r$, and $c_i \in \Sigma \cup \Pi \cup X \cup \{*\}$, as

$$L(\pi) = L(c_1) \ldots L(c_r),$$

the concatenation being defined as $L(c_1) \ldots L(c_k) = \{\gamma_1 \ldots \gamma_r | \gamma_1 \in L(c_1), \ldots, \gamma_r \in L(c_r)\}$. A string $\alpha$ *matches* a pattern $\pi$ if $\alpha \in L(\pi)$. The class of languages that can be expressed using generalised regular patterns is a subset of regular languages. ¿From now on, by 'pattern' we will understand a generalised regular pattern.

We introduce the following pattern classification based on limiting the pattern alphabet to different combinations of sets $\Sigma$, $\Pi$, $X$, and $\{*\}$, and limiting whether the symbol $*$ can appear in an arbitrary place in a pattern, or only at the beginning and at the end. For the patterns where $*$ can appear only at the beginning and at the end, i.e. for patterns of the type $\pi = *\pi'*$, $\pi = *\pi'$, $\pi = \pi'*$ and $\pi = \pi'*$ we will distinguish between the following cases restricting the alphabet of $\pi'$:

**A:** $\pi' \in \Sigma^*$,

**B:** $\pi' \in (\Sigma \cup \{x\})^*$,

**C:** $\pi' \in (\Sigma \cup \Pi)^*$,

**D:** $\pi' \in (\Sigma \cup X)^*$,

**E:** $\pi' \in (\Sigma \cup \Pi \cup X)^*$,

In the second case when $*$ can appear in an arbitrary place we define classes:

**F:** $\pi \in (\Sigma \cup \{*\})^*$,

**G:** $\pi \in (\Sigma \cup \Pi \cup \{*\})^*$,

**H:** $\pi \in (\Sigma \cup \Pi \cup X \cup \{*\})^*$.

Note that these classes can be partially ordered in a lattice, **A** being the bottom element and **H** the top element. The class **A** corresponds to *substring patterns*, and the class **F** to *regular patterns* of Shinohara [Shi83]. The class **E** corresponds to the class of patterns most frequently used in PROSITE database. In PROSITE notation the leading and closing $*$ symbols are not used, $*\pi*$ being written simply as $\pi$. The notation permits attachment of the pattern to the beginning or the end of a sequence by using leading $<$ or closing $>$ symbol, thus $\pi$ becomes $<\pi>$. the individual symbols from the pattern alphabet are separated

Any of the classes containing the alphabet $\Pi$ can be further refined by choosing a particular alphabet $\Pi$ (i.e. particular subsets $K_1, \ldots K_n$). We can also restrict the solution space by considering patterns of a specific (or up to a specific) length.

### 2.2.3 Defining the string functions via patterns

The simplest way to define a *classification (conservation)* function using a pattern $\pi$ is

$$f(\sigma) = \begin{cases} \text{TRUE} & \text{if } \sigma \in L(\pi) \\ \text{FALSE} & \text{otherwise.} \end{cases} \qquad \textbf{(a)}$$

An extension of this approach is to allow for approximate matching between the pattern and the string. A measure of the distance between two strings can be used (e.g. edit distance [Ste94]). If $dist(\sigma_1, \sigma_2)$ is the distance between two strings $\sigma_1$, and $\sigma_2$, then the distance between a string and a pattern can be defined as $Dist(\pi, \sigma) = \min_{\sigma' \in L(\pi)} dist(\sigma', \sigma)$. Hence the definition of classifier/conservation function $f$ can be generalised to include approximate matching:

$$f(\sigma) = \begin{cases} \text{TRUE} & \text{if } Dist(\pi, \sigma) \leq const \\ \text{FALSE} & \text{otherwise,} \end{cases} \qquad \textbf{(b)}$$

for some given constant *const*. More advanced measures than edit distance (or analogous similarity measures) have been used, for example in database similarity search programs, e.g., [**?**, **?**]. These involve use of amino acid substitution matrices [Day78, HH92].

i.e. class **A**)

modelling power represents

Patterns also can be used for defining classification/conservation functions in more complicated ways than just exact (**a**), or approximate (**b**) membership of a language. We are aware of two more ways reported in the literature:

**c** membership of a union of pattern languages (i.e. $f(\sigma) = \text{TRUE}$, if $\sigma \in L(\pi_1) \cup \ldots \cup L(\pi_n)$ where $\pi_1, \ldots, \pi_n$ are patterns of some class) [AFSA94, TSLM$^+$95, BJUV96].

**d** using decision trees over patterns (for the definition of a decision tree see [AMS+93]).

Method **c** is useful when a family contains several subfamilies and whilst there are strong conserved patterns in each subfamily, the family as a whole has only very weak conserved patterns. Note that the learning of unions of patterns languages from positive examples is closely related to *unsupervised learning* – the task of simultaneously finding a set of patterns, and subfamilies (i.e. subsets) of the training set, so that each subfamily shares a distinct pattern.

functions, definition

Since most of the information about the function for the considered classes of string functions is given by the underlying patterns, we will from now on often refer to patterns instead of functions.

### 2.2.4 Do the solution spaces give sufficient expressive power?

The PROSITE database contains many examples of protein families which have diagnostic consensus patterns of the type defined above. For example, the zinc finger proteins all have cysteine or histidine amino acids in certain positions. In these cases the class of generalised regular patterns may be sufficient. In many other families the features conserved between the sequences are more subtle. Probably there does not exist any consensus pattern diagnostic for the helix-turn-helix family. Patterns of correlations exist (e.g. pseudo-knots in RNA secondary structures) that give crossing dependencies taking us even beyond what can be compactly described using context-free grammars.

TRUE/FALSE and

The choice of a target class $\mathcal{C}$ for our approximation functions $f$ (i.e. the first part of our approach) is a trade-off between what classes of functions are expressive enough to allow description of the crucial biological features of sequences, and for what classes of functions we can develop an efficient algorithm for finding string functions from examples, i.e. what classes can be learned efficiently. For a particular problem, we want to choose a class $\mathcal{C}$ as small as possible for efficiency reasons, but at the same time $\mathcal{C}$ should be expressive enough to contain a function $f$ approximating $g$.

After we have chosen the solution space $\mathcal{C}$, the problem of finding the function $f$ from positive and negative examples $S_+$ and $S_-$ has some similarity to PAC-learning (probably approximately correct learning) [Val84, SA95]. The difference between our case and PAC learning is that PAC learning requires approximation within some given precision, while in our case it may not be possible. Also, a potentially infinite number of examples are assumed to be available in PAC learning, while in our case we have only finite number of examples, and frequently this number is quite small. Moreover, it is possible that the shortest description of any good approximation function $f$ is simply a comprehensive collection of all the positive examples. In this case the task of learning from biosequences is not so much that of approximating a hypothetical function $g$, but rather just discovering interesting properties of $g$.

hypothesises.

### 2.3 Ranking discovered patterns and functions

classification function is In most of this subsection we assume that a conservation or classification function is defined directly from a single pattern using method **a** (exact match) or **b** (approximate match). In this way we are able to talk about ranking patterns instead of string functions, thus making it easier to relate the discussion to papers describing pattern discovery methods.

A fitness measure can be defined as a function

$$F(\pi, S) \to \mathcal{R}$$

that takes two arguments and returns a real. The first argument is a pattern from a particular pattern class. The second argument is the training set. The value of the function should show how good the pattern is in respect to the training set. Sometimes the values can be normalised to $[0, 1]$.

If we have chosen a very restricted solution space (e.g. based on patterns of class **A** of a fixed length), the fitness measure can be very simple, particularly in the case of clean data. It can have just two values 0 and 1, assigning 1 to patterns matching all the positive example sequences, and to none of the negative example sequences for the classification problem. In the noisy case we can use this simple two value fitness measure by assuming a certain maximum level of noise (say, 30%), and assigning 1 to the patterns correctly classifying at least the remaining portion (i.e. 70%) of the training set.

We rely heavily on the choice of the solution space in the case of this simple fitness measure. If we choose too general a solution space, for instance including the regular expression $\cup_{s \in S_+} s$, it can lead to overfitting the training set and the string functions obtained may be not useful at all for prediction. The heavy burden of making the right choice of the solution space can be relieved by using a more sophisticated fitness measure that depends not only on how well the pattern fits the training data, but also different includes as a factor an *a priori* rating of the patterns judging from their syntactic form. We can express $F(\pi, S) = F_1(\pi, S) \cdot F_2(\pi)$, where $F_1(\pi, S)$ shows how well the pattern $\pi$ fits the training data and $F_2(\pi)$ gives an *a priori* rating of the pattern. The dependence on the training data can still be a 0/1-value function: $F_1(\pi, S) = 1$, if the function classifies (conserves) the given percentage of sequences in the training set correctly, $F_1(\pi, S) = 0$ otherwise.

The function $F_2$ can be based on Occam's razor principle (see [Hut94]). This principle says that when several theories explain (past) observations equally well, the simplest theory is better. It is possible to prove formally that under certain assumptions the simplest patterns are likely "to predict the future better". In the case of the classification problem the evaluation can be based on Occam's Razor directly: simpler (e.g. shorter) patterns correctly classifying the same number of examples should be *a priori* rated higher.

In the case of the conservation problem we can use the *information content* of the pattern [JCH95] for *a priori* pattern rating. Information content means the amount of information provided by the knowledge that the pattern matches a sequence. Amongst those patterns which match the same number $n$ of sequences, we give a higher rating to those patterns with a greater information content. Since the information content of the pattern generally grows with the

length (complexity) of the pattern, this may seem to be the opposite of Occam's Razor; we discuss the close relationship between the information content and the Occam's Razor principle later on.

A weakness of using a 0/1-valued function is that the approach then depends heavily on the choice of the noise threshold. For the classifier problem a more subtle measure $F_1$ can be based on how many of the respectively positive and negative examples match the pattern $P$. and For each of the string functions $f_1, \ldots, f_n$ discovered by an algorithm, we count the number of *false positives*, i.e. $s \in S_-$ and $f(s)$ =TRUE, and *false negatives*, i.e. $s \in S_+$, and $f(s)$ = FALSE. We also define the number of true positives ($TP$) as the number of sequences $s$ in $S_+$ for which $f(s)$ =TRUE, and the number of true negatives ($TN$) as the number of sequences in $S_-$ for which $f(s)$ =FALSE. The *sensitivity* of a function $f$ [LWS+93] can be defined as

$$Sn(f) = \frac{TP}{TP + fn} \tag{2}$$

where $fn$ is the number of false negatives. Similarly, the *specificity* of the function can be defined as

$$Sp(f) = \frac{TN}{TN + fp} \tag{3}$$

where $fp$ is the number of false positives. The proposed functions may be ranked according to sensitivity, specificity, or some combination[4].

For the conservation problem a way of scoring the patterns can be based on the *statistical significance* of the patterns (e.g. [WAG84, Sta89a, **?**, SD95]), defined as follows. atistical,1995:Sewell:pattern). Suppose $p_1, \ldots, p_n$ are patterns such that each $p_i$ matches a subset $S_i$ of $S_+$. Then, for a pattern $p_i$, the *pattern probability* is the probability that $p_i$ matches at least $|S_i|$ out of $|S_+|$ random sequences (of the same length and composition as the sequences in $S+$) purely by chance. In this analysis the sequences and the sequence positions are assumed to be independent. The pattern significance can be defined as the reverse of pattern probability, thus patterns having lower probability should be ranked higher. Note that statistical significance of the pattern will increase either if the information content of the pattern increases or if the pattern matches more sequences. In this way both aspects of the fitness measure are taken into account.

A closely related way is to measure the information content of the *block*[5] defined by the substrings in the training set matching a pattern [**?**, LAB+93, Sta89b]. This can be viewed as a generalised size of this block. A problem in this case is that a relatively small number of very similar sequences can result in a block of very large size. Therefore the assumption of some minimal number of sequences needed to match the pattern is still required. A way of avoiding this is considered in [JHH96].

[4]One possible combined measure is the *correlation coefficient* between two sets; (1) the set of sequences in $T_+$, and (2) the set of sequences in the set $T_+ \cup T_-$ for which the function $f$ returns TRUE. The correlation coefficient is $C = (TP \cdot TN - fp \cdot fn)/\sqrt{(TP + fp) \cdot (fp + TN) \cdot (TN + fn) \cdot (fn + TP)}$, which is 1.0 when there are no false positives or negatives, 0.0 when $f$ is random with respect to $S_+$, and $S_-$, and $-1.0$ when there are only false positives and false negatives.

[5]A block is defined as a local ungapped alignment, i.e. it is a set of segments of identical length "put on top of each other" giving an alignment [**?**].

define the model) [WAG84, NG94]. finding a pattern used. itself, same different used.

Another theoretically substantiated balanced measure for the conservation problem can be based on the *minimum description length* (MDL) principle [Ris78, LV93, BUV96]. Considering string functions defined from a pattern (**a** in our classification) or a set of patterns (**c**), the MDL principle says that the best pattern (or set of patterns) is the one that minimises the sum of

- the length (in bits) of the description of the pattern(s); and
- the length (in bits) of the sequences when encoded with the help of the pattern(s).

of Note that this corresponds to Occam's Razor principle, and that this gives a way of applying the principle to the conservation problem. A fitness measure based on the MDL principle is developed for conservation functions which are defined by unions of patterns (i.e. case **c**) [BJUV96]. A coding scheme is introduced using PROSITE type patterns for compression of sequences. The fitness value of a conservation function is defined to be proportional to the compression of the training set that can be achieved using the set of patterns defining the function. It is shown that this fitness measure can be expressed as the sum of the information contents of each pattern [JCH95] times the number of sequences matching the respective pattern, minus a correction independent of the number of matched strings.

## 2.4 Search algorithms and guarantees

The specifications of the problems **C1** - **N2** were rather informal, leaving a number of notions undefined, including "high likelihood" and "most sequences". Any learning algorithm for solving these problems is effectively designed for a specific target class $\mathcal{C}$ and uses a specific fitness measure $F$. The input of such an algorithm is the training set $S$, and the algorithm is required, given the training set $S$, to produce a set of patterns $P$ from the class $\mathcal{C}$ such that the fitness value $F(\pi, S)$, for $\pi \in P$ is "relatively" high. If it can be proved that the algorithm will produce the specified portion of the patterns with the highest fitness value $F(\pi, S)$ among all the patterns in $\mathcal{C}$ (i.e. either the fittest pattern, or a given number or percentage of the fittest patterns, or all the patterns with a fitness higher than a given constant), then it is said to be *guaranteed* to find the best pattern (or best patterns).

problems is to

The success in the prediction of the properties of unknown sequences relies on choosing an appropriate pattern class $\mathcal{C}$ and a good fitness measure $F$. Evaluating this is outside the scope of a mathematical definition, and should rely either on experimental evaluation or on the judgement of an expert.

A set of sequences which is not included in the training set may be used as a test set in order to evaluate how good the discovered patterns are at predicting family membership of unknown sequences. A random subset of the family (and a subset of the set of sequences outside the family) can be used as a training set. The resulting patterns can be tested against the rest of the sequences, evaluating for example the sensitivity, or specificity, of the patterns. This is

a standard evaluation method which is used in experimental machine learning. However, the number of known sequences is too small for the performance of this kind of experiment for most families known at present.

It is an advantage if the algorithm produces several alternative classifier or conservation functions which can be presented to the user. In order to test a new method for pattern discovery, it can also be applied to well known families to show that it is able to recover already known conserved patterns. The ultimate test of a pattern is to check experimentally whether it corresponds to some region conserved in the family for structural and/or functional reasons.

Various algorithms for finding classification or conservation functions having high fitness according to the given measure from the given class are discussed in the next section.

# 3. Algorithms

In this section we first discuss algorithms for finding classification and conservation functions defined directly from generalised regular patterns using methods **a** and **b** (the exact and approximate match of a pattern) defined in Section 2.2.3. At the end of this section we briefly describe methods for finding functions defined using=20 unions and decision trees of patterns (methods **c** and **d**).=20

At the highest level we can divide the pattern discovery=20 algorithms in two groups.=20 The first, which we call pattern-driven (PD) approaches, is based on=20 enumerating candidate patterns in a given solution space and=20 picking out the ones with high fitness.=20 The second, which we call sequence driven (SD)=20 approaches, comprises algorithms that try to find=20 patterns by comparing given strings and looking for local similarities = between=20 them.=20 For instance an SD algorithm may be based on constructing a local=20 multiple alignment of given sequences and then extracting the patterns=20 from the alignment by combining the segments common to most of the = sequences.=20

The advantage of the PD approaches is that in this way it is possible to = =20 guarantee finding the best patterns up to some limited size, almost regardless of the total length of the examples. The reason for this is that it is usually possible to organise the algorithm so that it=20 is linear-time in this length. On the other hand=20 the size of the pattern-space is exponential in the length of the = patterns.=20 PD algorithms guaranteed to find the pattern with the highest fitness = value,=20 have worst case time complexity exponential in the length of the = patterns.=20 Thus, guaranteed PD algorithms can only find patterns of limited = complexity. It is also possible to combine PD and SD approaches in a single = algorithm.=20

It may be possible to discover patterns of an almost arbitrary size by SD algorithms. A weakness of the SD approach is that in general it is impossible to = guarantee optimality of the results without sacrificing efficiency.=20 The reason for this is that the algorithmic problems on which=20 precise comparison of multiple sequences can be based (e.g. the problem=20 of constructing the optimal multiple alignment or finding the longest=20 common subsequence) are NP-hard=20 [WJ94, GJ79], and therefore SD approaches have to = be based=20 on=20 heuristics.=20 In general SD algorithms tend to work well if the sequences are = sufficiently=20

similar.=20

In the following subsections we describe in more detail=20 the basic ideas of PD, SD and combined approaches to pattern discovery.=20 We present the basic information about=20 each algorithm separately in Appendix A, and present some sample patterns discovered by some of the algorithms in Appendix B. =20

### 3.1. Pattern driven approaches

The general framework of pattern driven algorithms can be formulated=20 as follows:

- define the solution space $\mathcal{C}$ (i.e., a set of patterns) and the fitness measure,
- enumerate the patterns in the solution space,=20
- calculate the fitness of each pattern with respect to the=20 given examples,=20
- report the fittest patterns.=20

The most straightforward implementation of the PD approach is explicit=20 enumeration of all the patterns from the pattern space one by one. =20 For instance, if the patterns are subwords=20 (i.e. of the class **A** defined in section 2.2.2)=20 of length 3 in the alphabet=20 {`a, c, g, t`}, and if we use=20 $\{0,1\}$ as a=20 value of fitness measure, assigning 1 to the patterns which correctly=20 classify the given portions of the training sequences and 0 to the others, then the algorithm can enumerate=20 all the words `aaa, aac, aag,`=20 ..., `ttt`, calculate their fitness by counting in how many examples=20 each is present, and output the words that have the fitness value 1. =20 that=20 what is meant is the

In the simplest case=20 of substring patterns (class **A**) this approach was first applied =20 in the early 1980's=20 [QWK82, WAG84]=20 and later in [Sta89b]. The search space is limited to fixed length=20 patterns and the algorithms count the number of=20 sequences in the training set that approximately match the pattern[6]. The best pattern is selected on the basis of its fitness to the the "neighbourhoods".. For *a priori* ranking of the patterns, Waterman et al. [WAG84] estimate the statistical significance of = the discovered patterns, while Staden [Sta89b] calculates a measure of the information content of the block consisting of the segments that approximately match. Recently, Wolferstetter et al. [WFHW96] have used a similar method coupled with a user-friendly Web-interface for discovering conserved motifs in genome regulatory regions, and = employ a fitness measure based on information content.

suffix-tree[7] and=20

[6]For this, a notion of the set of neighbours of a string, i.e. strings that are similar to the given string, is introduced, and=20 a count is made of in how many examples each substring or any of its neighbours is present.=20 fitness.

[7]GSTs are a generalisation of the=20 length of time currently finding

The straight-forward enumeration approach can be easily extended to more complicated patterns. Smith et al. [SAC90] have used this approach for discovering patterns containing characters from the basic alphabet =20 and also wildcards.=20 The algorithm enumerates all possible patterns consisting of three = conserved=20 positions

Note that SD algorithms that are based on combining pairs of patterns=20 in each step (as in the example above) usually cannot be guaranteed=20 to find the fittest patterns common to

with constant spacings within a pre-set range, i.e., patterns = in=20 class **B** from section 2.2.2. The patterns considered can be written in the form $a_1$-x($d_1$)-$a_2$-x($d_2$)-$a_3$, where $a_1$, $a_2$, and $a_3$ are characters from the basic alphabet, and $d_1$ and $d_2$ are the number of wildcard=20 characters $x$ in between them. 20. In this case only the numbers of sequences that exactly contain the pattern are counted. The user provides the minimum number of sequences that = should=20 contain the pattern in order for it to be considered. The patterns are evaluated using a heuristic fitness measure, and the patterns with the highest fitness are reported in the end. [SAC90] also uses elements of the SD approach to extend the patterns found by enumeration (see section 3.3).

An interesting extension of this method has been reported in=20 [SNO95], which in addition permits the discovery of patterns containing flexible length wildcards (i.e. = patterns=20 of the class **D**). =20 =20 An obvious problem in this straight-forward enumeration is that of = efficiency. The size of the search space for patterns of length $l$ grows as=20 $O(|\Sigma|^l)$. However, the number of patterns can be reduced if we impose some restrictions on the pattern class. For instance, in the method of Smith et al. [SAC90] there are $20 \times 10 \times 20 \times 10 \times 20 = 3D800000$ candidate patterns of the type $a_1$-x($d_1$)-$a_2$-x($d_2$)-$a_3$ to be checked if the distance range = is 10 (i.e., $0 \leq d_1 < 10$ and $0 \leq d_2 < 10$) and $a_i \in \Sigma_p$. However this number becomes impractical for more general pattern = classes. Therefore some method for pruning the solution space, either by a provably accurate method, or by using heuristics should be found if we want to increase the size and complexity of the patterns.

### 3.1.1. Methods for limiting the search space prior to = the=20 search

It is possible to limit the space of the patterns to be explored by = using=20 information extracted from the training set. Effectively this means using an SD algorithm=20 as a pre-processing step for a PD method.

One simple heuristic for limiting the search space is based on an assumption that the patterns that are present approximately (within some distance) in many sequences are likely to be present in an exact = form in at least some. This is not strictly true because the most fit pattern may be a kind of average, e.g. Steiner's sequence[8], itself not present in a single sequence. However this is not very likely if a sufficient number of sequences are given. Therefore a heuristic can be based on the enumeration of only those substrings that are present in at least in one sequence. This reduces the search space drastically, since there are only $O(N^2)$ substrings for a set of strings with total length $N$. If the length of the substrings is bounded by $l$, then there are only $O(lN)$ patterns to be considered, instead of $O(|\Sigma|^l)$.=20 This approach has been used by Saqi and Sternberg [SS94],=20 where a=20 statistical significance measure has also been used for sorting out the interesting patterns[9].=20 =20 approach.=20

This heuristic can be taken even further by selecting a random subset of the training set, if the number of sequences in both the total training set and the subset = is large enough. In this case it is statistically likely that any substring = which occurs approximately in a sufficient number of sequences=20 in the training set will occur in an exact form more than some number of times $n$ in the random subset[10]. Therefore we can enumerate only those substrings which are present in the subset more than $n$ times. Moreover, the strings in the subset can be represented as a=20 *generalised suffix tree* [Hui92], and then the potential candidates for the pattern can be selected in linear time=20 [WMS$^+$94]. Thus the algorithm becomes linear time in the length of the sequences and the patterns. =20 approach.

A different way of limiting the search space is proposed by Jonassen [Jon96]. Here a *pattern graph* is defined. A path in this graph = corresponds to a set of patterns, and a depth-first search strategy is used to = search for the paths corresponding to patterns matching at least $N_{min}$ of = the=20 positive examples with the highest fitness. It is possible to derive a pattern graph from an existing multiple=20 sequence alignment, for instance an alignment of a subset of the = sequences=20 in $S_+$, so that only patterns consistent with the alignment are = considered.=20 This gives a smaller search space, and can be considered as an = SD-element.

### 3.1.2. Methods for pruning during the search

A rigorous approach for pruning the search space can be based on representing it as a tree and pruning the subtrees rooted with patterns having a fitness under some threshold.=20 For instance, if we are looking for simple patterns=20 (i.e., of the class **A**) in the alphabet `{a,c,g,t}`, then a part of the pattern space can be=20 represented as the tree in figure **??**.

This tree can be traversed in either a breadth-first or a depth-first = manner; both of these ways have been explored by Sagot et al. [SVS95b]. The breadth-first approach can be more time-economic in practice because

$S_+$ in the end, even if each=20 iteration step is guaranteed to find the fittest pattern common to the =20 pair of sets that are combined in the step.=20 (This is similar to building multiple alignment by pairwise=20 alignments, which also cannot guarantee finding the optimal =20

= =20 the pruning of the search tree can be more efficient than in depth-first = search.=20 For instance, for a substring `act` to be present in a sufficient=20 number of sequences, both substrings `ac` and `ct` should have=20 been found earlier if the search is done breadth-first. If depth-first search was used, we would only require that the substring `ac` should have been found earlier. Unfortunately breadth-first search can realistically only be applied to very short patterns, because the number of the patterns to be remembered=20 grows very fast using this method. Therefore for practical purposes=20 depth-first search is used. A very efficient implementation of this idea =A0 for substring patterns is of the the Kar-Miller-Rosenberg (KMR) [KMR72] algorithm [11].

This idea is extended by Sagot et al. [SVS95b] to find more complicated patterns of class **C** containing symbols denoting groups of the basic alphabet. Unfortunately the efficiency of the algorithm=20 decreases when amino acid symbols are present in many groups. Also, if some groups are very large (containing many basic symbols) this will slow the search down because extending patterns with such = symbols=20 will often produce conserved patterns, and hence large parts of the = search=20 tree have to be explored. For this reason [SVS95b] does = not allow wildcard characters.

On the other hand it can be seen that dealing with wildcard characters should be quite easy.=20 For instance, if we have found that the fitness of the=20 pattern `tgc` is high enough, then by using wildcards the patterns can be potentially extended not only to patterns `tgca`, ..., `tgct`, but also to patterns `tgc`$x$`a`, ..., `tgc`$x$`t`, `tgc`$xx$`a`, = 20 ..., `tgc`$xx$`t`, `tgc`$xxx$`a`, ..., ... and each of them should be checked for fitness. Neuwald and Green presents a method using this approach. They apply a a pruning mechanism based on a measure of statistical significance=20 of the patterns, avoiding to explore extensions of patterns with low significance thus speeding up the depth-first search significantly [NG94]. Also, they introduce a=20 new, so called *block* data structure and use this to very efficiently find the set of substrings matching each pattern, thus speeding up the depth-first search. [NG94] allows group=20 characters in patterns (i.e., the alphabet $\Pi$) for groups consisting of pairs of amino-acids, hence giving patterns of class = **C**.

In a later paper, Sagot and Viari [**?**] have presented an alternative approach which uses a depth-first search to discover patterns containing ambiguous symbols as well as wildcards. In principal, one need not specify beforehand which groups of letters are to be used in ambiguous pattern positions. In practice, this works for nucleotide sequences, however, for protein sequences (where there are $2^{20} - 1$ possible non-empty=20 groups) one needs to define a priori the ambiguous sybmols to be used. For each possible ambiguous symbol one can set an upper limit=20 on the number of occurences of this symbol in an individual pattern.=20 This is used together with a minimum percentage of sequences to=20 contain a pattern, to prune the search. Also, if two patterns match the same sequence segments, and one is a generalisation of the other, only extensions of the least general pattern is explored further.

Jonassen et al. [JCH95] describes an algorithm where the use of a depth-first search strategy combined with the block data structure is pushed even further. This algorithm is able to discover patterns having both ambiguous positions (groups of amino acids) and flexible spacings (gaps), giving patterns in the class **E**. The user defines restrictions on the kind of patterns that can be discovered, effectively defining a subclass of **E**. The algorithm sets out to find all patterns in this subclass matching at least some minimum number $N_{min}$ of the positive sequences. The search tree is pruned so that extensions of patterns matching less than $N_{min}$ sequences, are not analysed. The algorithm works in two phases, and normally during the first phase only patterns consisting of single letter positions and wildcards are considered. The best patterns found during the first phase are passed on to a second phase where they are subjected to an either exhastive or heuristic search where ambiguous pattern symbols might be added.=20 A fitness measure for patterns is defined, and the algorithm is guaranteed to find the highest scoring patterns within the subclass of **E** that match at least $N_{min}$ of the positive = examples.

In addition to using an SD-element to limit the search space (see the = previous subsection), Jonassen [Jon96] also introduces=20 branch-and-bound and heuristics to make the pruning of the search tree=20 more efficient. This speeds the search significantly, especially for = sets=20 of quite similar sequences. In both [JCH95, Jon96] an SD element is used to specialise patterns discovered in the depth-first search.

Experiments clearly show that pruning the search space in combination = with=20 efficient data structures substantially increases efficiency of the algorithms. Nevertheless the algorithms are still worst-case exponential in the length of the patterns, and no nontrivial speed-up over the straight-forward algorithms has been proved theoretically.

Note that although all the algorithms reported here are for=20 the conservation problem, the same algorithms

alignment in the end.)=20 =20 inevitable.=20 =20 The earliest SD algorithms for finding a regular pattern (class **F**) common to a set of=20 strings that we are aware of were developed by the computational = learning=20 community, and do not have a direct relations to bio-computing=20 [Shi83, Nix83]. These algorithms are = based on finding the longest common subsequence (LCS)[12] for pairs of sequences. The algorithm begins by finding the LCS of the two shortest sequences, and in the following steps takes = the current shortest sequence and finds its LCS with the result of the = previous=20 steps. Although this algorithm is not

can be used for=20 the classification problem by using an appropriate fitness measure.=20 This has been used for instance by Ogiwara et = al. [OUSK92].

## 3.2. Sequence-driven approaches

The common elements of the sequence driven approaches can be=20 summarised as follows:=20

- For sequences $s_1, \ldots, s_k \in S_+$ make sets=20 $P_{\{s_1\}} = 3D\{s_1\}, P_{\{s_2\}} = 3D\{s_2\}, \ldots, = P_{\{s_k\}} = 3D\{s_k\}$.
- Iterate: choose $i$ and $j$ in some given way and=20 combine the sets $P_{S_i}$ and $P_{S_j}$ into a new set=20 $P_{S_i \cup S_j}$ such that $P_{S_i \cup S_j}$ is a set of patterns with high fitness that matches all (or most of the) sequences from $S_i \cup S_j$.=20 In general more than two sets may be combined in one iteration step.=20
- In the end we obtain a set $P_{S_+}$ of patterns conserved in $S_+$ (or = in most of S+).

Various SD methods differ in

1. the way the sets to be combined are chosen (i.e., in the methods for=20 choosing $i$ and $j$ in the iteration step),
2. how the combination is done (dynamic programming, heuristics)=20 and how the (fittest) patterns are chosen,
3. the particular pattern space and the representation of patterns (e.g. local alignments may be used to represent the patterns),
4. whether one, most, or all patterns/alignments are kept.

For example, suppose we are given three sequences:

$$s_1 = 3D \texttt{ AWCEFGHJKLM} \tag{4}$$

$$s_2 = 3D \texttt{ EFGOPAWRJKLS} \tag{5}$$

and=20

$$s_3 = 3D \texttt{ TAWUVOPHJKL} \tag{6}$$

According to the SD approach we make three initial sets of patterns=20 $P_{\{s_1\}} = 3D\{\texttt{AWCEFGHJKLM}\}$,=20 $P_{\{s_2\}} = 3D\{\texttt{EFGOPAWRJKLS}\}$, and=20 $P_{\{s_3\}} = 3D\{\texttt{TAWUVOPHJKL}\}$. Suppose, for instance, that the chosen order for joining these sets is = that=20 first $P_{\{s_1\}}$ and $P_{\{s_2\}}$ are joined and then the result is=20 joined with $P_{\{s_3\}}$. Further, let=20 the method of choosing the common patterns be such that=20 after the first step we get a set of two patterns=20 $P_{\{s_1,s_2\}} = 3D\{\texttt{*AW*JKL*}, \texttt{*EFG*JKL*}\}$ (note that both these patterns match $s_1$ and $s_2$,=20 and that both patterns are the longest in the sense that no extension=20 matches $s_1$ and $s_2$).=20 In the next step, joining $P_{\{s_1,s_2\}}$ with $P_{\{s_3\}}$, the algorithm may detect = that only=20 the first pattern: `*AW*JKL*` is shared by the third sequence and hence will=20 obtain $P_{\{s_1,s_2,s_3\}} = 3D\{\texttt{*AW*JKL*}\}$.=20 This pattern is the fittest in the sense that it is the longest=20 regular pattern which matches all three sequences.=20

Joining the pattern sets can be done, for example, by using=20 dynamic = programming [SK83, Ste94, CLR90].=20 These algorithms can often be constructed so that=20 they guarantee finding the fittest patterns common to the given pair=20 of sets for a given fitness function.=20 If negative examples are given (i.e. the classification problem), these = can be taken into account in the iteration step (e.g. the patterns present=20 in negative examples may be excluded from the resulting set). The problems of pattern discovery and local=20 multiple sequence alignment are very closely related, and=20 some SD algorithms store local alignments instead of patterns during the iteration. =20 which=20 (5) AWCDEFHIJKLM

[12]By a subsequence of two sequences $a_1 \ldots a_n$, and $b_1 \ldots b_m$ we mean a sequence $c_1 \ldots c_k$, such that there exists $i_1 < \ldots < i_k$ and $j_1 < \ldots < j_k$ for which $c_1 \ldots c_k = 3Da_{i_1} \ldots a_{i_k} = 3Db_{j_1} \ldots b_{j_k}$.

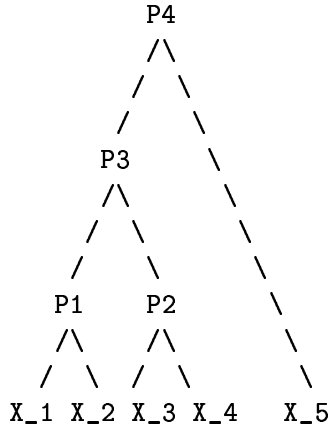


Figure 4: Example dendrogram for sequences=20 $X_1, X_2, X_3, X_4, X_5$. Pairs $X_1$, $X_2$, and $X_3$, $X_4$ are=20 the most similar among themselves,=20 and the sequence=20 $X_5$ is the most different from any of the other sequences.=20 The algorithm aligns=20 $X_1$ to $X_2$, obtaining $P_1$, $X_3$ to $X_4$, obtaining $P_2$, then $P_1$ to $P_2$ obtaining $P_3$, and finally, $P_3$ to $X_5$ obtaining=20 $P_4$.=20 Patterns $P_1$, $P_2$, $P_3$, and=20 $P_4$ =20 match sequences which are below each of them,=20 thus P4 matches all the sequences.

guaranteed to find the LCS of the set of=20 sequences, it has been proved in [Shi83, Nix83] that in some rigorously defined learning model=20 the method will produce the "right" pattern in polynomial-time in the total length of sequences. =20

### 3.2.1. Best pair comparison based heuristics

An algorithm for finding patterns in=20 biosequences based on pairwise comparison is given in Smith and Smith [SS90].=20 This approach uses the fact that pairs of sequences, as well as pairs of sequences and patterns of the type considered,=20 and pairs of patterns, can be aligned by dynamic programming algorithms.=20 The algorithm also exploits the fact that=20 the characters of the basic alphabet (i.e., $\Sigma_p$) can be organised in partially ordered hierarchical groups. hierarchy,

First, an estimated phylogenetic tree (the so called *dendrogram*)=20 is built using the estimated relative distances among the sequences.=20 For instance, a possible dendrogram of sequences =20 $X_1, X_2, X_3, X_4$, $X_5$ where pairs $X_1$, $X_2$, and $X_3$, $X_4$ =20 are the sequences most similar among themselves, but the sequence=20 $X_5$ is the most different from any of the other,=20 is given in figure 4. The pairs (sequence, sequence), or in later stages (sequence, pattern) = or=20 (pattern, pattern) are aligned at each node of the dendrogram starting=20 bottom-up, and a common pattern is obtained from=20 each pair via dynamic programming. =20

The result of aligning two characters is the character denoting the smallest possible group in the hierarchy containing both characters, which may already denote a group of basic characters. The scoring is positive, but decreases with groups higher up in the hierarchy[13]. Gaps are penalised as $w = 3Dw_0 + w_e \cdot k$, where $w_0$ is the gap opening penalty, $w_e$ is the gap extension penalty,

[13]In the AACC hierarchy, a match to a basic character is scored +3, at the next levels +2 and +1, and a match to a wildcard is scored 0.

=20

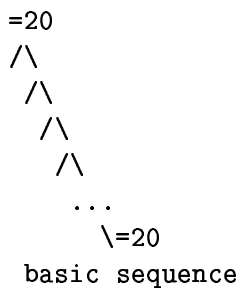


Figure 5: A dendrogram where one sequence is chosen as a = basic=20 sequence=20 and all other sequences are aligned against it

and $k$ is the gap length. If, while aligning a pattern to a pattern, two gaps are aligned, only gap extensions (if needed) are penalised, but not the gap opening.

Note that pairwise alignments are guaranteed to give an optimal (i.e. the most specific) pattern common to the two sequences/patterns aligned, but this does not give any guarantee about the optimality of patterns higher up in the dendrogram with = respect to all given sequences. In addition to the pattern that is common to all sequences, the algorithm also obtains patterns common to subsets of=20 related sequences, therefore the algorithm can be also used for = classification (in fact for unsupervised learning).=20

A different heuristic is developed by = Roytberg [Roy92]. One sequence is selected as the *basic sequence*,=20 and all the other sequences (the so called *serial sequences*) are aligned against it. This approach corresponds to a dendrogram of the type given in=20 figure 5. The algorithm finds the substrings in the basic sequence that have=20 approximate matches in all,=20 or in a specified percentage, of the serial sequences, additionally ensuring that the respective substrings from the serial sequences are = similar to each other[14].

### 3.2.2. All pair comparison heuristics

A heuristic based on finding pairwise similarities between all pairs of=20 sequences is described by Schuler et al. [SAL91].=20 The algorithm begins by comparing all pairs of input=20 sequences. It locates for each pair the substrings that score high=20 enough, thus obtaining the so called 2-blocks[15]. Next it attempts to extend such 2-blocks to three sequences.=20 For this it checks all pairs of 2-blocks having one sequence=20 in common, and 3-blocks are extracted from=20 those with similar enough parts in all three sequences.=20 Then the same idea is applied=20 to 3-blocks to extend them to 4-blocks and so on. Theoretically there may be exponentially many blocks to try, but in practice if the=20 threshold for similarity scores has

[14]Note that the similarities are not necessarily transitive, i.e. the fact that some substring $A$ from the basic sequence is similar to a = =20 substring $B$ in a serial sequence $X_b$ and to a substring $C$ in a serial = sequence=20 $X_c$, does not necessarily mean that $B$ is similar to $C$.

[15]By an *n-block* we mean an array of $n$ substrings of=20 equal length.

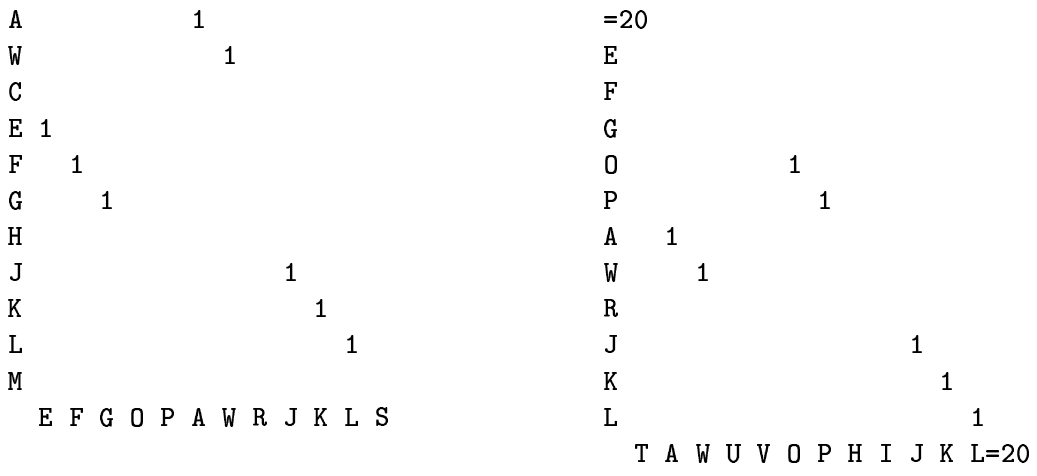


Figure 6: Dot-matrices=20 of strings=20 (4)-(5) and (5)-(6) (0 elements are left blank in the example)

been=20 set high enough, the number of hypothesis is manageable.=20 A very similar approach is also described by Brodsky et al. [BVK+92].

A heuristic representing pairwise alignments by so called *dot-matrices* is described by Vingron and = Argos [VA91].=20 Given a pair of sequences $b_1 \dots b_l$ and $c_1 \dots c_m$, a dot matrix $A = 3D[a_{i,j}]$ is a matrix of size $l \times m$ with = elements=20 $a_{i,j}$ defined as follows: $a_{i,j} = 3D1$ if $b_i = 3Dc_j$, otherwise $a_{i,j} = 3D0$= . For instance,=20 the dot matrices 4-5 and = 5-6=20 for the sequences (4), (5),=20 and (6) are given in figure=20 6.=20

The algorithm calculates dot matrices for all pairs of sequences and filters out similarities (non-zero entries in the matrices) that are not consistent with the other dot matrices by using=20 *Boolean multiplication*[16]. For instance, the result of Boolean multiplication=20 for the matrices in figure 6 is given in figure 7.=20 Note that the resulting matrix is different from 4–6 in that=20 only substrings present in all three sequences, namely=20 `AW` and `JKL` =20 have 1 in the respective positions (and not the matching character `= H`).=20

In general, if sequences=20 $X_1, \dots, X_n$ are given, there exist $n(n-1)/2$ dot=20 matrices $M_{1,1}$, $M_{1,2}$, ..., $M_{1,n}$, $M_{2,2}$, = $M_{2,3}$,=20 ..., ..., $M_{n,n}$. The matrix resulting from the Boolean=20 multiplication $M^*_{k,m} = 3D M_{k,l} \times M_{l,m}$ shows which = substrings are=20 common to all three sequences $X_k$, $X_l$, and $X_m$.=20 By fixing $k$ and $m$, and taking all possible $l$'s (not equal to $k$=20 or $m$) we can find substrings common to all strings.=20 It is also possible to find all substrings=20 common to $X_k$ and $X_l$ and at least a given number of other sequences by similar algebraic matrix manipulations. Vingron and Argos [VA91] describe a heuristic=20 based on such matrix manipulations for finding "significant"=20 (i.e. with relatively high fitness) substrings common to a majority of the sequences.=20

More general dot matrices can also been defined using real instead of boolean values repre-

[16] A *Boolean multiplication* of such matrices is defined the same way as ordinary matrix multiplication, except that instead of=20 summation, the Boolean summation is used (i.e. 0+0=3D0, 0+1=3D1, 1+0=3D1,= =20 and 1+1=3D1).

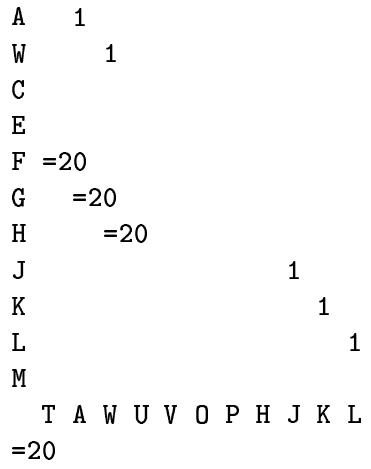


Figure 7: Resulting dot-matrix from Boolean multiplication = =20 (7)–(8) × =20 (8)–(9)

senting=20 the similarity scores between the positions. However, it should be noted that if a non-transitive similarity relation = is=20 used (e.g., defined from PAM or Blosum matrices), the algorithm may find sets of substrings some of which are not similar to each other (the = algorithm guarantees 3-consistencies, but not $k$-consistencies for $k > 3$ [Fre78]).

After the dot matrices have been filtered, a directed graph is = constructed=20 with one node for each possible local alignment of similar substrings, = and edges between all nodes. Two special nodes (source and sink) are added,=20 the source node corresponding to all sequence starts being aligned, and the sink node to all sequence ends being aligned. Each edge=20 $u \rightarrow v$ in the graph is given a weight depending on the = relative=20 positions of the substrings corresponding to nodes $u$ and $v$. Dijkstra's algorithm [AHU83] is used to find the = highest scoring path from the source to the sink node. This path defines a = partial global alignment of the sequences.

It should be noted that an extremely efficient algorithm for finding the longest substrings common to at least $k$ out of $n$ given sequences can be based on *suffix-trees* [McC76, Ukk92, Hui92].=20 This gives an algorithm that is linear-time in the total length=20 of the examples and independent of the length of the patterns. However, no very efficient=20 algorithms=20 are currently known which use suffix trees to find approximately conserved substrings. A related approach use directed acyclic word graphs (DAWGs) instead of suffix trees [**?**].

### 3.2.3. Algorithms for the classification problem

SD algorithms can also be used in the case when both positive and = negative=20 sequences are given (i.e. for the classification problem). This has been done by Kudo et al. [KKASI92].=20 The approach is primarily designed for finding patterns at gene splice-site 5-th end, and since all such sites have a fixed position of = 100% conserved GT, the sequences can be pre-aligned by aligning GT.=20 The target language in [KKASI92] is=20 a union of subwords, either with wildcards (i.e. **B c** and without = $*$), or in the more general case containing arbitrary combinations of

basic characters=20 (i.e. **C c** and without $*$).=20 The algorithm finds the least general set of patterns=20 that covers all the positive examples, and does not contain any negative examples in iterative steps. Each step of the = iteration =20 introduces wildcard characters in order to unify some positive examples, but so that none of the negative = examples is matched. More precisely,=20 it attempts to unify by the introduction of wildcards in non-matching = positions,=20 first pairs of positive examples, then triples =20 from sequences contained in the successful pairs, then quadruples from successful triples, etc, until such unification is no longer=20 possible without the inclusion of any negative examples.=20

The problem of finding patterns from positive and negative examples has also recently been studied by Tateishi et al.=20 [TM95, TMM95].=20 They use a somewhat different definition of the classification=20 problem. The positive and negative examples=20 are provided in pairs $(pos_1, neg_1), (pos_2, neg_2), = \ldots, (pos_n, neg_n)$,=20 and the aim is to find a classification function able to=20 distinguish between $pos_i$ and $neg_i$ for each $i$, but not=20 necessarily between $pos_i$ and $neg_j$ for $i \neq j$.=20 They show that the problem of finding a pattern maximising correctly classified pairs is still NP-hard. A greedy algorithm=20 for approximating the solution for a simple pattern class=20 **C** from pre-aligned sequences having the same length is given.=20 Some heuristics for more complicated patterns are also proposed.=20

therefore we=20 approaches is

### 3.3. Combined approaches

The most obvious way to combine PD and SD approaches is to use=20 SD for refining (expanding or combining) the patterns found by PD = search. This can be done in two ways. The first way is to

- use a PD approach for spotting some candidate patterns,=20
- mark the position of the candidate patterns in the sequences,
- align the sequences so that the positions of candidate pattern are = aligned together,=20
- finally extend the candidate pattern=20 while the fitness of the emerging pattern is increasing.

=20

This method is used in [SAC90, JCH95].

The second way of refining patterns found by the PD approach=20 is by grouping similar=20 patterns together, aligning them, and trying=20 to generalise from them. For instance, if substrings=20 ...`aacaa`... and ...`aagaa` = ...=20 are found to be frequently occurring in sequences, then a common pattern = =20 `aa[cg]aa` can be obtained from them.=20 This kind of refinement is used in=20 [NG94, SS94] = =20

After having combined patterns, Neuwald and Green [NG94] additionally=20 calculate an *initial profile* from the (ungapped) alignment = defined by=20 the substrings matching a combined

pattern=20 (the simplest form of a profile is a position dependent scoring matrix, giving one score to each amino acid for each position in a segment to=20 match the profile).=20 The profile is iteratively refined by realigning the sequences to the profile, throwing away non-significant matches, and recalculating the profile. =20

A modification of the first combined approach is used by Landraud et al. [LAC89]. In this algorithm first of all a variation of KMR is used to find all substrings present in at least $k$ out of the given $n$ substrings. In the next step the substring having "the best" approximate similarities, in some precisely defined sense, in the remaining $n-k$ sequences, is picked out from the substrings found in the first step. The strings are aligned so that the respective substrings are aligned together in all the sequences. After that, the second step is repeated separately for the parts of the sequences that are to the left and to the right of the substrings used in the previous stage. This is repeated while possible, i.e. a divide-and-conquer strategy is used.

Another explicit way of combining PD and SD approaches is described by Ogiwara et al. [OUSK92]. The basic idea is to use a PD approach to find relatively short candidate substrings, to transform the original sequences into different data structures consisting of these substrings joined by 'gaps', and finally to align the data structures obtained and to extract the common patterns. Let us consider this approach in some more detail.

The algorithm is for the classification problem, i.e., it uses both positive and negative examples. All words of the given length are enumerated in a PD manner and a count is taken of in how many positive examples and how many negative examples each is present. In practice tetra-, penta-, and hexapeptide patterns (i.e. substrings of length 4 to 6) are counted. Only those strings that are present in at least $f$ percent of positive examples, and in none of the negative examples are retained; in practice two cases: $f = 3D100\%$ and $f = 3D70\%$ are = considered.

Next the positions of these words and their nearest neighbours are marked in the positive examples. In this case "nearest neighbour" means having no more than one difference (insertion, deletion or substitution). Thus the examples are transformed to new structures of the type: $p_{1,j}\, g_{1,j}\, p_{2,j}\, g_{2,j}\, \ldots p_{n,j}$, where $p_{i,j}$ are the frequent substrings, and $g_{i,j}$ are integers equal to the distance between the starting positions of $i$-th and $i+1$-th substrings in the $j$-th example. In the second stage the transformed examples are aligned by using heuristics of pairwise alignment. The output comprises consensus patterns of the type $p_1 - x(min_1, max_1) - p_2 - x(min_2, max_2) - \ldots - p_n$, where $p_i$ are subwords, and $x(min_i, max_i)$ specifies the minimal and maximal distances (spacers) between the subwords. =20

In Henikoff et al. [HH91] a = combined PD=20 and SD algorithm=20 is developed for finding frequent blocks in protein databases. The first = stage=20 simply uses the algorithm of [SAC90], thus finding patterns and the respective =20 blocks in a PD manner, and then extends them (see the beginning of this subsection). =20 The positions of the patterns are marked on the initial examples.=20 Next the "best" set of patterns which occur in the same order without overlapping in a critical number of=20 sequences is found. Such an ordering is called a *path*. =20 A graph is constructed where nodes represent patterns, and an arc extends=20 from node $b_1$ to $b_2$ if pattern $b_1$ precedes pattern $b_2$ and = does not=20 overlap in at=20 least the critical number of sequences.=20 The graph is searched for the best path according to a defined scoring=20 scheme.=20 scoring=20 scheme see=20 Note that this step is similar to the last step of Vingron and Argos' = algorithm [VA91]. =20

### 3.4 Learning unions of patterns and decision trees

So far we have only considered algorithms for discovering string=20 functions defined by one pattern from a subclass of generalised regular patterns (i.e., functions of type **a** and **b**). However,=20 several algorithms have been reported that are able to discover=20 more complex string functions based on unions of regular patterns or decision-trees over regular patterns, i.e., functions of=20 type **c** and **d**. =20

Algorithms for discovering these classes of functions have been reported in [AMS+93, AKM+92, **?**, **?**], where the authors prove that the classes of these concepts can be *learned* from examples in polynomial time in the sense of inductive inference or PAC-learning. Unfortunately, the order of the=20 polynomials is too high and therefore various heuristics have to be used in practical applications. =20

Arikawa et = al. [AMS+93, AKM+92]=20 consider learning from both positive and negative examples, i.e. the classification problem. In [AMS+93] the method for learning decision trees first introduced by Quinlan [Qui86] is used. In [AKM+92] an=20 algorithm for learning so called *elementary formal systems* has been developed. In practice only a special case of elementary formal systems is used, which in fact is the union of a bounded number of regular patterns. A study is also made in [AMS+93] of how the indexing of the basic alphabet can be performed automatically = so positive=20 and negative that the classification of the positive and negative examples remains = correct. The problem is proved to be NP-hard, and a heuristic for its approximation is given.

learning -=20 $S_k$=20

In [AFSA94] a method for learning the union of an *a priori* bounded number of regular patterns from positive examples only is developed. The algorithm finds the=20 most specific union of pattern languages containing all positive=20 examples. Note that in this approach the fact that the number of=20 patterns in a union is bounded by some *a priori* given constant is essential, as otherwise the algorithm would=20 simply return the union of all sequences in the training set.=20 This sets an *a priori* limit on the number of subfamilies that can be discovered in the sequences.=20 In the noisy data case this also means an assumption about the level of noise.=20 [TSLM+95] has similar limitations. =20

An algorithm for=20 grouping the sequences and filtering out noise without an *a = priori* assumptions on the level of noise and the size of the=20 groups has been developed by Wu and Brutlag [WB95].=20 On the other hand this algorithm requires the sequences to be = pre-aligned.=20 The algorithm uses the so called *beam-search* method for splitting the training set into subfamilies in alternative ways=20 and generating the candidate pattern.=20

The algorithm for discovering unions of an unbounded number of patterns and without any assumption on the level of=20 noise from non-pre-aligned sequences=20 is developed in [BJUV96]. The patterns are=20 of the PROSITE type (i.e., class **E**).=20 The algorithm uses a fitness measure based on=20 the MDL principle (see section 2.2) to balance between how well=20 a set of patterns covers the given examples, and how compact it is. The algorithm is based on the pattern discovery=20 algorithm Pratt [Jon96] and on a greedy=20 set covering algorithm. The greedy algorithm guarantees finding=20 the union of patterns within a logarithmic multiplier=20 to

the optimum. Experiments=20 are reported showing that the algorithm correctly splits a family of=20 biosequences=20 into subfamilies discovering a strong pattern within each of the family, = and=20 thus=20 effectively performs unsupervised learning (see section 2).=20 =20

## 4. Conclusions

The aim of this work has been to give a survey of methods for the automatic discovery of patterns in biological sequences and to establish some systematisation of this area. For this reason we have designed a framework which we believe facilitates the understanding and analysis of approaches to this problem. This framework contains partial formalisations of the problems of pattern discovery and evaluation, and also classifications of pattern languages and algorithmic approaches. We have looked at algorithms for the discovery of deterministic patterns with expressive power inside the regular languages. From this set of algorithms we have chosen to describe some that we consider to be representative of the complete set. We have identified the main algorithmic ideas of each of these methods and have shown how these ideas relate to each other.

While dealing with these problems we have noticed that different authors employ very different computational experiments to test their algorithms and to convince the reader of the usefulness or superiority of their algorithms. The number and lengths of the sequences used, the types of sequence families and the ways in which the results are presented vary greatly, although the algorithms are frequently intended to solve the same problem. We believe that it would be beneficial for the field if an attempt were to be made to establish some systematisation regarding which experiments could be used for testing the algorithms. This would be useful for comparing the algorithms and choosing the best algorithm for a particular problem. It may also be useful to adopt some standard way of presenting those experimental results about new sequence families which have not yet been explored by earlier methods.

Although this survey shows that many nontrivial and efficient pattern discovery algorithms have been developed recently, biologists need considerably more powerful algorithms for efficient knowledge discovery in the growing volume of the biosequence data. Algorithms are required which are able to discover more complicated or subtle patterns in larger training sets containing unknown levels of noise. We hope that this survey will help to move the field forward towards these aims.

### Acknowledgements

The authors wish to thank Rickard Lathrop, Darrel Conklin, and Rein Aasland for careful reading and commenting earlier versions of this paper. Alvis Brāzma has been supported by the Finnish Centre for International Mobility (CIMO), the Latvian Council of Sciences (Grant Number 93.593), the Royal Society and the Human Capital and Mobility programme of the European Union. Inge Jonassen and Ingvar Eidhammer have been supported by grants from the Norwegian Research Council. David Gilbert was supported by a grant from the British Council.

# Appendix A: Algorithms and software

Key

| **Algorithms** | |
|---|---|
| Pattern | Pattern type (see section 2.2.4) |
| Pre | Prealigned [Y/N] |
| G | Guaranteed [Y/N] (see section 2.4) |
| +/- | Uses positive and/or negative example training sets |
| Domain | DNA/protein/Not Applicable |

| **Software** | |
|---|---|
| Name | Name of the software |
| Src/Ex | Source or executable |
| Platform | Runs on what platform |
| Obtain | Obtain from: a/ftp=anonymous ftp; A=authors; n/a=not available |

| Authors | Algorithms | | | | | Software | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pattern | Pre | G | +/- | Domain | Name | Src/Ex | Platform | Obtain |
| [Nix83] | Fa | N | N | + | N/A | | | | |
| [Shi83] | Fa | N | Y | + | N/A | | | | |
| [WAG84] | Ab | N | Y | + | protein, DNA | | | | |
| [LAC89] | Fb | N | N | + | protein, DNA | | | | |
| [Sta89b] | Ab | N | Y | + | DNA | Fortran77 | Vax VMS | n/a | |
| [SS90] | Ea | N | N | + | protein | | | | |
| [SAC90] | Ba | N | Y | + | protein | MOTIF | Turbo-C | IBMPC | n/a |
| [VA91] | Fa [FILLOG/SUM]; Fb [FILMAXAV] | N | N | + | protein | unkown | | | A |
| [KKASI92] | Ba, Ca | Y | Y | +/- | DNA | | | | |
| [OUSK92] | Fa | N | Y/N | +/- | protein | | | | |
| [Roy92] | Ab | N | N | + | protein, DNA | MuSCo | | IBMPC, IBM/370 | n/a avail |
| [AMS+93] | Fd | N | N | +/- | protein | | | | |
| [NG94] | Ca | N | N | + | protein | ASSET | Src | SPARC2 | a/ftp |
| [SS94] | Ca | N | N | + | protein | | | | |
| [WMS+94] | Fb | N | N | + | protein | DISCOVER, CLASSIFY | Ex | DOS, DEC Ultra, SunSPARC | A |
| [JCH95] | Ea | N | Y/N | + | protein | Pratt | Src (C) | dec-alpha, sparc10 | a/ftp |
| [Jon96] | Ea | N | N | + | protein | Pratt2 | Src (C) | dec-alpha, sparc10 | a/ftp |
| [SVS95b] | Ca | N | Y | + | protein | | | | |
| [SVS95a] | Ab | N | Y | + | protein, DNA | | | | |
| [TSLM+95] | Ec, Ed | N | N | +/- | protein | BONSAI | | | n/a |
| [WB95] | Ca | Y | N | + | protein | SEQCLASS | x, Common Lisp | Sun SPARC | n/a |

## Appendix B: Input sequences (positive examples) of patterns discovered by some of the reported algorithms

**Note:** We base our notation on that of PROSITE, augmented with some additional symbols.

### 1. Staden [Sta89b]

Examples: 88 *E.coli* promoter sequences, varying in length from 47 to 64, having a total length of 5238 characters.
The patterns found most frequently to be approximately present in the sequences are:

t-t-t-t-t-t

t-t-a-t-a-a
t-t-g-a-c-a
t-c-t-t-g-a
t-a-t-a-a-t
a-c-t-t-t-a
a-a-a-a-a-a
a-g-t-a-t-a

### 2. Smith and Smith [SS90]

Examples: 128 sequences of length between 141 and 147 from hemoglobin delta epsilon gamma beta major-chain sequences.
Pattern:

l-l-x(2)-a-x(3)-b-x(2)-c-x(5)-G-x-l-x-a-x-l-c-c-a-a-c-P-W-l-l-R-b- F-x(2)-F-G-x-c-x-l-x(3)-a-x(2)-l-x(2)-a-x(3)-G-x-i-a-x(3)-c-x(3)-c- x-l-c-l-x-a-x(3)-c-x(2)-L-S-l-x-H-x(3)-c-x(2)-l-x(2)-l-F-l-x-c-G- x(2)-c-a-x(2)-c-x(7)-F-x(4)-l-x(2)-c-l-i-c-x(3)-a-x(2)-p-L-x(3)-Y

Examples: 12 sequences from Trypsinogen/Venom serine proteases.
Pattern:

l-l-l-h-x-a-a-G-G-x(2)-C-x(2)-l-x(2)-P-b-x(3)-c-x(4)-i-x(0,1)-F-C- G-x-k-L-l-x(3)-W-V-a-k-A-p-H-C-x-l-x(2)-c-l-a-i-L-G-l-x(6)-l-x(2)- E-x-c-x(6)-c-x(2)-P-l-x-l-x(3)-c-l-l-x(0,1)-T-l-c-L-l-i-L-x(4)-l-x- l-a-x(2)-a-x-L-P-l-x(5)-G-l-x(3)-a-x-G-W-G-x(3)-l-g-x(5)-l-x(2)-l-C- x-l-x(2)-a-c-x-l-x(2)-C-l-x(2)-Y-x-G-x(0,1)-a-x(2)-l-x-c-C-x-G-c-c- l-G-G-x-D-k-C-x-G-D-S-G-G-P-a-a-x-l-G-x-c-Q-G-a-a-S-W-G-x(2,3)-C-A- x(4)-P-p-c-x(2)-l-V-c-l-b-a-x-W-l-l-l-x-a-A

The lower case letters denote classes from the AACC hierarchy as follows: a=[ILV], b=[FWY], c=[ILVFWYCM], h=[DE], i=[HKR], j=[NQ], k=[ST], l=[DEHKRNQSTBZ], p=[AG].

### 3. Smith et al. [SAC90]

Examples: 15 sequences from DNA integrases.
Pattern:

x(15)-H-x-L-R-H-x(2)-A-x(6)-G-x(6)-Q-x(2)-L-G-H-x(2)-l-x(2)-T-x(2)-Y-x(5)

### 4. Kudo et al. [KKASI92]

Positive examples: 496 pre-aligned DNA segments of length 9 from around the $5'$ splice site (three in the exon and six in the intron).
Negative examples: 1123 DNA segments of length 9 (all containing gt in position 4-5).
Some of the best patterns discovered are (in class B):

<x-a-g-g-t-a-a-x-x>
<a-a-g-g-t-x-a-g-x>
<c-x-x-g-t-a-a-g-x>

and (in class C)

<x-[agc]-[agc]-g-t-a-a-g-x>
<[agc]-x-[agc]-g-t-a-a-g-x>
<x-[agc]-x-g-t-a-a-g-[tgc]>

### 5. Ogiwara et al. [OUSK92]

Examples: sequences from cytochrome b5 family
A partially conserved pattern found:

H-P-G-G-E-E-V-L

Examples: sequences from a family of L-lactate dehydrogenase
A partially conserved pattern found:

P-V-D-[IV]-L-x(47)-G-[EQ]-H-G-D

Examples: sequences in a family of glyceraldehyde-3-phosphate dehydrogenases
A completely conserved pattern found:

G-F-G-R-I(0,1)-G-R-x(129,134)-S-N-A-S-C-T-T-N-[CS]-L-A-P- x(14)-[LM]-M-T-T-V-H-x(30,31)-T-G-A-A-[KR]-A-[VT]-x(92,95)- [SA]-W-Y-D-N-E

### 6. Saqi and Sternberg [SS94]

Examples: a set of heat shock proteins
Some of the patterns found:

x-G-G-G-T-F-D-[ILV]-[ST]-[ILV]
x-[ILV]-[FWY]-D-L-G-G-G-T-F-D-[ILV]
D-[LF]-G-G-G-T-F-D

Examples: a set of toxin proteins
Some of the patterns found:

x(2)-C-C-x(4)-C-x
D-R-C-C-x(2)-H-D-x-C

### 7. Neuwald and Green [NG94]

Examples: a set of 56 sequences with an average length 471 from acyltransferases
Some of the patterns found:

V-x-P-x(2)-[RQ]-x(4)-G-x(2)-L-[LM]
N-x(2)-A-x(3)-Y-x(3)-G-F

### 8. Wang et al. [WMS$^+$94]

Examples: 47 sequences of length 190-780 in a group of cyclic proteins
Some of the patterns found:

L-Q-L
I-A-S-K-Y-E-E
D-T-A-G-Q-E-*-L-V-G-N-K

**9. Sagot et al [SVS95a]**

Examples: 80 proteins belonging to the elongation family
46 patterns found

**10. Shoudai et al. [TSLM$^+$95]**

Examples: 3796 signal peptides indexed to three-letter alphabet $\Sigma_{hydro}$ of maximum length 32. Classified in three groups of sizes 2205, 640, and 603, by patterns:

2-*-2-*-0-2-*-1-*-0-2
1-0-*-0-*-0-*-2-1-*-0
2-2-2-*-1-2-*-1-2

where 0,1,2 stands for different amino acid groups in different patterns (see [TSLM$^+$95] for details).

**11. Jonassen et al. [JCH95]**

Examples: 241 protein sequences from the zinc finger c2h2 family, average length 393
Pattern:

C-x(2,4)-C-x(3)-[ILVFYC]-x(8)-H-x(3,5)-H

Examples: 164 protein sequences from the snake toxin family, average length 64
Pattern:

G-C-x(1,3)-C-P-x(8,10)-C-C-x(2)-[EPDN]

Examples: 27 protein sequences containing PHD finger, average length 874
Pattern:

C-x(2,4)-C-[YCEPGSDNQR]-x-[VMFWHTAPGSN]-x-H-x(2)-C-[ILVMFYHTCA]-x(11)- [YWCEPGSDNQ]-x(2)-[IFHCAPGSDN]

## References


[AFSA94] H. Arimura, R. Fujino, T. Shinohara, and S. Arikawa. Protein Motif Discovery from Positive Examples by Minimal Multiple Generalization over Regular Patterns. In *Proc. of the 5th Genome Informatics Workshop*, pages 39–48, 1994.

[AHU83] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *Data Structures and Algorithms.* Addison-Wesley, 1983.

[AKM+92] S. Arikawa, S. Kuhara, S. Miyano, A. Shinohara, and T. Shinohara. A learning algorithm for elementary formal systems and its experiments on identification of transmembrane domains. In *Proc. 25th Hawaii Int. Conf. on System Sci.*, pages 675–684, 1992.

[AMS+93] S. Arikawa, S. Miyano, A. Shinohara, S. Kuhara, Y. Mukouchi, and T. Shinohara. A Machine Discovery from Amino Acid Sequences by Decision Trees over Regular Patterns. *New Generation Computing*, pages 361–375, 1993.

[Bai92] A. Bairoch. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Research*, 20:2013–2018, 1992.

[BB94] P. Bucher and A. Bairoch. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. In *Proc. of Second International Conference on Intelligent Systems for Molecular Biology*, pages 53–61, 1994.

[BCHM94] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. M. McClure. Hidden Markov Models of Biological Primary Sequence Information. *Proc. Natl. Acad. Sci USA*, 91:1059–1063, Feb 1994.

[BJUV96] A. Brazma, I. Jonassen, E. Ukkonen, and J. Vilo. Discovering patterns and subfamilies in biosequences. In *Proc. of Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 34–43. AAAI Press, 1996.

[BUV96] A. Brazma, E. Ukkonen, and J. Vilo. Discovering Unbounded Unions of Regular Pattern Languages from Positive Examples. In *Proceedings of 7th Annual International Symposium on Algorithms and Computation (ISAAC-96), Lecture Notes in Computer Science*, page (in press), December 1996.

[BVK+92] L. I. Brodsky, A. V. Vassilyev, Y. L. Kalaydzidis, Y. S. Osipov, R. L. Tatuzov, and S. I. Feranchuk. Genebee: the program package for biopolymer structure analysis. In S. Gindikin, editor, *Mathematical methods of analysis of biopolymer sequences, DIMACS series in discrete mathematics and theoretical computer science, volume 8*. American Mathematical Society, 1992.

[CLR90] T. H. Cormen, C. E. Leicerson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

[CWC92] S. C. Chan, A. K. C. Wong, and D. K. Y. Chiu. A Survey of Multiple Sequence Comparison Methods. *Bulletin of Mathematical Biology*, 54(4):563–598, 1992.

[Day78] M. O. Dayhoff. *Atlas of protein sequence and structure*, volume 5. National Biomedical Research Foundation, 1978.

[Fre78] Eugene C. Freuder. Synthesizing constraint expressions. *Communications of the ACM*, 21(11):958–966, 1978.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, NY, 1979.

[GME87] M. Gribskov, M. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A*, 84:4355–4358, 1987.

[HH91] S. Henikoff and J. G. Henikoff. Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, 19(23):6565–6572, 1991.

[HH92] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:100915–100919, 1992.

[Hui92] L. C. K. Hui. Color Set Size Problem with Application to String Matching. In *Proc. of Combinatorial Pattern Matching*, pages 230–243. Springer-Verlag, 1992.

[Hut94] A. Hutchinson. *Algorithmic learning.* Clarendon Press, 1994.

[JCH95] I. Jonassen, J. F. Collins, and D. G. Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Science*, 4(8):1587–1595, 1995.

[JHH96] I. Jonassen, C. Helgesen, and D. G. Higgins. Scoring function for pattern discovery programs taking into account sequence diversity. Reports in Informatics 116, Dept. of Informatics, University of Bergen, February 1996.

[Jon96] I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. Reports in Informatics 118, Dept. of Informatics, University of Bergen, March 1996.

[KBM+94] A. Krogh, M. Brown, I. S. Mian, K. Sjoelander, and D. Haussler. Hidden Markov model in computational biology. Applications to protein modelling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

[KKASI92] M. Kudo, S. Kitamura-Abe, M. Shimbo, and Y. Iida. Analysis of context of $5'$-splice site sequences in mammalian mRNA precursors by subclass method. *CABIOS*, 8(4):367–376, 1992.

[KLP92] T. Kristensen, R. S. Lopez, and H. Prydz. An estimate of the sequencing error frequency in the DNA sequence databases. *DNA Seq.*, 2:343–346, 1992.

[KMR72] R. M. Karp, R. E. Miller, and A. L. Rosenberg. Rapid identification of repeated patterns in strings, trees and arrays. In *4th ACM Symposium on Theory of Computing*, pages 125–136, 1972.

[LAB+93] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science*, 262:208–214, Oct 1993.

[LAC89] A. M. Landraud, J-F. Avril, and P. Chretienne. An Algorithm for Finding a Common Structure Shared by a Family of Strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):890–895, Aug 1989.

[LV93] M. Li and P. Vitanyi. *An introduction to Kolmogorov complexity and its applications.* Springer-Verlag, New York, 1993.

[LWS+93] R. Lathrop, T. Webster, R. Smith, P. Winston, and T. Smith. Integrating AI with sequence analysis. In L. Hunter, editor, *Artificial Intelligence and Molecular Biology*, pages 211–258. AAAI Press/The MIT Press, 1993.

[McC76] E. M. McCreight. A space–economical suffix tree construction algorithm. *Journal of the ACM*, 23:262–272, 1976.

[NG94] A. F. Neuwald and P. Green. Detecting Patterns in Protein Sequences. *Journal of Molecular Biology*, 239:689–712, 1994.

[Nix83] R. P. Nix. *Editing by Example.* PhD thesis, Yale University, Xerox Palo Alto Research Center, California, USA, August 1983.

[OUSK92] A. Ogiwara, I. Uchiyama, Y. Seto, and M. Kanehisa. Construction of a dictionary of sequence motifs that characterize groups of related proteins. *Protein Engineering*, 5(6):479–488, 1992.

[Qui86] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[QWK82] C. Queen, M. N. Wegman, and L. J. Korn. Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. *Nucleic Acids Research*, 10:449–456, 1982.

[Ris78] J. Rissanen. Modeling by the shortest data description. *Automatica-J.IFAC*, 14:465–471, 1978.

[Roy92] M. A. Roytberg. A search for common patterns in many sequences. *CABIOS*, 8(1):57–64, 1992.

[SA95] T. Shinohara and S. Arikawa. Pattern inference. In K. P. Jantke and S. Lange, editors, *Algorithmic learning for knowledge-based systems, GOSLER final report*, pages 259–291. Springer-Verlag, 1995.

[SAC90] H. O. Smith, T. M. Annau, and S. Chandrasegaran. Finding sequence motifs in groups of functionally related proteins. In *Proc. Natl. Acad. Sci. USA*, pages 826–830, Jan 1990.

[SAL91] G. D. Schuler, S. F. Altschul, and D. J. Lipman. A workbench for multiple alignment construction and analysis. *PROTEINS: Structure, Function, and Genetics*, 9:180–190, 1991.

[SD95] R. F. Sewell and R. Durbin. Method for calculation of probability of matching a bounded regular expression in a random data string. *Journal of Computational Biology*, 2:25–31, 1995.

[Shi83] T. Shinohara. Polynomial time inference of extended regular pattern languages. *LNCS*, 147:115–127, 1983.

[SK83] D. Sankoff and J. B. Kruskal. *Time warps: string edits, and macromolecules: the theory and practice of sequence comparison.* Addison-Wesley, 1983.

[SNO95] M. Suyama, T. Nishioka, and J. Oda. Searching for common sequence patterns among distantly related proteins. *Protein Engineering*, 8(11):1075–1080, 1995.

[SS90] R. F. Smith and T. F. Smith. Automatic generation of primary sequence patterns from sets of related protein sequences. In *Proc. Natl. Acad. Sci. USA*, pages 118–122, Jan 1990.

[SS94] M. A. S. Saqi and M. J. E. Sternberg. Identification of sequence motifs from a set of proteins with related function. *Protein Engineering*, 7(2):165–171, 1994.

[Sta89a] R. Staden. Methods for calculating the probabilities of finding patterns in sequences. *CABIOS*, 5:89–96, 1989.

[Sta89b] R. Staden. Methods for discovering novel motifs in nucleic acid sequences. *CABIOS*, 5(4):293–298, 1989.

[Ste94] G. A. Stephen. *String searching algorithms.* World Scientific, 1994.

[SVS95a] M-F. Sagot, A. Viari, and H. Soldano. A distance-based block searching algorithm. In C. Rawlings et al, editor, *Proc. of Third International Conference on Intelligent Systems for Molecular Biology*, pages 322–331, Menlo Park, California, July 1995. AAAI Press.

[SVS95b] M-F. Sagot, A. Viari, and H. Soldano. Multiple sequence comparison: a peptide matching approach. In Z. Galil and E. Ukkonen, editors, *Proc. of 6th Annual Symposium, CPM (Lecture Notes in Computer Science 937)*, pages 366–385. Springer, July 1995.

[Tay86] W. R. Taylor. Identification of protein sequence homology by consensus template alignment. *Journal of Molecular Biology*, 188:233–258, 1986.

[TM95] E. Tateishi and S. Miyano. A greedy strategy for finding motifs from positive and negative examples. Technical Report RIFIS-TR-CS-118, Research Institute of Fundamental Information Science, Kyushu University, Japan, Aug 1995.

[TMM95] E. Tateishi, O. Maruyama, and S. Miyano. Extracting best consensus motifs from positive and negative examples. Technical Report RIFIS-TR-CS-115, Research Institute of Fundamental Information Science, Kyushu University, Japan, Aug 1995.

[TSLM+95] M T. Shoudai, M. Lappe, S. Miyano, A. Shinohara, T. Okazaki, S. Arikava, T. Uchida, S. Shimozono, T. Shinohara, and S. Kuhara. BONSAI Garden: parallel knowledge discovery system for amino acid sequences. In C. Rawlings et al, editor, *Proc. of Third International Conference on Intelligent Systems for Molecualar Biology*, pages 359–366, Melono Park, California, Jouly 1995. AAAI Press.

[Ukk92] E. Ukkonen. Constructing Suffix Trees On–line in Linear Time. *Information Processing 92*, 1:484–492, 1992.

[VA91] M. Vingron and P. Argos. Motif Recognition and Alignment for Many Sequences by Comparison of Dot–matrices. *Journal of Molecular Biology*, 218:33–43, 1991.

[Val84] G. L. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[WAG84] M. S. Waterman, R. Arratia, and D. J. Galas. Pattern Recognition in Several Sequences: Consensus and Alignment. *Bulletin of Mathematical Biology*, 46(4):515–527, 1984.

[WB95] T. D. Wu and D. L. Brutlag. Identification of protein motifs using conserved amino acid properties and partitioning techniques. In C. Rawlings et al, editor, *Proc. of Third International Conference on Intelligent Systems for Molecular Biology*, pages 402–410, Melno Park, California, July 1995. AAAI Press.

[WFHW96] F. Wolferstetter, K. French, G. Herrmann, and T. Werner. Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *CABIOS*, 12(1):71–80, 1996.

[WJ94] L. Wang and T. Jiang. One the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994.

[WMS+94] J. T. L. Wang, T. G. Marr, D. Shasha, B. A. Shapiro, and G-W. Chirn. Discovering active motifs in sets of related protein sequences and using them for classification. *Nucleic Acids Research*, 22(14):2769–2775, 1994.