

# Probabilistic Arithmetic Automata and their Applications\*

Tobias Marschall<sup>†</sup>   Inke Herms<sup>‡</sup>   Hans-Michael Kaltenbach<sup>§</sup>  
 Sven Rahmann<sup>†</sup>

November 29, 2010

**Abstract.** We present *probabilistic arithmetic automata (PAAs)*, a general model to describe chains of operations whose operands depend on chance, along with two different algorithms to exactly calculate the distribution of the results obtained by such probabilistic calculations. PAAs provide a unifying framework to approach many problems arising in computational biology and elsewhere. Here, we present five different applications, namely (1) pattern matching statistics on random texts, including the computation of the distribution of occurrence counts, waiting time and clump size under HMM background models; (2) exact analysis of window-based pattern matching algorithms; (3) sensitivity of filtration seeds used to detect candidate sequence alignments; (4) length and mass statistics of peptide fragments resulting from enzymatic cleavage reactions; and (5) read length statistics of 454 sequencing reads. The diversity of these applications indicates the flexibility and unifying character of the presented framework.

While the construction of a PAA depends on the particular application, we single out a frequently applicable construction method for pattern statistics: We introduce *deterministic arithmetic automata (DAAs)* to model deterministic calculations on sequences, and demonstrate how to construct a PAA from a given DAA and a finite-memory random text model. We show how to transform a finite automaton into a DAA and then into the corresponding PAA.

---

\*Parts of this article have been published in conference proceedings [29, 47, 23, 48, 50]. An extended version of one of these articles [50] has been submitted to a journal. A preprint is available on arXiv [49].

<sup>†</sup>Bioinformatics for High-Throughput Technologies, Computer Science 11, TU Dortmund, 44221 Dortmund, Germany

<sup>‡</sup>Genome Informatics, Faculty of Technology, Bielefeld University, 33501 Bielefeld, Germany

<sup>§</sup>Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland

# 1 Introduction

In many applications, processes can be modeled as chains of operations working on operands that are drawn probabilistically. As an example, let us consider a simple dice game. Suppose you have a bag containing three dice, a 6-faced, a 12-faced, and a 20-faced die. Now a die is drawn from the bag, rolled, and put back. This procedure is repeated  $n$  times. In the end one may, for example, be interested in the distribution of the maximum number observed. Many variants can be thought of, for instance, we might start with a value of 0 and each die might be associated with an operation, e.g. the spots seen on the 6-faced die might be subtracted from the current value and the spots on the 12-faced and 20-faced dice might be added. In addition to the distribution of values after  $n$  rolls, we can ask for the distribution of the waiting time for reaching a value above a given threshold.

The goal of this article is to establish a general formal framework, referred to as *probabilistic arithmetic automata (PAAs)*, to directly model such systems and answer the posed questions. We emphasize that we are not interested in simulation studies or approximations to these distributions, but in an exact computation up to machine accuracy. Further, we show that problems from diverse applications, especially from computational biology, can be conveniently solved with PAAs in a unified way, whereas they are so far treated heterogeneously in the literature. This article is a substantially revised and augmented version of several extended abstracts that introduced the PAA framework [47] and outlined some of the applications presented here [29, 23, 48, 50].

Let us give an overview of the application domains of PAAs considered in this paper. We begin with the field of pattern matching statistics. Biological sequence analysis is often concerned with the search for structure in long strings like DNA, RNA or amino acid sequences. Frequently, “search for structure” means to look for patterns that occur very often. An important point in this process is to sensibly define a notion of “very often”. One option is to consult the statistical significance of an event: Suppose we have found a certain pattern  $k$  times in a given sequence. What is the probability of observing  $k$  or more matches just by chance? The answer to this question depends on the used null model, i.e. the notion of “by chance”. It turns out that the PAA framework paves the way to using quite general null models; finite-memory text models as used in this article comprise i.i.d. models, Markovian models of arbitrary order and character-emitting hidden Markov models (HMMs).

The PAA framework can also be applied to the exact analysis of algorithms. Traditionally, best case, average case, and worst case behavior of algorithms are considered. In contrast, we construct PAAs to compute the whole exact distribution of costs of arbitrary window-based pattern matching algorithms like Horspool’s or Sunday’s algorithm. For these algorithm, we present (perhaps surprising) exemplary results on short patterns and moderate text lengths.

Another application arises when searching for a (biological) query sequence, such as a DNA or protein sequence, in a comprehensive database. The goal is to quickly retrieve all sufficiently similar sequences. Heuristic methods use so-called *alignment seeds* in order to first detect candidate sequences which are then investigated more carefully. To

evaluate the quality of such a seed, one computes its sensitivity or hitting probability, i.e. the fraction of all desired target sequences it hits. Similarly, we ask which fraction of non-related sequences is hit by a seed by chance, as this quantity directly translates into unnecessary subsequent alignment work. The stochasticity arises from directly modelling alignments of similar (or non-similar) sequences rather than modelling the sequences themselves. We use the PAA framework to compute the match distribution and, in particular, the sensitivity of certain filtration seeds under finite-memory null models.

Next, we investigate protein identification by mass spectrometric analysis. If one can describe the typical length and mass of peptide fragments measured by the mass spectrometer, one can define a reliable comparison of so-called peptide mass fingerprints, based on an underlying null model. In this article, we compute the joint length-mass distribution of random peptide fragments. Furthermore, we calculate the occurrence probability of fragments in a given mass range, i.e., the probability that cleaving a random protein of a given length yields at least one fragment within this mass range. This probability aids the interpretation of mass spectra as it gives the significance of a measured peak. Again, our framework permits to use arbitrary finite-memory protein models.

The final application concerns DNA sequencing: The task of determining a DNA sequence has seen great technological progress over the last decades. For a particular sequencing technology (“454 sequencing”), the length of a sequenced DNA fragment depends on the order of its characters. We compute the exact length distribution of such fragments. With this distribution we can specify the technology settings that yield the longest reads on average if statistical properties of the genome under consideration are known, improving the sequencing performance by up to 10%.

Most of the mentioned applications have, in some form or another, previously been discussed in the literature, but never been identified as instances of the same abstract scheme. The contribution of this article is twofold. On the one hand, we introduce a generic framework unifying the view on the presented applications. On the other hand, we show that, in all considered applications, not only known results can be reproduced using PAAs, but new achievements are made. Since the range of applications is quite diverse, we give further references to relevant literature in each section separately.

**Organization of the Article.** In the first part of this article we set up the PAA framework. Specifically, we formally introduce PAAs in Section 2 and give generic algorithms in Section 3. We consider waiting time problems on PAAs in Section 4. In Section 5, *deterministic arithmetic automata* and *finite-memory text models* are defined and shown to be a convenient means of specifying a PAA.

Having the framework in place, we explore several application domains. Section 6 deals with the statistics of patterns on random texts. In Section 7, PAAs are employed for the analysis of window-based pattern matching algorithms. We cover the field of alignment seed statistics in Section 8. Then, in Section 9, we apply PAAs to mass statistics of fragments resulting from enzymatic digestion. The optimization of read

lengths in 454 sequencing is discussed in Section 10. We conclude the article with a summarizing discussion in Section 11.

The relations between this article and preliminary versions published in conference proceedings are as follows. Sections 2, 3 and parts of Section 6 are based on [47]. Section 5 is based on [50]. Section 6.4 contains material from [48]. Sections 7, 8, and 9 are based on [50], [23], and [29], respectively.

An extended version of [50] has been submitted for consideration in a special issue of a journal, a preprint is available as [49]. There, the analysis of pattern matching algorithms is further generalized and applied to more algorithms. Here, we include a basic example in Section 7 to give another example of the utility of PAAs.

**Notation and Conventions.** The natural numbers (without zero) are denoted  $\mathbb{N}$ ; to include zero, we write  $\mathbb{N}_0$ . Throughout the article,  $\Sigma$  is a finite alphabet; as usual,  $\Sigma^*$  denotes the set of all finite strings. Indices are zero-based, i.e.  $s = s[0] \dots s[|s| - 1]$  for  $s \in \Sigma^*$ . Substrings, prefixes, and suffixes are written  $s[i \dots j] := s[i] \dots s[j]$ ,  $s[..i] := s[0] \dots s[i]$ , and  $s[i..] := s[i] \dots s[|s| - 1]$ , respectively. All stochastic processes considered in this article are discrete. Therefore, appropriate probability spaces can always be constructed; we do not clutter notation by stating them explicitly. By  $\mathbb{P}$ , we refer to a probability measure;  $\mathcal{L}(X)$  denotes the distribution of the random variable  $X$ . Iverson brackets are written  $\llbracket \cdot \rrbracket$ , i.e.  $\llbracket A \rrbracket = 1$  if the statement  $A$  is true and  $\llbracket A \rrbracket = 0$  otherwise.

## 2 Probabilistic Arithmetic Automata

In this section, we define *probabilistic arithmetic automata (PAAs)* in order to formalize chains of operations with probabilistic operands. PAAs can be interpreted as generalized Markov Additive Processes (MAP) [18, 19] in the discrete case.

**Definition 2.1** (Probabilistic Arithmetic Automaton). A probabilistic arithmetic automaton  $\mathcal{P}$  is a tuple

$$\mathcal{P} = (\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu = (\mu_q)_{q \in \mathcal{Q}}, \theta = (\theta_q)_{q \in \mathcal{Q}}),$$

where

- $\mathcal{Q}$  is a finite set of states,
- $q_0 \in \mathcal{Q}$  is called start state,
- $T : \mathcal{Q} \times \mathcal{Q} \rightarrow [0, 1]$  is a transition function with  $\sum_{q' \in \mathcal{Q}} T(q, q') = 1$  for all  $q \in \mathcal{Q}$ , i.e.  $(T(q, q'))_{q, q' \in \mathcal{Q}}$  is a stochastic matrix,
- $\mathcal{V}$  is a set called value set,
- $v_0 \in \mathcal{V}$  is called start value,
- $\mathcal{E}$  is a finite set called emission set,

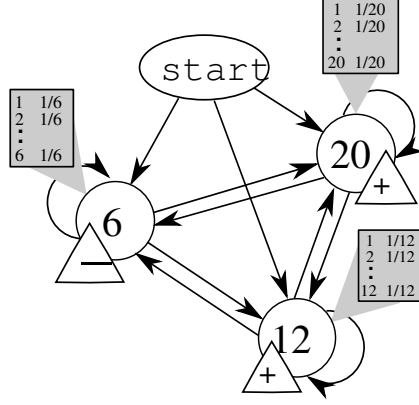


Figure 1: Illustration of a PAA for the dice example. Each of the three dice is represented by a state (circles). Emission distributions are depicted as gray boxes. Each arrow stands for a possible state transition, each transition having a probability of  $1/3$ . The state's operations are given in triangles next to the state. For the start state, emission distribution and operation are not drawn.

- each  $\mu_q : \mathcal{E} \rightarrow [0, 1]$  is an emission distribution associated with state  $q$ ,
- each  $\theta_q : \mathcal{V} \times \mathcal{E} \rightarrow \mathcal{V}$  is an operation associated with state  $q$ .

We attach the following semantics: At first, the automaton is in its start state  $q_0$ , as for a classical deterministic finite automaton (DFA). In a DFA, the transitions are triggered by input symbols. In a PAA, the transitions are purely probabilistic;  $T(q, q')$  gives the chance of going from state  $q$  to state  $q'$ . Note that the tuple  $(\mathcal{Q}, T, \delta_{q_0})$  defines a Markov chain on state set  $\mathcal{Q}$  with transition matrix  $T$ , where the initial distribution  $\delta_{q_0}$  is the Dirac distribution assigning probability 1 to  $\{q_0\}$ .

While going from state to state, a PAA performs a chain of calculations on a set of values  $\mathcal{V}$ . It starts with value  $v_0$ . Whenever a state transition is made, the entered state, say state  $q$ , generates an emission from  $\mathcal{E}$  according to the distribution  $\mu_q$ . The current value and this emission are then subject to the operation  $\theta_q$ , resulting in the next value from the value set  $\mathcal{V}$ . Notice that the Markov chain  $(\mathcal{Q}, T, \delta_{q_0})$ , together with the emission set  $\mathcal{E}$  and the distributions  $\mu = (\mu_q)_{q \in \mathcal{Q}}$ , defines a hidden Markov model (HMM). In the context of HMMs, however, the focus usually rests on the sequence of emissions, whereas we are interested in the value resulting from a chain of operations on these emissions.

By introducing PAAs, we emphasize that many applications can naturally be modelled as a chain of operations whose result is of interest. When compared to Markov chains, PAAs do not offer an increase in expressive power. In fact, from a theoretical point of view, every PAA might be seen as a Markov chain on the state space  $\mathcal{Q} \times \mathcal{V}$ . Thus, we advocate PAAs not because of their expressive power but for their merits as a modelling technique. As we shall see, the framework lends itself to many applications and often

allows simple and intuitive problem formulations. Before we formalize the introduced semantics in Definition 2.3, let us come back to the dice example from Section 1.

**Example 2.2** (Dice). *We model each of the three dice as a PAA state. All transition probabilities equal  $1/3$  and the emissions have uniform distributions over the number of faces of the respective dice. If we are interested in the maximum, each state's operation is the maximum. In general, we can associate individual operations with each state, for instance: “sum the numbers from the 12- and 20-faced dice and subtract the numbers seen on the 6-faced die”. The value set is  $\mathbb{Z}$ , the start value is naturally  $v_0 = 0$ . The corresponding PAA is illustrated in Figure 1.*

**Definition 2.3** (Stochastic processes induced by a PAA). *For a given PAA  $\mathcal{P} = (\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu, \theta)$ , we denote its state process by  $(Q_t^{\mathcal{P}})_{t \in \mathbb{N}_0}$ . It is defined to be a Markov chain with  $Q_0^{\mathcal{P}} \equiv q_0$  and*

$$\begin{aligned} & \mathbb{P}(Q_{t+1}^{\mathcal{P}} = q_{t+1} \mid Q_t^{\mathcal{P}} = q_t, \dots, Q_0^{\mathcal{P}} = q_0) \\ &= \mathbb{P}(Q_{t+1}^{\mathcal{P}} = q_{t+1} \mid Q_t^{\mathcal{P}} = q_t) = T(q_t, q_{t+1}) \end{aligned} \quad (1)$$

for all  $q_0, \dots, q_{t+1} \in \mathcal{Q}$ . Further, we define the emission process  $(E_t^{\mathcal{P}})_{t \in \mathbb{N}_0}$  through

$$\begin{aligned} & \mathbb{P}(E_t^{\mathcal{P}} = e \mid Q_0^{\mathcal{P}} = q_0, \dots, Q_t^{\mathcal{P}} = q_t, E_0^{\mathcal{P}} = e_0, \dots, E_{t-1}^{\mathcal{P}} = e_{t-1}) \\ &= \mathbb{P}(E_t^{\mathcal{P}} = e \mid Q_t^{\mathcal{P}} = q) = \mu_q(e), \end{aligned} \quad (2)$$

i.e. the current emission depends solely on the current state. We use  $(Q_t^{\mathcal{P}})_{t \in \mathbb{N}_0}$  and  $(E_t^{\mathcal{P}})_{t \in \mathbb{N}_0}$  to define the process of values  $(V_t^{\mathcal{P}})_{t \in \mathbb{N}_0}$  resulting from the performed operations:

$$V_0^{\mathcal{P}} \equiv v_0 \quad \text{and} \quad V_t^{\mathcal{P}} = \theta_{Q_t^{\mathcal{P}}}(V_{t-1}^{\mathcal{P}}, E_t^{\mathcal{P}}). \quad (3)$$

If the considered PAA is clear from the context, we omit the superscript  $\mathcal{P}$  and write  $(Q_t)_{t \in \mathbb{N}_0}$ ,  $(V_t)_{t \in \mathbb{N}_0}$ , and  $(E_t)_{t \in \mathbb{N}_0}$ , respectively.

### 3 Computing the State-Value Distributions of PAAs

We describe two algorithms to compute the distribution of resulting values. In other words, we seek to calculate the distribution  $\mathcal{L}(V_n)$  of the random variable  $V_n$  for a given  $n$ . The idea is to compute the joint distribution  $\mathcal{L}(Q_n, V_n)$  and then to derive the sought distribution by marginalization:

$$\mathbb{P}(V_n = v) = \sum_{q \in \mathcal{Q}} \mathbb{P}(Q_n = q, V_n = v). \quad (4)$$

For the sake of a shorter notation, we define  $f_t(q, v) := \mathbb{P}(Q_t = q, V_t = v)$  for  $t \in \mathbb{N}_0$ ,  $q \in \mathcal{Q}$ ,  $v \in \mathcal{V}$ .

A slight complication arises when  $\mathcal{V}$  is infinite. However, for each  $t$ , the range of  $V_t$  is finite, as it is a function of the states and emissions up to time  $t$ , and these are finite sets. We define  $\mathcal{V}_t := \text{range } V_t$  and  $\vartheta_n := \max_{0 \leq t \leq n} |\mathcal{V}_t|$ . Clearly  $\vartheta_n \leq (|\mathcal{Q}| \cdot |\mathcal{E}|)^n$ . Therefore all actual computations are on finite sets. As we will see, in many applications,  $\vartheta_n$  grows only polynomially (even linearly) with  $n$ . In the following, we shall understand  $\mathcal{V}$  as the appropriate union of  $\mathcal{V}_t$  sets. Running times of algorithms are given in terms of  $\vartheta_n$ .

### 3.1 Basic Algorithm

We discuss an algorithm to compute the distribution  $f_n = \mathcal{L}(Q_n, V_n)$ . A basic recurrence relation follows from Definitions 2.1 and 2.3.

**Lemma 3.1** (State-value recurrence). *For a given PAA, the state-value distribution can be computed by*

$$f_0(q, v) = \begin{cases} 1 & \text{if } q = q_0 \text{ and } v = v_0, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

and

$$f_{t+1}(q, v) = \sum_{q' \in \mathcal{Q}} \sum_{(v', e) \in \theta_q^{-1}(v)} f_t(q', v') \cdot T(q', q) \cdot \mu_q(e), \quad (6)$$

where  $\theta_q^{-1}(v)$  denotes the inverse image set of  $v$  under  $\theta_q$ .

*Proof.* Equation (5) follows directly from (1) and (3). Let us verify Equation (6):

$$\begin{aligned} f_{t+1}(q, v) &= \mathbb{P}(Q_{t+1} = q, V_{t+1} = v) \\ &= \sum_{q' \in \mathcal{Q}} \sum_{v' \in \mathcal{V}} \sum_{e \in \mathcal{E}} \mathbb{P}(Q_{t+1} = q, V_{t+1} = v, Q_t = q', V_t = v', E_{t+1} = e) \\ &= \sum_{q' \in \mathcal{Q}} \sum_{v' \in \mathcal{V}} \sum_{e \in \mathcal{E}} \mathbb{P}(Q_{t+1} = q, V_{t+1} = v, E_{t+1} = e \mid Q_t = q', V_t = v') \cdot f_t(q', v') \\ &= \sum_{q' \in \mathcal{Q}} \sum_{v' \in \mathcal{V}} \sum_{e \in \mathcal{E}} [\theta_q(v', e) = v] \cdot \mathbb{P}(Q_{t+1} = q, E_{t+1} = e \mid Q_t = q', V_t = v') \cdot f_t(q', v') \\ &= \sum_{q' \in \mathcal{Q}} \sum_{(v', e) \in \theta_q^{-1}(v)} \underbrace{\mathbb{P}(Q_{t+1} = q, E_{t+1} = e \mid Q_t = q', V_t = v')}_{(*)} \cdot f_t(q', v'). \end{aligned}$$

We further evaluate the expression (\*):

$$\begin{aligned} (*) &= \mathbb{P}(Q_{t+1} = q, E_{t+1} = e \mid Q_t = q', V_t = v') \\ &= \underbrace{\mathbb{P}(E_{t+1} = e \mid Q_t = q', Q_{t+1} = q, V_t = v')}_{\stackrel{(i)}{=} \mu_q(e)} \cdot \underbrace{\mathbb{P}(Q_{t+1} = q \mid Q_t = q', V_t = v')}_{\stackrel{(ii)}{=} T(q', q)}, \end{aligned}$$

where (i) is true because of (2) and (3) and (ii) follows from the fact that  $(Q_t)_{t \in \mathbb{N}_0}$  is a Markov chain.  $\square$

We start with the distribution  $f_0$  and calculate the subsequent distributions by applying Equation (6) until we obtain the desired  $f_n$ . A straightforward implementation of Equation (6) results in a *pull-strategy*; that means each entry in the table representing  $f_{t+1}$  is calculated by “pulling over” the required probabilities from table  $f_t$ . Note that this approach makes it necessary to calculate  $\theta^{-1}$  in a preprocessing step. In order to avoid this, we may implement a *push-strategy*, meaning that we iterate over all entries in  $f_t$  rather than  $f_{t+1}$  and “push” the encountered summands over to the appropriate

---

**Algorithm 1** PAADIST

---

**Input:**  $f_0 = \mathcal{L}(Q_0, V_0)$ ,  $n \in \mathbb{N}_0$ , space for two tables of size  $|\mathcal{Q}| \times \vartheta_n$

**Output:**  $f_n = \mathcal{L}(Q_n, V_n)$

```
1: for  $t = 1$  to  $n$  do
2:   initialize  $f_t(q, v) \equiv 0$  for all  $q \in \mathcal{Q}$ ,  $v \in \mathcal{V}_t$ 
3:   for all  $q \in \mathcal{Q}$  and  $v \in \mathcal{V}_{t-1}$  do
4:     for all  $q' \in \mathcal{Q}$  and  $e \in \mathcal{E}$  do
5:        $v' \leftarrow \theta_{q'}(v, e)$ 
6:        $f_t(q', v') \leftarrow f_t(q', v') + f_{t-1}(q, v) \cdot T(q, q') \cdot \mu_{q'}(e)$ 
7:     end for
8:   end for
9: end for
10: return  $f_n$ 
```

---

places in table  $f_{t+1}$ ; in effect, we just change the order of summation. Algorithm 1 shows the push-strategy in detail.

In the course of the computation, we have to store two distributions,  $f_t$  and  $f_{t+1}$ , at a time. Once  $f_{t+1}$  is calculated,  $f_t$  can be discarded. Since the table at time  $t$  has a size of  $|\mathcal{Q}| \times |\mathcal{V}_t|$ , the total space consumption is  $\mathcal{O}(|\mathcal{Q}| \cdot \vartheta_n)$ . Computing  $f_t$  from  $f_{t-1}$  takes  $\mathcal{O}(|\mathcal{Q}| \cdot |\mathcal{V}_t| + |\mathcal{Q}|^2 \cdot |\mathcal{V}_{t-1}| \cdot |\mathcal{E}|)$  time, as can be seen from Algorithm 1. We arrive at the following lemma.

**Lemma 3.2.** *Given a PAA  $(\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu, \theta)$ , the distribution of values  $\mathcal{L}(V_n)$  can be computed in  $\mathcal{O}(n \cdot |\mathcal{Q}|^2 \cdot \vartheta_n \cdot |\mathcal{E}|)$  time and  $\mathcal{O}(|\mathcal{Q}| \cdot \vartheta_n)$  space.*

## 3.2 Doubling Technique

If  $n$  is large, executing the above algorithm may be slow. In this section, we present an alternative algorithm that may be favorable for large  $n$ . To derive this algorithm, we consider the conditional probability

$$U^{(t)}(q_1, q_2, v_1, v_2) := \mathbb{P}(Q_{t_0+t} = q_2, V_{t_0+t} = v_2 \mid Q_{t_0} = q_1, V_{t_0} = v_1). \quad (7)$$

Note that  $U^{(t)}$  does not depend on  $t_0$ , because transition as well as emission probabilities do not change over “time” (a property called *homogeneity*). Once  $U^{(n)}$  is known, we can simply read off the desired distribution  $\mathcal{L}(Q_n, V_n)$ :

$$\mathbb{P}(Q_n = q, V_n = v) = U^{(n)}(q_0, q, v_0, v). \quad (8)$$

The following lemma shows how  $U^{(t)}$  can be computed.

**Lemma 3.3.** *Let  $(\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu, \theta)$  be a PAA and  $(Q_t)_{t \in \mathbb{N}_0}$  and  $(V_t)_{t \in \mathbb{N}_0}$  its state and value process, respectively. Then,*

$$U^{(1)}(q_1, q_2, v_1, v_2) = T(q_1, q_2) \cdot \sum_{\substack{e \in \mathcal{E}: \\ \theta_{q_2}(v_1, e) = v_2}} \mu_{q_2}(e) \quad (9)$$



and, for all  $t_1 \in \mathbb{N}_0$  and  $t_2 \in \mathbb{N}_0$ ,

$$U^{(t_1+t_2)}(q_1, q_2, v_1, v_2) = \sum_{q' \in \mathcal{Q}} \sum_{v' \in \mathcal{V}} U^{(t_1)}(q_1, q', v_1, v') \cdot U^{(t_2)}(q', q_2, v', v_2). \quad (10)$$

Using these recurrences, the distribution of values  $\mathcal{L}(V_n)$  can be computed in  $\mathcal{O}(\log n \cdot |\mathcal{Q}|^3 \cdot \vartheta_n^3)$  time and  $\mathcal{O}(|\mathcal{Q}|^2 \cdot \vartheta_n^2)$  space.

*Proof.* Equation (9) follows from Definition 2.3, while Equation (10) follows from the Chapman-Kolmogorov Equation for homogeneous Markov chains when the PAA is seen as a Markov chain with state space  $\mathcal{Q} \times \mathcal{V}$ . Computing  $U^{(t_1+t_2)}$  from  $U^{(t_1)}$  and  $U^{(t_2)}$  takes  $\mathcal{O}(|\mathcal{Q}|^3 \cdot \vartheta_n^3)$  time, as follows from Equation (10). On the other hand, one step suffices to obtain  $U^{(2t)}$  from  $U^{(t)}$ . Thus, we can compute all  $U^{(2^b)}$  for  $0 \leq b \leq \lceil \log(n) \rceil$  in  $\lceil \log(n) \rceil$  steps, which in turn can be combined into  $U^{(n)}$  in at most  $\lceil \log(n) \rceil$  steps.  $\square$

We note that the doubling technique is asymptotically faster if  $\vartheta_n$  is  $o(\sqrt{n/\log n})$  and  $\mathcal{Q}$  and  $\mathcal{E}$  are fixed.

## 4 Waiting Times

Besides calculating the distribution of values after a fixed number of steps, we can ask for the distribution of the number of steps needed to reach a certain value or a certain state. Such *waiting time* problems play an important role in many applications. We discuss examples in sections 6 and 10. A classical treatment of waiting time problems is given in [21]. Applications to occurrence problems in texts are reviewed in [59].

**Definition 4.1** (Waiting time for a value). *The waiting time for a set of target values  $\mathcal{T} \subset \mathcal{V}$  is a random variable defined as  $W_{\mathcal{T}} := \min\{t \in \mathbb{N}_0 \mid V_t \in \mathcal{T}\}$  if this set is not empty, and defined as infinity otherwise.*

While  $\mathbb{P}(W_{\mathcal{T}} \geq t)$  may be nonzero for all  $t \in \mathbb{N}$ , we are frequently only interested in the distribution up to a fixed time  $n$ . Then, of course, the exact values of  $\mathcal{L}(W_{\mathcal{T}})(t) = \mathbb{P}(W_{\mathcal{T}} = t)$  are unknown for  $t > n$ , but their total probability  $\mathbb{P}(W_{\mathcal{T}} > n)$  is known, and  $n$  is typically chosen such that this total probability remains below a desired threshold.

**Lemma 4.2.** *Let  $(\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu, \theta)$  be a PAA and  $\mathcal{T} \subset \mathcal{V}$ . Then, the probabilities  $\mathcal{L}(W_{\mathcal{T}})(0), \dots, \mathcal{L}(W_{\mathcal{T}})(n)$  can be computed in  $\mathcal{O}(n \cdot |\mathcal{Q}|^2 \cdot |\mathcal{E}| \cdot (\vartheta_n - |\mathcal{T}|))$  time and  $\mathcal{O}(|\mathcal{Q}| \cdot (\vartheta_n - |\mathcal{T}|))$  space. Alternatively, this can be done using  $\mathcal{O}(\log n \cdot |\mathcal{Q}|^3 \cdot (\vartheta_n - |\mathcal{T}|)^3)$  time and  $\mathcal{O}(|\mathcal{Q}|^2 \cdot (\vartheta_n - |\mathcal{T}|)^2)$  space.*

*Proof.* We construct a modified PAA by defining a new value set  $\mathcal{V}' := (\mathcal{V} \setminus \mathcal{T}) \cup \{\bullet, \circ\}$ , assuming (without loss of generality) that  $\bullet, \circ \notin \mathcal{V}$ , and new operations

$$\theta'_q(v, e) := \begin{cases} \theta_q(v, e) & \text{if } v \notin \{\bullet, \circ\} \text{ and } \theta_q(v, e) \notin \mathcal{T}, \\ \bullet & \text{if } v \notin \{\bullet, \circ\} \text{ and } \theta_q(v, e) \in \mathcal{T}, \\ \circ & \text{if } v \in \{\bullet, \circ\} \end{cases}$$

for all  $q \in \mathcal{Q}$ . Let  $V'_t$  be the modified value process. Using the modified PAA, the probability of waiting time  $t$  can be expressed as

$$\mathbb{P}(W_{\mathcal{T}} = t) = \mathbb{P}(V'_t = \bullet).$$

Runtime and space bounds follow from Lemmas 3.2 and 3.3.  $\square$

Besides waiting for a set of values, we may also wait for a set of states. Since a PAA's state process does not depend on emission and value processes, the remainder of this section solely concerns the Markov chain  $(\mathcal{Q}, T, \delta_{q_0})$ , which is part of the PAA  $(\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu, \theta)$ . Waiting times in Markov chains are a well-studied topic with particular interest in pattern occurrences [59] and queuing theory [13]. For completeness and implementation purposes within the PAA framework, we briefly restate the construction here.

**Definition 4.3** (Waiting time for a state). *The waiting time for a set of target states  $\mathcal{S} \subset \mathcal{Q}$  is a random variable defined as  $W_{\mathcal{S}} := \min\{t \in \mathbb{N}_0 \mid Q_t \in \mathcal{S}\}$  if this set is not empty and defined as infinity otherwise.*

**Lemma 4.4.** *Let  $(\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu, \theta)$  be a PAA,  $\alpha : \mathcal{Q} \rightarrow [0, 1]$  be a probability distribution on  $\mathcal{Q}$ , and  $\mathcal{S} \subset \mathcal{Q}$  be a set of target states. Consider the Markov chain  $(\mathcal{Q}, T, \alpha)$ , let  $(Q'_t)_{t \in \mathbb{N}_0}$  be its state process, and let  $W'_{\mathcal{S}} := \min\{t \in \mathbb{N}_0 \mid Q'_t \in \mathcal{S}\}$  be the waiting time for states  $\mathcal{S}$ . Then  $\mathcal{L}(W'_{\mathcal{S}})(0), \dots, \mathcal{L}(W'_{\mathcal{S}})(n)$  can be computed in  $\mathcal{O}(n \cdot |\mathcal{Q}|^2)$  time and  $\mathcal{O}(|\mathcal{Q}|)$  space, or in  $\mathcal{O}(\log n \cdot |\mathcal{Q}|^3)$  time and  $\mathcal{O}(|\mathcal{Q}|^2)$  space using the doubling technique. If  $\alpha = \delta_{q_0}$ , then  $W_{\mathcal{S}} = W'_{\mathcal{S}}$ .*

*Proof.* As in the proof of Lemma 4.2, we introduce an aggregation state  $\bullet$  to replace  $\mathcal{S}$  and an absorbing state  $\circ$  to “flush”  $\bullet$ . Then  $\mathbb{P}(W'_{\mathcal{S}} = t) = \mathbb{P}(Q'_t = \bullet)$ .  $\square$

When  $\alpha = \delta_{q_0}$ , the above lemma yields the waiting time for the first event of reaching one of the states in  $\mathcal{S}$ . We further consider the waiting time of a return event

$$W_{\mathcal{S}}^{t_0} := \min\{t \in \mathbb{N} \mid Q_{t_0+t} \in \mathcal{S}\}.$$

If the Markov chain is aperiodic and irreducible, it has a unique stationary state distribution, against which the PAA state distribution converges exponentially fast. We can then use Lemma 4.4 to compute

$$\lim_{t \rightarrow \infty} \mathbb{P}(W_{\mathcal{S}}^t = t' \mid Q_t \in \mathcal{S}) \quad \text{for each } t' \in \mathbb{N}$$

by choosing  $\alpha$  in Lemma 4.4 as the stationary distribution restricted to  $\mathcal{S}$ .

## 5 PAAs Based on Random Sequences

We discuss the construction of PAAs modelling the deterministic processing of random sequences. That means we assume to be given a mechanism that processes sequences

character by character and deterministically computes a value for a given string. A pattern matching algorithm that computes the number of matches in a given sequence might serve as an example. In this section, we ask for the distribution of resulting values when a *deterministic* computation is applied to *random* strings. Therefore we first define *text models* and *deterministic arithmetic automata* to represent random texts and deterministic computations, respectively, and then combine both into a PAA.

## 5.1 Random Text Models

Given an alphabet  $\Sigma$ , a random text is a stochastic process  $(S_t)_{t \in \mathbb{N}_0}$ , where each  $S_t$  takes values in  $\Sigma$ . A text model  $\mathbb{P}$  is a probability measure assigning probabilities to (sets of) strings. It is given by (consistently) specifying the probabilities  $\mathbb{P}(S_0 \dots S_{|s|-1} = s)$  for all  $s \in \Sigma^*$ . We only consider finite-memory models in this article which are formalized in the following definition.

**Definition 5.1** (Finite-memory text model). *A finite-memory text model is a tuple  $(\mathcal{C}, c_0, \Sigma, \varphi)$ , where  $\mathcal{C}$  is a finite state space (called context space),  $c_0 \in \mathcal{C}$  a start context,  $\Sigma$  an alphabet, and  $\varphi : \mathcal{C} \times \Sigma \times \mathcal{C} \rightarrow [0, 1]$  with  $\sum_{\sigma \in \Sigma, c' \in \mathcal{C}} \varphi(c, \sigma, c') = 1$  for all  $c \in \mathcal{C}$ . The random variable giving the text model state after  $t$  steps is denoted  $C_t$  with  $C_0 \equiv c_0$ . A probability measure is now induced by stipulating*

$$\mathbb{P}(S_0 \dots S_{n-1} = s, C_1 = c_1, \dots, C_n = c_n) := \prod_{i=0}^{n-1} \varphi(c_i, s[i], c_{i+1})$$

for all  $n \in \mathbb{N}_0$ ,  $s \in \Sigma^n$ , and  $(c_1, \dots, c_n) \in \mathcal{C}^n$ .

The idea is that the model given by  $(\mathcal{C}, c_0, \Sigma, \varphi)$  generates a random text by moving from context to context and emitting a character at each transition, where  $\varphi(c, \sigma, c')$  is the probability of moving from context  $c$  to context  $c'$  and thereby generating the letter  $\sigma$ .

Note that the probability  $\mathbb{P}(S_0 \dots S_{|s|-1} = s)$  is obtained by marginalization over all context sequences that generate  $s$ . This can be efficiently done, using the decomposition of the following lemma.

**Lemma 5.2.** *Let  $(\mathcal{C}, c_0, \Sigma, \varphi)$  be a finite-memory text model. Then,*

$$\mathbb{P}(S_0 \dots S_n = s\sigma, C_{n+1} = c) = \sum_{c' \in \mathcal{C}} \mathbb{P}(S_0 \dots S_{n-1} = s, C_n = c') \cdot \varphi(c', \sigma, c)$$

for all  $n \in \mathbb{N}_0$ ,  $s \in \Sigma^n$ ,  $\sigma \in \Sigma$  and  $c \in \mathcal{C}$ .

*Proof.* We have

$$\begin{aligned}
& \mathbb{P}(S_0 \dots S_n = s\sigma, C_{n+1} = c) \\
&= \sum_{c_1, \dots, c_n} \mathbb{P}(S_0 \dots S_n = s\sigma, C_1 = c_1, \dots, C_n = c_n, C_{n+1} = c) \\
&= \sum_{c_1, \dots, c_n} \prod_{i=0}^{n-1} \varphi(c_i, s[i], c_{i+1}) \cdot \varphi(c_n, \sigma, c) \\
&= \sum_{c_n \in \mathcal{C}} \left( \sum_{c_1, \dots, c_{n-1}} \prod_{i=0}^{n-1} \varphi(c_i, s[i], c_{i+1}) \right) \cdot \varphi(c_n, \sigma, c) \\
&= \sum_{c_n \in \mathcal{C}} \mathbb{P}(S_0 \dots S_{n-1} = s, C_n = c_n) \cdot \varphi(c_n, \sigma, c).
\end{aligned}$$

Renaming  $c_n$  to  $c'$  yields the claimed result.  $\square$

Similar text models are used in [36], where they are called probability transducers. In the following, we refer to a finite-memory text model  $(\mathcal{C}, c_0, \Sigma, \varphi)$  simply as text model, as all text models considered in this article are special cases of Definition 5.1.

For an i.i.d. model, we set  $\mathcal{C} = \{\varepsilon\}$  and  $\varphi(\varepsilon, \sigma, \varepsilon) = p_\sigma$  for each  $\sigma \in \Sigma$ , where  $p_\sigma$  is the occurrence probability of letter  $\sigma$  (and  $\varepsilon$  may be interpreted as an empty context). For a Markovian text model of order  $r$ , the distribution of the next character depends only on the  $r$  preceding characters (fewer at the beginning); thus we set  $\mathcal{C} := \bigcup_{i=0}^r \Sigma^i$ . The conditional follow-up probabilities are given by

$$\begin{aligned}
& \mathbb{P}(S_i = s[i] \mid S_{i-1} = s[i-1], \dots, S_0 = s[0]) \\
&= \begin{cases} \varphi(s[.i-1], s[i], s[.i]) & \text{if } i < r, \\ \varphi(s[i-r \dots i-1], s[i], s[i-r+1 \dots i]) & \text{if } i \geq r. \end{cases}
\end{aligned}$$

This notion of text models also covers variable order Markov chains as introduced in [62], which can be converted into equivalent models of fixed order. Text models as defined above have the same expressive power as character-emitting HMMs, that means, they allow to construct the same probability distributions. For a given HMM, we can construct an equivalent text model by using the same state space (contexts) and setting  $\varphi(c, \sigma, c') := T(c, c') \cdot \mu_{c'}(\sigma)$ , where  $T$  and  $\mu_{c'}$  are the HMM's transition function and emission distribution attached to state  $c'$ , respectively. When, on the other hand, a text model  $(\mathcal{C}, c_0, \Sigma, \varphi)$  is given, we construct an equivalent HMM by using  $\mathcal{C}^2$  as state space and setting

$$T((c_1, c_2), (c'_1, c'_2)) := \begin{cases} \sum_{\sigma \in \Sigma} \varphi(c_2, \sigma, c'_2) & \text{if } c_2 = c'_1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mu_{(c_1, c_2)}(\sigma) := \varphi(c_1, \sigma, c_2).$$

## 5.2 Deterministic Arithmetic Automata (DAAs)

In order to model deterministic calculations on sequences, we define a deterministic counter-part to PAAs.

**Definition 5.3** (Deterministic Arithmetic Automaton, DAA). *A deterministic arithmetic automaton is a tuple*

$$\mathcal{D} = (\mathcal{Q}, q_0, \Sigma, \delta, \mathcal{V}, v_0, \mathcal{E}, (\eta_q)_{q \in \mathcal{Q}}, (\theta_q)_{q \in \mathcal{Q}}),$$

where  $\mathcal{Q}$  is a finite set of states,  $q_0 \in \mathcal{Q}$  is the start state,  $\Sigma$  is a finite alphabet,  $\delta : \mathcal{Q} \times \Sigma \rightarrow \mathcal{Q}$  is called transition function,  $\mathcal{V}$  is a set of values,  $v_0 \in \mathcal{V}$  is called the start value,  $\mathcal{E}$  is a finite set of emissions,  $\eta_q \in \mathcal{E}$  is the emission associated to state  $q$ , and  $\theta_q : \mathcal{V} \times \mathcal{E} \rightarrow \mathcal{V}$  is a binary operation associated to state  $q$ . Further, we define the associated joint transition function

$$\bar{\delta} : (\mathcal{Q} \times \mathcal{V}) \times \Sigma \rightarrow (\mathcal{Q} \times \mathcal{V}), \quad \bar{\delta}((q, v), \sigma) := (\delta(q, \sigma), \theta_{\delta(q, \sigma)}(v, \eta_{\delta(q, \sigma)})).$$

We extend the definitions of  $\delta$  and  $\bar{\delta}$  inductively to  $\Sigma^*$  in their second argument by setting

$$\begin{aligned} \delta(q, \varepsilon) &:= q && \text{for the empty string } \varepsilon, \\ \delta(q, x\sigma) &:= \delta(\delta(q, x), \sigma) && \text{for all } x \in \Sigma^* \text{ and } \sigma \in \Sigma, \\ \bar{\delta}((q, v), \varepsilon) &:= (q, v), \\ \bar{\delta}((q, v), x\sigma) &:= \bar{\delta}(\bar{\delta}((q, v), x), \sigma). \end{aligned}$$

When  $\bar{\delta}((q_0, v_0), s) = (q, v)$  for some  $q \in \mathcal{Q}$  and  $s \in \Sigma^*$ , we say that  $\mathcal{D}$  computes value  $v$  for input  $s$  and define  $\text{value}_{\mathcal{D}}(s) := v$ .

Informally, a DAA starts with the state-value pair  $(q_0, v_0)$  and reads a sequence of symbols from  $\Sigma$ . Being in state  $q$  with value  $v$ , upon reading  $\sigma \in \Sigma$ , the DAA performs a state transition to  $q' := \delta(q, \sigma)$  and updates the value to  $v' := \theta_{q'}(v, \eta_{q'})$  using the operation and emission of the new state  $q'$ .

For each state  $q$ , the emission  $\eta_q$  is fixed and could be dropped from the definition of DAAs. In fact, one could also dispense with values and operations entirely and define a DFA over state space  $\mathcal{Q} \times \mathcal{V}$ , performing the same operations as a DAA. However, we intentionally include values, operations, and emissions to emphasize the connection to PAAs.

## 5.3 Constructing PAAs from DAAs and Text Models

We now formally state how to convert a DAA into a (restricted) PAA, where each emission distribution is deterministic (assigning probability 1 to a particular value), given a text model.

**Lemma 5.4** (DAA + Text model  $\rightarrow$  PAA). *Let  $(\mathcal{C}, c_0, \Sigma, \varphi)$  be a text model and  $\mathcal{D} = (\mathcal{Q}^{\mathcal{D}}, q_0^{\mathcal{D}}, \Sigma, \delta, \mathcal{V}, v_0, \mathcal{E}, (\eta_q)_{q \in \mathcal{Q}^{\mathcal{D}}}, (\theta_q^{\mathcal{D}})_{q \in \mathcal{Q}^{\mathcal{D}}})$  be a DAA. Then, define*

- a state space  $\mathcal{Q} := \mathcal{Q}^{\mathcal{D}} \times \mathcal{C}$ ,
- a start state  $q_0 := (q_0^{\mathcal{D}}, c_0)$ ,
- transition probabilities

$$T((q, c), (q', c')) := \sum_{\sigma \in \Sigma: \delta(q, \sigma) = q'} \varphi(c, \sigma, c'), \quad (11)$$

- (deterministic) emission probability vectors

$$\mu_{(q, c)}(e) := \begin{cases} 1 & \text{if } e = \eta_q, \\ 0 & \text{otherwise,} \end{cases}$$

for all  $(q, c) \in \mathcal{Q}$ .

- operations  $\theta_{(q, c)}(v, e) := \theta_q^{\mathcal{D}}(v, e)$  for all  $(q, c) \in \mathcal{Q}$ .

Then,  $\mathcal{P} = (\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, \mu = (\mu_q)_{q \in \mathcal{Q}}, \theta = (\theta_q)_{q \in \mathcal{Q}})$  is a PAA with

$$\mathcal{L}(V_t) = \mathcal{L}(\text{value}_{\mathcal{D}}(S_0 \dots S_{t-1}))$$

for all  $t \in \mathbb{N}_0$ , where  $S$  is a random text according to the text model  $(\mathcal{C}, c_0, \Sigma, \varphi)$ .

*Proof.*  $\mathcal{P}$  is a PAA by Definition 2.1. As in Section 3, we define  $f_t(q, v) := \mathbb{P}(Q_t = q, V_t = v)$ . To prove  $\mathcal{L}(V_t) = \mathcal{L}(\text{value}_{\mathcal{D}}(S_0 \dots S_{t-1}))$ , we show that

$$f_t((q^{\mathcal{D}}, c), v) = \sum_{s \in \Sigma^t} [\bar{\delta}((q_0^{\mathcal{D}}, v_0), s) = (q^{\mathcal{D}}, v)] \cdot \mathbb{P}(S_0 \dots S_{t-1} = s, C_t = c) \quad (12)$$

for all  $q^{\mathcal{D}} \in \mathcal{Q}^{\mathcal{D}}$ ,  $c \in \mathcal{C}$ ,  $v \in \mathcal{V}$ , and  $t \in \mathbb{N}_0$ . For  $t = 0$ , Equation (12) is correct by definitions of PAAs, DAAs and text models. For  $t > 0$  we prove it inductively. Assume (12) to be correct for all  $t'$  with  $0 \leq t' < t$ .

$$f_t(\underbrace{(q^{\mathcal{D}}, c)}_{=: q}, v) \quad (13)$$

$$= \sum_{q' \in \mathcal{Q}} \sum_{(v', e) \in \theta_q^{-1}(v)} f_{t-1}(q', v') \cdot T(q', q) \cdot \mu_q(e) \quad (14)$$

$$= \sum_{q' \in \mathcal{Q}} \sum_{(v', e) \in \mathcal{V} \times \mathcal{E}} [\theta_{q^{\mathcal{D}}}^{\mathcal{D}}(v', e) = v] \cdot f_{t-1}(q', v') \cdot T(q', q) \cdot [\eta_{q^{\mathcal{D}}} = e] \quad (15)$$

$$\begin{aligned} &= \sum_{q'^{\mathcal{D}} \in \mathcal{Q}^{\mathcal{D}}} \sum_{c' \in \mathcal{C}} \sum_{(v', e) \in \mathcal{V} \times \mathcal{E}} [\theta_{q'^{\mathcal{D}}}^{\mathcal{D}}(v', e) = v] \cdot [\eta_{q'^{\mathcal{D}}} = e] \cdot f_{t-1}(q', v') \\ &\quad \cdot \sum_{\sigma \in \Sigma} [\delta(q'^{\mathcal{D}}, \sigma) = q^{\mathcal{D}}] \cdot \varphi(c', \sigma, c) \end{aligned} \quad (16)$$

$$\begin{aligned}
&= \sum_{s \in \Sigma^{t-1}} \sum_{\sigma \in \Sigma} \sum_{q^{\mathcal{D}} \in \mathcal{Q}^{\mathcal{D}}} \sum_{c' \in \mathcal{C}} \sum_{(v', e) \in \mathcal{V} \times \mathcal{E}} \llbracket \theta_{q^{\mathcal{D}}}^{\mathcal{D}}(v', e) = v \rrbracket \cdot \llbracket \eta_{q^{\mathcal{D}}} = e \rrbracket \\
&\quad \cdot \llbracket \delta(q^{\mathcal{D}}, \sigma) = q^{\mathcal{D}} \rrbracket \cdot \llbracket \bar{\delta}((q_0^{\mathcal{D}}, v_0), s) = (q^{\mathcal{D}}, v') \rrbracket \\
&\quad \cdot \mathbb{P}(S_0 \dots S_{t-2} = s, C^{t-1} = c') \cdot \varphi(c', \sigma, c)
\end{aligned} \tag{17}$$

$$\begin{aligned}
&= \sum_{s\sigma \in \Sigma^t} \sum_{q^{\mathcal{D}} \in \mathcal{Q}^{\mathcal{D}}} \sum_{(v', e) \in \mathcal{V} \times \mathcal{E}} \llbracket \theta_{q^{\mathcal{D}}}^{\mathcal{D}}(v', e) = v \rrbracket \cdot \llbracket \eta_{q^{\mathcal{D}}} = e \rrbracket \cdot \llbracket \bar{\delta}((q_0^{\mathcal{D}}, v_0), s) = (q^{\mathcal{D}}, v') \rrbracket \\
&\quad \cdot \llbracket \delta(q^{\mathcal{D}}, \sigma) = q^{\mathcal{D}} \rrbracket \cdot \mathbb{P}(S_0 \dots S_{t-1} = s\sigma, C_t = c)
\end{aligned} \tag{18}$$

$$= \sum_{s\sigma \in \Sigma^t} \llbracket \bar{\delta}((q_0^{\mathcal{D}}, v_0), s\sigma) = (q^{\mathcal{D}}, v) \rrbracket \cdot \mathbb{P}(S_0 \dots S_{t-1} = s\sigma, C_t = c) \tag{19}$$

In the above derivation, step (13)→(14) follows from (6). Step (14)→(15) follows from the definitions of  $\theta_q$  and  $\mu_q$ . Step (15)→(16) uses the definitions of  $T$  and  $\mathcal{Q}$  in Lemma 5.4. Step (16)→(17) uses the induction assumption. Step (17)→(18) uses Lemma 5.2. The final step (18)→(19) follows by combining the four Iverson brackets summed over  $q^{\mathcal{D}}$  and  $(v', e)$  into a single Iverson bracket.  $\square$

**Remark 5.5.** *In the above lemma, states having zero probability of being reached from  $q_0$  may be omitted from  $\mathcal{Q}$  and  $T$ .*

**Lemma 5.6** (PAA from DAA; Construction time and space).

1. *For a PAA constructed according to Lemma 5.4, the value distribution  $\mathcal{L}(V_n)$ , or the joint state-value distribution, can be computed with  $\mathcal{O}(n \cdot |\mathcal{Q}^{\mathcal{D}}| \cdot |\Sigma| \cdot |\mathcal{C}|^2 \cdot \vartheta_n)$  operations using  $\mathcal{O}(|\mathcal{Q}^{\mathcal{D}}| \cdot |\mathcal{C}| \cdot \vartheta_n)$  space. The same statement holds for computing the waiting time distribution up to time  $n$ .*
2. *If for all  $c \in \mathcal{C}$  and  $\sigma \in \Sigma$ , there exists at most one  $c' \in \mathcal{C}$  such that  $\varphi(c, \sigma, c') > 0$ , then the time is bounded by  $\mathcal{O}(n \cdot |\mathcal{Q}^{\mathcal{D}}| \cdot |\Sigma| \cdot |\mathcal{C}| \cdot \vartheta_n)$ .*
3. *Using the doubling technique, the distributions can be computed in  $\mathcal{O}(\log n \cdot |\mathcal{Q}^{\mathcal{D}}|^3 \cdot |\mathcal{C}|^3 \cdot \vartheta_n^3)$  time and  $\mathcal{O}(|\mathcal{Q}^{\mathcal{D}}|^2 \cdot |\mathcal{C}|^2 \cdot \vartheta_n^2)$  space.*

*Proof.*

1. From Lemma 3.2, we obtain bounds for time and space complexity of  $\mathcal{O}(n \cdot |\mathcal{Q}|^2 \cdot \vartheta_n \cdot |\mathcal{E}|)$  and  $\mathcal{O}(|\mathcal{Q}| \cdot \vartheta_n)$ , respectively. By construction,  $|\mathcal{Q}| \leq |\mathcal{Q}^{\mathcal{D}}| \cdot |\mathcal{C}|$ . Recall that Lemma 3.2 is based on Algorithm 1. The loops in lines 1 and 3 together account for a factor of  $\mathcal{O}(n \cdot |\mathcal{Q}| \cdot \vartheta_n)$  in the time complexity. A factor of  $\mathcal{O}(|\mathcal{Q}| \cdot |\mathcal{E}|)$  is caused by the inner loop in line 4. However, since the constructed PAA has deterministic (i.e. Dirac distributed) emissions, we do not need to iterate over all  $e \in \mathcal{E}$  and save a factor of  $|\mathcal{E}|$ . Furthermore, we only need to iterate over all states reachable in one step. For each  $q \in \mathcal{Q}$ , there exist at most  $|\Sigma| \cdot |\mathcal{C}|$  such states by construction of the PAA. Therefore, the inner loop in line 4 can be modified to take  $\mathcal{O}(|\Sigma| \cdot |\mathcal{C}|)$  time, yielding the claimed runtime bound.

2. If for all  $c \in \mathcal{C}$  and  $\sigma \in \Sigma$ , there exists at most one  $c' \in \mathcal{C}$  such that  $\varphi(c, \sigma, c') > 0$ , then at most  $|\Sigma|$  different states are reachable from each state  $q \in \mathcal{Q}$ . The claimed runtime follows by the same arguments as above.
3. Alternative time and space complexities for the doubling algorithm follow directly from Lemma 3.3.

□

This concludes the derivation of the PAA framework and construction methods. The following sections are devoted to several applications.

## 6 Pattern Matching Statistics

One application of the introduced framework are pattern matching statistics. An algorithm that searches for a pattern  $p$  maps strings to the number of matches. That means, it deterministically processes a string and, by doing so, computes a value. In this section, we ask for the distribution of the number of occurrences of a given pattern in a random text.

In computational biology, one searches for patterns that occur often and hypothesizes that these patterns carry biological meaning. Mere abundance, however, does not necessarily imply that the found pattern is meaningful. Short patterns like **AC** and degenerate patterns like **ANNNNNNC** will naturally occur quite often in a given stretch of DNA (here **N** is a wildcard character meaning aNy nucleotide). A better approach to quantify a pattern's overrepresentation is to consult the statistical significance: Suppose we have found a certain pattern  $k$  times in a given sequence. What is the probability of observing  $k$  or more matches just by chance? Precisely this question can be answered by computing the distribution of occurrence counts. Given a suitable null model, a procedure to compute the significance of a pattern is a powerful tool in the context of motif discovery, as it allows the comparison of different patterns regardless of their structure and length.

There are many different types of patterns that are relevant in computational biology, such as single strings, sets of strings, Prosite patterns<sup>1</sup>, consensus strings together with a distance measure and a distance threshold, abelian patterns, position weight matrices in connection with a threshold, etc. All these pattern types may be seen as ways to concisely describe a finite sets of strings. Thus, all these patterns may be expressed in the form of deterministic finite automata (DFAs) that recognize the respective string set. As our method is based on this DFA representation, it is very general and flexible regarding the pattern type.

Besides specifying a pattern, one has to decide how overlaps are to be handled. We refer to the used strategy as *counting scheme*. The easiest case is to disallow overlaps at all; we call this scheme *non-overlapping count*. Consequently, we define the *overlapping count* to be the number of substrings that match the given pattern. In the case of a

---

<sup>1</sup>used in the Prosite database (see Hulo et al. [26]). A syntax description can be found under <http://www.expasy.org/tools/scanprosite/scanprosite-doc.html>.



set of strings without any restrictions, this scheme makes counting more complicated as some words may be substrings of others. Many authors avoid the problem—at least partly—by simply counting the positions where at least one pattern ends, which we refer to as *match position count*.

## 6.1 Related Work and New Results

The topic of statistics of words on random texts has been studied extensively. An overview is provided in the book by Lothaire [40]. Chapter 6 (“Statistics on Words with Applications to Biological Sequences”), which is particularly interesting in this context, is based on the overview article by Reinert et al. [59].

In many approaches, a generating function is derived for the sought quantity. Then, typically using symbolic Taylor expansion, the concrete values can be computed. This procedure is, for instance, described by Régnier [58], who gives formulas for mean, variance and higher statistical moments of the exact occurrence count distribution. This approach has the advantage of additionally allowing asymptotic analysis. Her framework is general enough to admit Markovian sources as well as finite sets of patterns to be treated in the overlapping as well as in the non-overlapping case. Closely related is the approach of Nicodème et al. [53], who present an algorithmic chain to compute the distribution of the match position count for regular expressions. Lladser et al. [39] recently reviewed the field. Their main concern is to bring together involved concepts in a consistent and rigorous manner. They make the connection to the classical field of automata theory and pattern matching explicit and therefore speak of *probabilistic pattern matching*. Furthermore, they describe the relation between finite automata and Markov chains in terms of the *Markov chain embedding* technique. Related approaches to compute the exact p-values based on automata are developed in [9, 55]. Another dynamic programming approach was presented in [71]. It is used to compute exact p-values for position weight matrices describing transcription factor binding sites (TFBS).

In this section, we provide a unifying framework for the efficient computation of pattern matching statistics. In contrast to existing approaches, overlaps are handled correctly and arbitrary finite-memory text models (including HMMs) can be used.

## 6.2 Constructing DAAs from DFAs

As usual, we define a deterministic finite automaton (DFA) to be a tuple  $(\mathcal{Q}, \Sigma, \delta, q_0, \mathcal{F})$ , where  $\mathcal{Q}$  is a finite state space,  $\Sigma$  is a finite alphabet,  $\delta : \mathcal{Q} \times \Sigma \rightarrow \mathcal{Q}$  is a transition function,  $q_0$  is a start state, and  $\mathcal{F} \subset \mathcal{Q}$  is a set of accepting states. Again, we extend  $\delta$  to  $\Sigma^*$  in its second argument, as for DAAs in Definition 5.3. Suppose a pattern is given in the form of a DFA, that means, a string is an instance of the pattern if it is accepted by the DFA. We seek to calculate the distribution of the number of occurrences of this pattern.

To use a DFA for pattern matching, that is, to find all instances of a pattern in a (long) text, we construct an automaton that accepts not only all strings matching the pattern, but all strings that have a suffix matching the pattern (see [52]). For a pattern given in

the form of a DFA or non-deterministic finite automaton (NFA), it is always possible to construct a DFA for this task as follows. First, we add a self-transition labeled with the whole alphabet  $\Sigma$  to the automaton's start state to ensure that the start state remains active all the time. The result is an NFA, which can be made deterministic again by employing the classical subset construction (explained for example in [52]). As we see in Sections 6.2.1 to 6.2.2, there are often simpler and more direct ways to build the sought DFA.

When a DFA reads a text, it is in an accepting state whenever (at least) one instance of the pattern ends. The number of these events equals the *match position count* as defined above. To count the number of times a DFA  $(\mathcal{Q}, \Sigma, \delta, q_0, \mathcal{F})$  is in an accepting state, we define emissions

$$\eta_q := \begin{cases} 1 & \text{if } q \in \mathcal{F}, \\ 0 & \text{otherwise.} \end{cases}$$

We call a tuple  $(\mathcal{Q}, \Sigma, \delta, q_0, (\eta_q)_{q \in \mathcal{Q}})$  *counting DFA*.

To compute the distribution of the *overlapping count* instead of the match position count, we have to take into account that more than one match can end at a position in the text. For each state, we therefore need to know how many matches end when it is entered. As we will see, this is never a problem when the pattern represents a finite set of strings<sup>2</sup>. Formally, we define emissions  $\eta = (\eta_q)_{q \in \mathcal{Q}}$ , where  $\eta_q \in \mathbb{N}_0$  gives the number of matches to be counted upon entering state  $q$ .

To obtain the *nonoverlapping count*, the automaton can be modified accordingly: We change the outgoing transitions of each accepting state of the match position count automaton to act as if they originate from the start state  $q_0$ . Therefore, we define a modified transition function  $\delta'$  by  $\delta'(q, \sigma) := \delta(q, \sigma)$  if  $q \notin \mathcal{F}$ , and  $\delta'(q, \sigma) := \delta(q_0, \sigma)$  otherwise.

As described above, all three counting schemes can be realized by counting DFAs when the pattern is a finite set of strings. The main idea now is to construct a DAA from this counting DFA by adding the emissions generated in each state, and turn this into a PAA using Lemma 5.4. For practical computations, it is often sufficient to truncate the values at a constant  $M \in \mathbb{N}$ . (Note that the match position count is always bounded by the length of the processed text, so  $\vartheta_n = \Theta(n)$ .) The proof of the following theorem contains the details of the DFA  $\rightarrow$  DAA  $\rightarrow$  PAA construction.

**Theorem 6.1.** *Let a counting DFA  $D = (\mathcal{Q}, \Sigma, \delta, q_0, (\eta_q)_{q \in \mathcal{Q}})$  and a text model  $(\mathcal{C}, c_0, \Sigma, \varphi)$  be given, and let  $(S_t)_{t \in \mathbb{N}_0}$  be a random text distributed according to that model. Then, the (truncated) distribution of accumulated counts*

$$\mathcal{L} \left( \min \left\{ M, \sum_{i=0}^{n-1} \eta_{\delta(q_0, S_0 \dots S_i)} \right\} \right) \quad (20)$$

---

<sup>2</sup>If the pattern represents an infinite set, such as a regular expression containing a star operator, there may be states in  $\mathcal{F}$  that can be entered when  $i$  or  $j$  matches end for  $i \neq j$ , or for which the number of overlapping matches to be counted might be unbounded. In this case, one is essentially restricted to the match position count. Since we only consider finite sets, this is not an issue here.

can be computed in  $\mathcal{O}(n \cdot |\mathcal{Q}| \cdot |\mathcal{C}|^2 \cdot |\Sigma| \cdot M)$  time and  $\mathcal{O}(|\mathcal{Q}| \cdot |\mathcal{C}| \cdot M)$  space. If for all  $c \in \mathcal{C}$  and  $\sigma \in \Sigma$ , there exists at most one  $c' \in \mathcal{C}$  such that  $\varphi(c, \sigma, c') > 0$ , then the runtime is bounded by  $\mathcal{O}(n \cdot |\mathcal{Q}| \cdot |\mathcal{C}| \cdot |\Sigma| \cdot M)$ . Alternatively, (20) can be computed in  $\mathcal{O}(\log n \cdot |\mathcal{Q}|^3 \cdot |\mathcal{C}|^3 \cdot M^2)$  time and  $\mathcal{O}(|\mathcal{Q}|^2 \cdot |\mathcal{C}|^2 \cdot M)$  space.

*Proof.* Given the counting DFA  $(\mathcal{Q}, \Sigma, \delta, q_0, (\eta_q)_{q \in \mathcal{Q}})$ , we use it to construct the DAA  $\mathcal{D} = (\mathcal{Q}, q_0, \Sigma, \delta, \mathcal{V}, v_0, \mathcal{E}, (\eta_q)_{q \in \mathcal{Q}}, (\theta_q)_{q \in \mathcal{Q}})$ , where  $\mathcal{V} := \{0, \dots, M\}$ ,  $v_0 := 0$ ,  $\mathcal{E} := \{\eta_q : q \in \mathcal{Q}\}$ , and all operations are truncated additions:

$$\theta_q(v, e) := \begin{cases} M & \text{if } v + e \geq M, \\ v + e & \text{otherwise} \end{cases}$$

for all  $q \in \mathcal{Q}$ . For this DAA  $\mathcal{D}$ , we have

$$value_{\mathcal{D}}(s) = \min \left\{ M, \sum_{i=0}^{n-1} \eta_{\delta(q_0, s[1..i])} \right\}$$

for all  $s \in \Sigma^*$ .

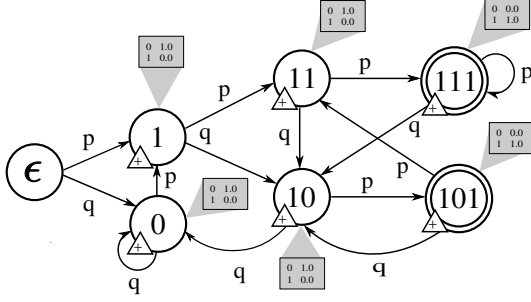
We apply Lemma 5.4 to this DAA and text model in order to construct the PAA. The runtime and space bounds for the basic algorithm follow directly. To obtain the bounds for the alternative doubling algorithm, namely  $\mathcal{O}(\log n \cdot |\mathcal{Q}|^3 \cdot |\mathcal{C}|^3 \cdot M^2)$  time and  $\mathcal{O}(|\mathcal{Q}|^2 \cdot |\mathcal{C}|^2 \cdot M)$  space, we exploit that the operations  $\theta_q$  are (almost) additions in this case. Thus,  $U^{(t)}(q_1, q_2, v_1, v_2) = U^{(t)}(q_1, q_2, v_3, v_4)$  if  $v_1 \leq v_2 < M$ ,  $v_3 \leq v_4 < M$  and  $v_2 - v_1 = v_4 - v_3$ . Thus, we can fix  $v_1 = 0$  and thereby save a factor of  $|\mathcal{V}| = M + 1$  in time and space. The special cases for  $v_2 = M$  or  $v_4 = M$  can be accommodated in the same bounds.  $\square$

Before turning the DFA into a PAA, one may wish to minimize it. Using an algorithm by Hopcroft [24], a classical DFA can be minimized in  $\mathcal{O}(|\mathcal{Q}| \log |\mathcal{Q}|)$  time for an alphabet of constant size, where  $\mathcal{Q}$  is the set of states. Refer to Knuutila [32] for a tutorial-like introduction and a variant that runs in  $\mathcal{O}(|\Sigma| \cdot |\mathcal{Q}| \log |\mathcal{Q}|)$  time when the alphabet size is not considered to be a constant. Hopcroft's algorithm can be adapted to minimize counting DFAs by using the partition induced by the different emissions as an initial partition, i.e. states with the same emission are grouped together.

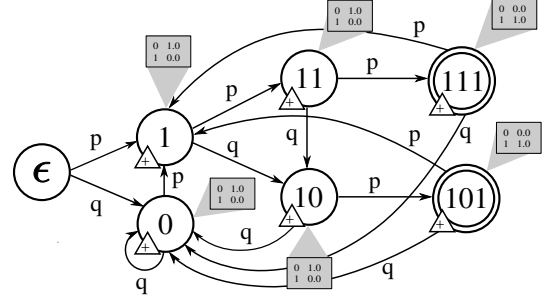
In the following, we review some concrete pattern classes important in practice, particularly in computational biology.

### 6.2.1 Finite Sets of Strings

Assume that the pattern is given in the form of a finite set of strings. In this situation, an Aho-Corasick automaton [1], which essentially is a DFA, can be built. It can be constructed in linear time by either using the algorithm given in the original paper or by employing a recent elegant algorithm based on the suffix tree of the reverse strings [20]. The emissions  $\eta_q$  (number of matches) can directly be read off the Aho-Corasick automaton's output function for all states  $q$ .



(a) PAA for overlapping hits



(b) PAA for non-overlapping hits

Figure 2: PAAs for the distribution of matches of the pattern set  $\{101, 111\}$ , assuming an i.i.d. text model over alphabet  $\Sigma = \{0, 1\}$  with probability  $p$  for character 1, and  $q := 1 - p$ . The start state is denoted  $\epsilon$ . Each state is associated with the operation “+” and a Dirac emission distribution, shown in the gray boxes.

Figure 2 shows a PAA computing the distribution of the number of (a) overlapping matches and (b) non-overlapping matches of the pattern set  $\{101, 111\}$  in an i.i.d. text model over the alphabet  $\{0, 1\}$ , where the probability of seeing 1 is  $p$ .

### 6.2.2 Finite Sets of Generalized Strings

Generalized strings are finite sequences of sets of characters over an alphabet  $\Sigma$ , for example  $[abc][ac][ab]$  (which matches  $aaa, ccb$  but not  $aba$ ). Rather than enumerating all strings matching a generalized string, we can directly construct an NFA that recognizes all strings ending with an instance of it. The NFA corresponding to one generalized string is just a linear chain of states; a start state plus one state for each position, where the start state is additionally equipped with a self-transition. The NFA for the set of generalized strings can be constructed by merging all individual start states into one common start state. The next step towards a PAA is to build a DFA. To obtain a DFA, we employ the classical subset construction. Although, in the worst case, it results in an exponential increase in the number of states, this method is feasible in many practical cases. In fact, the construction procedure can be modified such that it always results in the minimal DFA [46]. By the subset construction, a set  $B_q$  of NFA states corresponds to each DFA state  $q$ . The number of final NFA states in  $B_q$  equals the number of matching generalized strings that end when DFA state  $q$  is entered, giving us the number of matches  $\eta_q$  to be emitted by  $q$ .

### Prosit Patterns

Prosit is a database of biologically meaningful amino acid motifs (see Hulo et al. [26]). Prosit patterns can be seen as generalized strings with the extension that, for each position, a “multiplicity range” can be specified. In the pattern  $A-x(2,3)-C$ , for example,

an A is followed by either two or three arbitrary characters followed by a C. We translate every Prosite pattern into a set of generalized strings. The above example would result in the two patterns A-x-x-C and A-x-x-x-C. This set can then be dealt with as explained above.

We implemented the algorithms in Java and ran them on a Intel Core 2 Duo 2.66GHz, 4GB RAM, running Linux to assess practicability. Release 20.17 of Prosite contains 1319 patterns, 16 of which refer to the start or ending of a sequence. Those entries were ignored, leaving a database of 1303 patterns. For 42 patterns (3.2%) the computation did not succeed due to memory limitations. This can happen if either the Prosite pattern translates into too many generalized strings or if the DFA resulting from the subset construction grows too large. For 1236 of the 1261 remaining patterns, the subset construction was completed within 2 seconds while the computation took 69.9 seconds for the “worst pattern”. The resulting automata were minimized using Hopcroft’s algorithm. Many automata, however, already were minimal or close to minimal; for 1209 automata the minimized automaton was larger than half the size of the original automaton. The majority of resulting minimal automata were of reasonable size: We obtained 1198 automata with less than 10000 states, among which 1036 had less than 500 states.

To give an impression of the runtimes to be expected when computing the distribution of the overlapping occurrence count, consider the pattern

C-x-H-R-[GAR]-x(7,8)-[GEKVI]-[NERAQ]-x(4,5)-C-x-[FY]-H

from the Prosite database. It results in an automaton with 462 states. Assuming  $M = 50$  (maximum number of occurrences of interest) and  $n = 1000$  (text length), computing the distribution of the occurrence count took 1 second.

### 6.3 Waiting Time for Pattern Occurrences

The waiting time for the first occurrence of a pattern equals the waiting time for a state that emits a match. Therefore, its distribution can directly be computed by applying Lemma 4.4. The waiting time for a subsequent occurrence can be computed by choosing  $\alpha$  in Lemma 4.4 to be the equilibrium distribution restricted to all match states, i.e. those states that emit a match.

### 6.4 Clump Size Distribution

We already saw how to use a DFA recognizing a given pattern to construct a PAA. Through this method, we could accurately account for possible self-overlaps of patterns. The structure of self-overlaps was implicitly encoded in the DFA. In this section, we explicitly work out a pattern’s tendency to overlap itself by computing its clump size distribution  $\Psi$ . As detailed in [48], the exact clump size distribution of a pattern is, besides its theoretical value, useful for the construction of compound Poisson approximations. Compound Poisson approximations have also been discussed in [61, 70, 60].

**Definition 6.2.** *Given a sequence  $s \in \Sigma^*$  and a pattern  $p$ , a clump is a maximal set of overlapping occurrences of  $p$  in  $s$ .*

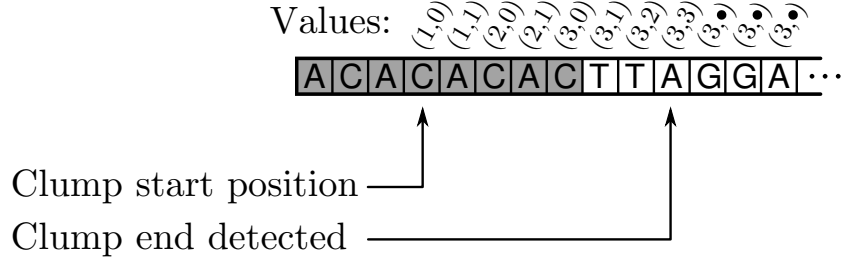


Figure 3: A clump (shaded gray) of three occurrences of the pattern **ACAC**. By definition, the clump starts at the last character of the first occurrence. The values computed by the clump size DAA are shown above the string. As soon as the second counter reaches  $m - 1 = 3$ , the clump has ended.

For example, let  $p := \text{ACA}$  and  $s := \text{GACACATTACAAA}$ . Then  $s$  contains three occurrences of  $p$  in two clumps (underlined). By definition, a clump consists of at least one match. We call the position of match's last character *match position* and consider the first match position in a clump. Further, we call the distribution of PAA states at such positions *clump start distribution* and denote it  $\gamma$ ; i.e. given that  $j$  is the first match position in a clump, then  $\mathbb{P}(Q_j = q) =: \gamma(q)$ , which is asymptotically independent of  $j$  under certain assumptions. For now, we assume  $\gamma$  to be known and come back to the task of its calculation later.

If  $m \geq 2$  is the length of the given motif, then a clump ends if  $m - 1$  consecutively visited states do not emit a match. That means we need to keep track of (a) the number of non-match states consecutively visited and (b) the number of matches the clump contains so far. The PAA framework allows this by modifying the PAAs described in the previous subsections. We define a new value set  $\mathcal{V}' := \mathbb{N}_0 \times \{0, \dots, m - 1, \bullet\}$  with the start value  $v'_0 := (0, 0)$  and attach the following semantics: If we are in state  $q$  and the current value is  $(h, x)$ , we have seen  $h$  matches in the current clump and the last of these matches occurred  $x$  steps in the past; i.e. if  $x = 0$ , a match has been emitted from the current state. The special value  $x = \bullet$  indicates that the clump has ended. We define the operations accordingly:

$$\theta'_q : ((h, x), e) \mapsto \begin{cases} (h + e, 0) & \text{if } e > 0 \text{ and } x \in \{0, \dots, m - 2\}, \\ (h, x + 1) & \text{if } e = 0 \text{ and } x \in \{0, \dots, m - 2\}, \\ (h, \bullet) & \text{if } x \in \{m - 1, \bullet\}. \end{cases}$$

In other words, if a match has been found ( $e > 0$ ), we increase the number of matches  $h$  by  $e$  and reset the distance to the last match to 0. Otherwise ( $e = 0$ , no match occurred),  $h$  remains unmodified, but the number of steps  $x$  since the last match is increased by one. See Figure 3 for an example. To incorporate the clump start distribution  $\gamma$ , we use one additional state  $q'_0$  that becomes the new start state; consequently, we set  $\mathcal{Q}' := \{q'_0\} \cup \mathcal{Q}$

and define the new transition function to be

$$T' : (q, q') \mapsto \begin{cases} \gamma(q') & \text{if } q = q'_0, \\ T(q, q') & \text{otherwise.} \end{cases} \quad (21)$$

The set  $\mathcal{V}'$  is infinite. As discussed in Section 3, this does not pose a problem as the range of each  $V_t$  is finite. Furthermore, for many applications it is sufficient to truncate the clump size distribution and use the value set  $\mathcal{V}'' := \{1, \dots, M\} \times \{0, \dots, m-1, \bullet\}$  along with adapted operations  $\theta''_q$ . Employing one of the algorithms shown in Sections 3.1 and 3.2, respectively, we can then calculate the joint state-value distributions  $\rho_t(q, h, x) := \mathbb{P}(Q_t = q, V_t = (h, x))$ . A clump ends if no new match has occurred  $m-1$  steps after the previous match. The clump size distribution  $\Psi$  is thus given by

$$\Psi(h) = \sum_{t=0}^{\infty} \sum_{q \in \mathcal{Q}} \rho_t(q, h, m-1). \quad (22)$$

To actually compute  $\Psi$ , we start with the initial table  $\rho_0$  and iteratively calculate the tables  $\rho_t$  for larger  $t$ . Each  $\rho_t$  contributes to the sought distribution through the inner sum from Equation (22) and we can successively add the contributions to an intermediate clump size distribution. Observe that the difference between the intermediate clump size distribution after iteration  $t$  and the exact one is bounded by

$$1 - \sum_{q \in \mathcal{Q}} \sum_{h=0}^M \rho_t(q, h, \bullet).$$

Thus, we iterate until this quantity drops under an accuracy threshold. The number of necessary steps, however, is bounded by  $\mathcal{O}(M \cdot m)$ , because a clump containing  $M$  matches can have a length of at most  $\mathcal{O}(M \cdot m)$ . In total, we need  $\mathcal{O}(|\Sigma| \cdot |\mathcal{C}| \cdot |\mathcal{Q}| \cdot M^2 \cdot m^3)$  time to compute the exact clump size distribution. Again, a factor of  $|\mathcal{C}|$  can be saved if for all  $c \in \mathcal{C}$  and  $\sigma \in \Sigma$ , there exists at most one  $c' \in \mathcal{C}$  such that  $\varphi(c, \sigma, c') > 0$ .

### State Distribution at Clump Start

Let us come back to computing the clump start distribution  $\gamma$  needed in Equation (21). As discussed in Section 2, the PAA's state process  $(Q_t)_{t \in \mathbb{N}_0}$  is a Markov chain and, hence, the classical theorems about existence of and convergence to an equilibrium distribution apply: Irreducibility and aperiodicity are sufficient for convergence to a unique equilibrium distribution. Assuming (a) that a pattern does not start with a wildcard and (b) all text model states  $c \in \mathcal{C}$  have positive occurrence probability, these conditions can be verified to be fulfilled by construction of the PAA.

We consider the joint distribution of state and steps since the last match position. Recall that  $E_t$  is the emission process of a PAA. We define  $L_t$  as the number of steps since we last encountered a match before step  $t$ . Thus

$$L_t := \min \{t' \in \{1, \dots, t\} \mid E_{t-t'} > 0\},$$

where we set  $L_t := -\infty$  if the set is empty, meaning that no match has occurred until step  $t-1$ . Again we use the PAA framework to compute the joint state-value distribution  $\mathcal{L}(Q_t, L_t)$  for any desired  $t$ . The clump start distribution is now given by

$$\gamma(q) = \lim_{t \rightarrow \infty} \mathbb{P}(Q_t = q \mid L_t \geq m, E_t > 0). \quad (23)$$

In practice, the limits for  $t \rightarrow \infty$  exist and converge in a few steps to double precision.

## 7 Analysis of Window-Based Pattern Matching Algorithms

The basic pattern matching problem is to find all occurrences of a *pattern* string in a (long) *text* string as fast as possible. Let  $n$  be the text length and  $m$  be the pattern length. The well-known Knuth-Morris-Pratt algorithm [31] reads each text character exactly once from left to right after preprocessing the pattern and needs a total of  $\Theta(n + m)$  character accesses. In contrast, the Boyer-Moore [10], Horspool [25] and Sunday [67] algorithms move a length- $m$  search window across the text and first compare its *last* character to the last character of the pattern. This often allows to move the search window by more than one position (at best, by  $m$  positions if the last window character does not occur in the pattern at all), for a best case of  $\Theta(m + n/m)$  but a worst case of  $\Theta(m + mn)$  character accesses. The worst case can be improved to  $\Theta(m + n)$ , but this makes the code more complicated and is seldom useful in practice. The Horspool algorithm and the variant of Sunday can be seen as modifications of the Boyer-Moore algorithm that are simpler to implement and additionally perform better in practice [52]. In general, a window-based algorithm that searches for a pattern  $p$  is characterized by

- a window size  $z$ ,
- a cost function  $\xi^p : \Sigma^z \rightarrow \mathbb{N}_0$  giving the cost caused by a window,
- a shift function  $shift^p : \Sigma^z \rightarrow \{1, \dots, m\}$  giving the number of positions the window can safely be shifted by.

The total cost of processing a text  $s \in \Sigma^n$  is denoted  $\xi^p(s)$ . In this section, we develop a methodology to calculate the exact distribution of  $\xi^p(S_0 \dots S_{n-1})$  when  $S$  is a random text and pattern  $p$  is fixed. This question has so far not been investigated, even though related questions have been answered in the literature. For example, [5, 6] analyze the expected number of character accesses for Horspool's algorithm. In [42] it is shown that the number of character accesses is asymptotically normally distributed for i.i.d. texts, and [65] extends this result to Markovian text models.

The technique we introduce here allows to compute the exact distribution of the total cost for arbitrary cost functions, general finite-memory text models as defined in Section 5, and applies to all window-based pattern matching algorithms. For concreteness, we consider Horspool's algorithm as given in Algorithm 2, that means, pattern and window are compared from right to left and the shift solely depends on the window's last character. Formally, Horspool's algorithm is characterized by



---

**Algorithm 2** HORSPOOL

---

**Input:** text  $s \in \Sigma^*$ , pattern  $p \in \Sigma^m$

**Output:** pair (number  $occ$  of occurrences of  $p$  in  $s$ , number  $cost$  of accesses to  $s$ )

```
1: pre-compute table  $shift[\sigma]$  for all  $\sigma \in \Sigma$ 
2:  $(occ, cost) \leftarrow (0, 0)$ 
3:  $t \leftarrow m - 1$ 
4: while  $t < |s|$  do
5:    $i \leftarrow 0$ 
6:   while  $i < m$  do
7:      $cost \leftarrow cost + 1$ 
8:     if  $s[t - i] \neq p[(m - 1) - i]$  then break
9:      $i \leftarrow i + 1$ 
10:  end while
11:  if  $i = m$  then  $occ \leftarrow occ + 1$ 
12:   $t \leftarrow t + shift[s[t]]$ 
13: end while
14: return  $(occ, cost)$ 
```

---

- $z := m$  (the window size equals the pattern length),
- $\xi^p(w) := \begin{cases} m, & \text{if } p = w, \\ \min \{i : 1 \leq i \leq m, p[m - i] \neq w[m - i]\}, & \text{otherwise,} \end{cases}$
- the shift depends on the position of the rightmost occurrence of  $w[m - 1]$  in  $p$ :

$$\begin{aligned} right^p(w) &:= \max [\{i \in \{0, \dots, m - 2\} : p[i] = w[m - 1]\} \cup \{-1\}], \\ shift^p(w) &:= (m - 1) - right^p(w). \end{aligned}$$

Here, we have used the number of character accesses as a measure of cost.

We now construct a PAA that allows to compute the cost distribution for arbitrary window-based pattern matching algorithms with respect to a random text of length  $n$  defined by a text model  $(\mathcal{C}, c_0, \Sigma, \varphi)$ . Note that we cannot construct a DAA and apply Lemma 5.4, as the considered algorithms can read several characters “at once”, while a DAA is only capable of processing one character at a time. In the defined model of window-based algorithms, shift and cost depend solely on the current window. Therefore, we model each possible window as a state (although concrete algorithms may permit smaller state spaces). Additionally, we need to keep track of the current text model state after generating the current window. We use the state space

$$\mathcal{Q} := \{(\varepsilon, c_0)\} \cup (\Sigma^z \times \mathcal{C}),$$

where  $(\varepsilon, c_0) =: q_0$  is the start state and any other state  $(w, c)$  corresponds to window content  $w$  and text model state  $c$ . Note that, if the text model is a Markovian model of

order  $r \leq z$ , then its state is fully determined by the current window. That means, only  $|\Sigma|^z + 1$  different states are reachable, all others can be discarded. For arbitrary text models, however, more states might be reachable. The strategy to construct the PAA is to simulate the algorithm and count both the reached position in the text and the cost so far. To this end, we assume that the text length  $n$  is fixed and define the value set

$$\mathcal{V} := \{0, \dots, n-1, \bullet\} \times \{0, \dots, n \cdot \hat{\xi}_{\square}^p\},$$

where  $\hat{\xi}_{\square}^p := \max\{\xi_{\square}^p(w) : w \in \Sigma^z\}$ , with  $\hat{\xi}_{\square}^p = m$  for Horspool's algorithm. Value  $v = (t, \xi)$  corresponds to the current window ending at text position  $t$  (with  $\bullet$  indicating that the text has ended), having accumulated a cost of  $\xi$  so far. The start value is set to  $v_0 := (z-1, 0)$ .

Each state deterministically emits cost and shift of the associated window, therefore we set

$$\mathcal{E} := \{1, \dots, m\} \times \{0, \dots, \hat{\xi}_{\square}^p\},$$

where an emission of  $(t', \xi') \in \mathcal{E}$  indicates that the current window has caused a cost of  $\xi'$  and the window is to be shifted by  $t'$ . Formally, we define

$$\mu_{(w,c)}((t', \xi')) := \begin{cases} 1 & \text{if } t' = \text{shift}^p(w) \text{ and } \xi' = \xi_{\square}^p(w), \\ 0 & \text{otherwise,} \end{cases}$$

for all  $(w, c) \in \mathcal{Q}$ . Note that we could use other (non-Dirac) emission distributions if necessary. We might assume, for example, that a cache miss occurs with a certain probability and associate higher costs with this event. The operation  $\theta_q$  in each state is essentially an addition on  $\mathcal{V}$  with one exception: To indicate that the complete text has been processed, we use the special values  $(\bullet, \xi) \in \mathcal{V}$ . We define:

$$\theta_q((t, \xi), (t', \xi')) := \begin{cases} (t+t', \xi+\xi') & \text{if } t \neq \bullet \text{ and } t+t' < n, \\ (\bullet, \xi+\xi') & \text{if } t \neq \bullet \text{ and } t+t' \geq n, \\ (\bullet, \xi) & \text{otherwise.} \end{cases}$$

The last component to be specified is the transition function  $T$ . The shift, and therefore all possible target states, are determined by the current window contents, that means, the suffix of the current window must match the prefix of the subsequent window. We define an indicator function that tells whether two windows are compatible in this sense:

$$I(w, w') := \llbracket w[\text{shift}^p(w)..] = w'[..z-1-\text{shift}^p(w)] \rrbracket.$$

While  $I(w, w')$  tells whether a transition from  $w$  to  $w'$  is allowed, its probability is given by the text model.

$$T((w, c), (w', c')) := \begin{cases} \mathbb{P}(c_0 \xrightarrow{w} c') & \text{if } (w, c) = (\varepsilon, c_0), \\ I(w, w') \cdot \mathbb{P}\left(c \xrightarrow{w'[z-\text{shift}^p(w)..]} c'\right) & \text{otherwise,} \end{cases}$$

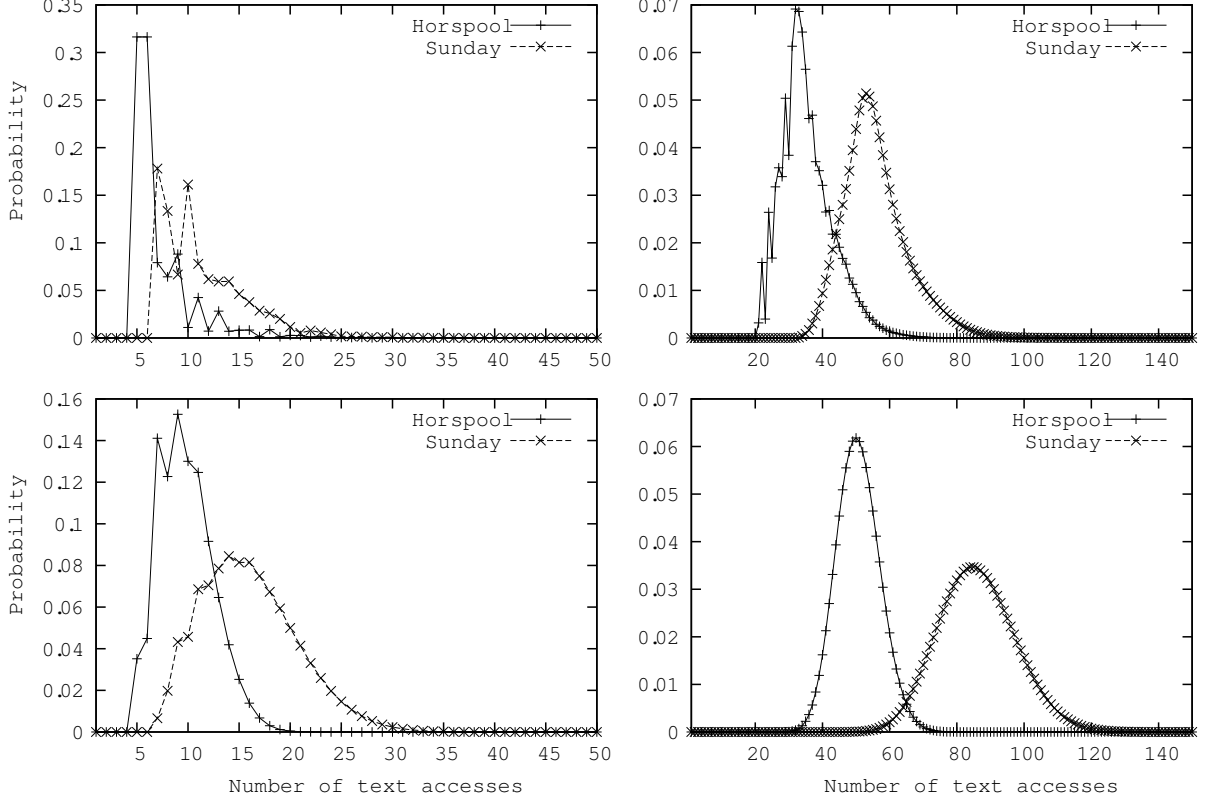


Figure 4: Exact distribution of the number of text accesses for Horspool's and Sunday's algorithms using a uniform i.i.d. text model over the alphabet  $\{A, C, G, T\}$ . Top vs. bottom row: pattern AAAAAA vs. ACAGC. Left vs. right column: text length 20 vs. 100.

where

$$\mathbb{P}\left(c'_0 \xrightarrow{\sigma_0 \dots \sigma_{j-1}} c'_j\right) := \sum_{c'_1, \dots, c'_{j-1} \in \mathcal{C}} \prod_{i=0}^{j-1} \varphi(c'_i, \sigma_i, c'_{i+1})$$

is the probability that the text model  $(\mathcal{C}, c_0, \Sigma, \varphi)$  produces the string  $\sigma[0] \dots \sigma[j-1] \in \Sigma^j$  while going from state  $c'_0 \in \mathcal{C}$  to state  $c'_j \in \mathcal{C}$  in  $j$  steps.

**Theorem 7.1.** *Let a window-based pattern matching algorithm specified by  $z$ ,  $\xi_{\square}^p$ , and  $\text{shift}^p$  and a text model  $(\mathcal{C}, c_0, \Sigma, \varphi)$  be given. The exact distribution of the cost of processing a random string of length  $n$  can be computed in  $\mathcal{O}(n^3 \cdot |\Sigma|^{2z} \cdot |\mathcal{C}| \cdot \hat{\xi}_{\square}^p)$  time and  $\mathcal{O}(n^2 \cdot |\Sigma|^z \cdot |\mathcal{C}| \cdot \hat{\xi}_{\square}^p)$  space. If the text model is Markovian of order  $r \leq z$ , then  $\mathcal{O}(n^3 \cdot |\Sigma|^{2z} \cdot \hat{\xi}_{\square}^p)$  time and  $\mathcal{O}(n^2 \cdot |\Sigma|^z \cdot \hat{\xi}_{\square}^p)$  space are sufficient.*

*Proof.* By construction, the PAA simulates the working of the specified algorithm. Assuming  $|\mathcal{E}| = \mathcal{O}(1)$ , runtime and space bounds follow from Lemma 3.2. If the text model is Markovian of order  $r \leq z$ , then only  $|\Sigma|^z + 1$  different states are reachable. The claimed time and space bound can then be met by using a reduced state space  $\mathcal{Q}' := \{\varepsilon\} \cup |\Sigma|^z$ .  $\square$

The exponential dependency on the window length  $z$  allows practical computations only for short patterns. For a pattern length of 5 and text lengths 20 and 100, a comparison of Horspool’s and Sunday’s algorithms is shown in Figure 4. The calculation took 1.8 seconds and 40.4 seconds for text length 20 and 100, respectively<sup>3</sup>. The plots reveal the combinatorial nature of the number of text accesses (which is asymptotically normally distributed) for short texts. It also reveals that Sunday’s algorithm needs more character accesses than Horspool’s algorithm for the example patterns.

## 8 Alignment Seed Sensitivity

We describe an application of the PAA framework to determine the quality of seeds used in homology search. Homologous biosequences have developed from a common ancestor and usually share high sequence similarity. In homology search, a sequence database is searched for a query sequence in order to find potential homologs, i.e. evolutionarily related sequences. To this end, each database sequence is compared with the query, which can be done by a local alignment method such as the Smith-Waterman algorithm [64].

### 8.1 Related Work and New Results

Since exact local alignment is too slow in practice, most heuristic homology search algorithms are based on a two-phase filtration technique [56, 2, 3, 30]. First, candidate sequences are selected that share a common pattern (“seed”) of matching characters with the query. These candidates (or “hits”) are then further investigated by an exact method. Initially, *contiguous seeds* (e.g., perfectly matching DNA 11-mers in the initial BLAST implementation) were used. PatternHunter (PH) by Ma et al. [41] was the first tool to systematically advocate and investigate *spaced seeds*: PH looks for 18-mers with at least 11 matching positions distributed as  $111*1**1*1**11*111$ , where 1 denotes a necessary match and \* denotes a don’t care position (match or mismatch). Over time, various seed models have been proposed in the literature, including consecutive seeds [56, 2], spaced seeds [41, 15, 11, 16], subset seeds [36], vector seeds [12], and indel seeds [43].

In the context of homology search, it is customary to model random *alignments* instead of random sequences. Typical models for such alignments may consist of several homology parameters and hence called *homology models*. They are described in Section 8.2. Different seeds in a class (e.g. all seeds with 11 match positions and length 18) can be compared according to their *sensitivity*, i.e. the probability to “hit” a random alignment of given length from a given homology model (see Definition 8.3 in Section 8.3 for a formal definition of “hit”).

A good seed exhibits high sensitivity for alignments that model evolutionarily related biosequences, and low sensitivity values for alignments that represent unrelated sequences. The latter property ensures that the seed does not detect too many *random*

---

<sup>3</sup>See <http://www.rahmannlab.de/software> for an implementation in JAVA. The experiments were run on an Intel Core 2 Quad CPU at 2.66GHz.

Table 1: Representative string  $\mathcal{A}$  of an ungapped alignment between two sequences.

Query	G	C	G	A	A	T	G	C	C	T
Database	G	C	C	A	A	C	G	C	T	T
$\mathcal{A}$	1	1	0	1	1	0	1	1	0	1

*hits*. Random hits decrease the efficiency of the filtration phase, since they are checked in vain for a significant alignment. An interesting finding was that the PH approach led to an increase in both sensitivity and filtration efficiency, compared to seeds of contiguous matches. Based on the observations in [41], the advantages of spaced seeds over consecutive seeds have been subsequently evaluated by many authors [15, 17, 38].

An extension to single seed models is the design of a multiple seed. This is a set of spaced seeds to be used simultaneously, such that a similarity is detected when it is found by (at least) one of the seeds. The idea to use a family of spaced seeds for BLAST-type DNA local alignment has been suggested by Ma et al. [41] and was implemented in PatternHunter II [37]. It has also been applied to local protein alignment in [14]. Recent approaches [33, 27] approximate the sensitivity of multiple spaced seeds by means of correlation functions. Since finding optimal multiple seeds is challenging, most authors concentrate on the design of efficient sets of seeds, leading to higher sensitivity than optimal single seeds [37, 35, 66].

When searching optimal seeds, one faces the following problems to evaluate candidate seeds:

**Problem 8.1** (Sensitivity computation). *Given a homology model, a target length  $n$ , and a set of seeds, what is the probability that a random alignment of length  $n$  is hit by the seed (at least once)?*

**Problem 8.2** (Hit distribution). *Given a homology model, a target length  $n$ , a set of seeds, and a maximal hit number  $K$ , what is the probability that a random alignment of length  $n$  is hit by the seed exactly  $k$  times (or at least  $k$  times), for each  $k = 0, \dots, K$ , when counting (a) overlapping hits, (b) non-overlapping hits?*

The second, more general question has not yet been investigated in the literature. As we show, the distribution is directly provided by the constructed PAA and allows the investigation of optimality criteria different from sensitivity alone.

## 8.2 Homology Models

We describe random alignments with known degree of similarity (e.g. a certain per cent identity value) by means of a *homology model*. A homology model generates representative strings  $\mathcal{A}$  over an alphabet  $\Sigma$  indicating the status of the alignment columns. In the simplest and most frequently studied case only substitution mutations and ungapped alignments are considered [41, 15, 11, 16, 17, 66]; see Table 1. That is  $\Sigma = \{0, 1\}$ , referring to matches (1) and mismatches (0).

Indel seeds, designed for gapped alignments, use the alignment alphabet  $\Sigma = \{0, 1, 2, 3\}$ , where additionally 2 denotes an insertion into the database sequence, and 3 indicates an insertion into the query sequence. There are various other alignment alphabets, e.g. the ternary alphabet representing a match or transition or transversion in DNA [54], or even larger alphabets to distinguish different pairs of amino acids in the case of proteins [14]. A representative string is modeled as a Markov chain  $(\Sigma, P, p^0)$  with a transition matrix  $P$  and an initial distribution  $p^0$  on the alphabet  $\Sigma$ .  $P$  and  $p^0$  are called *homology parameters*.

**Ungapped alignments.** We model ungapped alignments by an i.i.d. homology model with  $\Sigma = \{0, 1\}$ . In this case, the transition probability  $P(\sigma, \sigma')$  does not depend on  $\sigma$ . In particular,  $P$  takes the form  $\begin{pmatrix} 1-p & p \\ 1-p & p \end{pmatrix}$  for a *match probability*  $p \in [0, 1]$ , which quantifies the average identity of such alignments; for example,  $p \approx 0.3$  for unrelated,  $p \approx 0.95$  for closely related DNA sequences.

**Gapped alignments.** For gapped alignments, we use a first-order Markov chain. This is appropriate since the respective homology model should prohibit the pairs ‘23’ and ‘32’ in a representative string, because a substitution is more plausible than two consecutive indels. For our calculations, we used the transition matrix  $P$  proposed in [43]:

$$\begin{array}{c} \begin{matrix} & 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} \begin{pmatrix} p_0 & p_1 & p_g & p_g \\ p_0 & p_1 & p_g & p_g \\ p_0^* & p_1^* & p_g & 0 \\ p_0^* & p_1^* & 0 & p_g \end{pmatrix}, \end{array} \quad (24)$$

where  $p_0$  is the probability of a mismatch,  $p_1$  is the probability of a match, and  $p_g$  refers to the probability of a gap in the alignment. In order to obtain a stochastic transition matrix,  $p_\sigma^* = p_\sigma + p_g p_\sigma / (p_0 + p_1)$  for  $\sigma \in \{0, 1\}$  redistributes  $p_g$  to match and mismatch characters. The initial distribution is given by  $p^0 = (p_0, p_1, p_g, p_g)$ . Other transition probabilities are possible, e.g. if alignments with affine gap costs should be modeled.

### 8.3 Seed Models

A seed  $\pi = \pi[0]\pi[1]\dots\pi[L-1]$  is a string over an alphabet of “care” and “don’t care” characters. It represents alignment regions that indicate matches at the “care” positions. A seed is classified  $(L, \omega)$  by its *length*  $L = |\pi|$  and its *weight*  $\omega$ , which refers to the number of “care” positions.

A *contiguous seed* represents a region of contiguous matches, i.e.  $\pi[i] = 1$  for  $0 \leq i < L$ . A *spaced seed* is a string over the alphabet  $\Xi = \{1, *\}$ . The “care” positions are indicated by 1, while  $*$  refers to a match/mismatch wildcard. Reasonable seeds for the purpose of homology search always require  $\pi[0] = \pi[L-1] = 1$ . An *indel seed* according to Mak et al. [43] is a string over the alphabet  $\Xi = \{1, *, ?\}$ , where 1 and  $*$  are as above,

and ? stands for zero or one character from the alignment alphabet  $\Sigma = \{0, 1, 2, 3\}$ . Two consecutive ? symbols represent any character pair except ‘23’ or ‘32’. By means of this interpretation, the model explicitly allows for indels of variable size. For example, 1??1 may detect indels of size 0, 1, or 2. It hence tolerates 2, 1, or 0 match/mismatch positions.

A seed can thus be converted to a generalized string (see Section 6.2.2) over the alignment alphabet when we additionally allow to skip some characters ( $\epsilon$ -characters). In any case, a seed can be represented as a finite set of patterns over the alignment alphabet; this pattern set  $\mathcal{PS}(\pi)$  contains all instances of the generalized string.

**Definition 8.3** (Hit). *A hit of a seed  $\pi$  in an alignment  $\mathcal{A}$  is an occurrence<sup>4</sup> of a string from  $\mathcal{PS}(\pi)$  in  $\mathcal{A}$ . We call an ending position of a seed hit in  $\mathcal{A}$  a hit position.*

**Example 8.4.** *Consider the indel seed  $\pi = 1*1?1$ . It corresponds to the generalized string  $[1][01][1][\epsilon 0123][1]$ , and the instances are given by the pattern set*

$$\mathcal{PS}(\pi) = \{1011, 1111, 10101, 10111, 10121, 10131, 11101, 11111, 11121, 11131\}.$$

. *The seed hits the alignment string 1011011110 at positions 3, 6, 7, and 8, respectively.*

For a finite, non-empty set  $\Pi = \{\pi_1, \dots, \pi_m\}$  of spaced seeds, also called *multiple spaced seed*, the patterns are collected in  $\mathcal{PS}(\Pi) = \cup_{i=1}^m \mathcal{PS}(\pi_i)$ . A multiple seed is said to *hit*  $\mathcal{A}$ , if at least one of its components does.

## 8.4 PAAs for Seed Sensitivity and Applications

With the seed (set)  $\Pi$  represented as a pattern set  $\mathcal{PS}(\Pi)$ , and the alignment model being a finite memory text model (see Section 5.1), the problem has been reduced to computing the hit distribution of a finite set of patterns (Section 6.2.1). As mentioned in Section 6.2, both overlapping and non-overlapping seed hits can be considered. In fact, Figure 2 in Section 6.2.1 shows the PAA for seed  $\pi = 1*1$  with pattern set  $\mathcal{PS}(\pi) = \{111, 101\}$  in an i.i.d. text model to count (a) overlapping and (b) non-overlapping hits.

In contrast to previous work that only considers the sensitivity (probability of at least one hit, Problem 8.1), the PAA framework yields the entire match distribution (Problem 8.2). However, if only the sensitivity is desired, the value set can be reduced to  $\mathcal{V} = \{0, 1\}$ , reducing requirements to  $\mathcal{O}(n|\mathcal{Q}|^2)$  time and  $\mathcal{O}(|\mathcal{Q}|)$  space.

An example comparing a contiguous seed with the PH seed is shown in Table 2. For match probability  $p = 0.95$  (an alignment of highly similar sequences), a good seed should achieve a high sensitivity, or low probability for zero hits; indeed, the PH seed is almost two orders of magnitude better than the contiguous seed ( $6.7 \cdot 10^{-6}$  vs.  $4.1 \cdot 10^{-4}$  for  $k = 0$  hits). For  $p = 0.3$  (essentially a random alignment), a good seed should achieve a low sensitivity. Both sensitivity values are comparable ( $6.75 \cdot 10^{-5}$  for the contiguous seed;  $8.3 \cdot 10^{-5}$  for the PH seed). If we consider only candidates with at least two hits,

---

<sup>4</sup>A hit has previously been called a match or an occurrence, but here a match concerns a single position in an alignment, and the term “hit” seems more descriptive.

Table 2: Comparison of a seed requiring 11 contiguous matches in an alignment and the PH seed requiring at least 11 matches within 18 alignment columns at particular positions. The alignment model is ungapped i.i.d., with  $p$  being the probability of a match, and  $1 - p$  being the probability of a mismatch. We consider both highly similar sequences (top,  $p = 0.95$ ) and unrelated sequences (bottom,  $p = 0.3$ ). The tables show the probabilities for exactly  $k$  overlapping hits, for  $k = 0, \dots, 3$ , in a target region of length 64.

$p = 0.95$	$k = 0$	$k = 1$	$k = 2$	$k = 3$
111111111111	$4.1285 \cdot 10^{-4}$	$5.4005 \cdot 10^{-4}$	$8.4467 \cdot 10^{-4}$	0.0012
111*1**1*1**11*111	$6.7331 \cdot 10^{-6}$	$4.4978 \cdot 10^{-5}$	$1.6120 \cdot 10^{-4}$	$4.1669 \cdot 10^{-4}$
$p = 0.3$	$k = 0$	$k = 1$	$k = 2$	$k = 3$
111111111111	0.9999325	$4.7615 \cdot 10^{-5}$	$1.4025 \cdot 10^{-5}$	$4.1295 \cdot 10^{-6}$
111*1**1*1**11*111	0.9999170	$8.2780 \cdot 10^{-5}$	$2.2438 \cdot 10^{-7}$	$8.9947 \cdot 10^{-9}$

the PH seed retains its advantage for alignments of highly similar sequences and now outperforms the contiguous seed even for random alignments (probability of  $3.5 \cdot 10^{-5}$  for at least two matches for the contiguous seed, but only  $2.2 \cdot 10^{-7}$  for the PH seed).

The PAA framework provides a unifying method for computing seed sensitivity for different alignment models and seed models that were so far developed in an *ad-hoc* fashion in the literature. Furthermore, it is easily extended; let us mention two examples.

For a simple homology model with few parameters (say, the i.i.d. homology model for ungapped alignments with a single parameter  $p$ ), it is possible to evaluate the recurrence 6 symbolically, i.e., by representing the entries of the transition matrix  $T(q', q)$  and the probabilities of the state-value distribution  $f_t(q, v)$  as polynomials in the parameters. The resulting polynomial only needs to be computed once; then one can assess the sensitivity of a seed under different parameter values. This was previously presented by Mak et al. [44], without using the PAA approach.

The PAA framework can also be applied to design efficient sets of seeds. Similar to [66], we can successively find seeds that locally maximize the conditional sensitivity, given that the seeds already present in the set do not hit an alignment. Such a conditional sensitivity can be computed by making the existing seeds' accepting states absorbing without counting a match, so only instances of the current seed candidate are counted. Evaluating each seed candidate and picking the best one yields the seed to be added to the set.

## 9 Protein Fragment Mass Statistics

Peptide mass fingerprinting (PMF) is a technique for protein identification based on mass spectrometry (MS) and database search. Here, we present a PAA to compute the significance of protein identification by PMF. The protein of interest is enzymatically cleaved into smaller peptides, whose masses are determined by MS. Masses are measured



in Dalton (Da) or unified atomic mass units (u), where  $1 \text{ Da} = 1 \text{ u}$  equals one twelfth of the mass of an isolated atom of carbon-12 ( $^{12}\text{C}$ ) at rest and in its ground state, or  $1.66053878 \cdot 10^{-24} \text{ g}$ . The set of peptide masses, the so-called peptide mass fingerprint, is then used to query a database of known proteins. Each database sequence is processed *in silico* to obtain a theoretical fingerprint, and experimental and theoretical fingerprints are subsequently compared and scored. The highest-scoring database protein is the most likely candidate for the unknown sample.

In order to come up with a reasonable scoring for such a comparison, one is interested in the probability that a certain peptide fragment mass occurs by chance. For example, a mass of 445.2 Da is characteristic for the short fragment DVCK (aspartatic acid, valine, cysteine, lysine), which occurs in many known proteins, and is therefore not a helpful feature for protein identification.

Additionally, fragment masses vary for several reasons: The constituting atoms have different isotope masses; about 98.9% of all carbon atoms, for example, have six neutrons while about 1.1% have seven neutrons and are therefore heavier. Post-translational modifications (additions of chemical groups to some amino acids of the protein) may occur. Some cleavage sites may be missed by the protease, and one may observe one larger mass that corresponds to the sum of two expected masses.

Therefore, we are interested in the mass distribution of proteolytic fragments, or the joint length-mass distribution of such fragments, to subsequently answer questions about the probability that at least one fragment with a given approximate mass exists in a protein of given or typical length.

An approach chosen by many PMF software packages is to avoid probability computations and use empirical frequencies of fragments in large protein databases instead. Kaltenbach [28] introduced p-value-based scores that are database-independent, but based on a text model that represents protein sequences as i.i.d. random strings, i.e. as sequences of independent random variables that take values from the alphabet of amino acids according to frequencies estimated from the Swissprot database [7, 8].

Here, we generalize this result and present a PAA based on arbitrary finite-memory text models. First, we construct a DAA encoding the enzymatic cleavage reaction. Second, from this DAA and a model for random sequences of amino acids, a PAA is constructed by means of Lemma 5.4. The resulting PAA provides statistics of peptides usually measured in an MS experiment, in particular, the length distribution and the joint length-mass distribution of such fragments. Furthermore, we compute the probability that the peptide mass fingerprint of a random protein contains at least one peptide of a certain mass. This, in turn, provides a significance value for PMF identifications and can be used to derive a scoring scheme, as explored in [28]. We further incorporate the influence of isotopic distributions, incomplete cleavage, and post-translational modifications by appropriate modifications of the PAA.

## 9.1 PAA for Statistics of Proteolytic Fragments

Cleavage enzymes cut proteins at well-defined places which are often determined by one or two adjacent amino acids (see [69, 28]). For instance, the most widely used cleavage

agent Trypsin cleaves after lysine (K) and arginine (R) unless the next amino acid is proline (P). Such cleavage sites can be described by  $\Gamma\bar{\Pi}$  with *cleavage characters*  $\Gamma \subset \Sigma$  and *prohibition characters*  $\Pi \subset \Sigma$ , where  $\bar{\Pi}$  denotes the complement of  $\Pi$  in the amino acid alphabet  $\Sigma = \{A, \dots, Z\} \setminus \{B, J, O, U, X, Z\}$ .

Due to their distinct structure, we distinguish the first fragment  $F_1$  from following fragments  $F_+$ . While the first fragment may start with a prohibition character, following fragments do not. For i.i.d. text models, statistics for all following fragments are i.i.d. as well (assuming an infinite string length), see [69, 28]. For arbitrary finite-memory text models, however, statistics of fragments  $F_2, F_3, \dots$  may differ as well. In any case, statistics for fragments  $F_k$  converge as  $k$  grows to infinity. We write  $F_\star$  to denote an arbitrary fragment (either first or following). The last fragment does not necessarily end with a cleavage character because of the finiteness of protein sequences.

First, we construct a DAA  $\mathcal{D} = (\mathcal{Q}, q_0, \Sigma, \delta, \mathcal{V}, v_0, \mathcal{E}, (\eta_q)_{q \in \mathcal{Q}}, (\theta_q)_{q \in \mathcal{Q}})$  that sums up amino acid masses (for now, we ignore isotopes) until a cleavage point is encountered. We set  $\mathcal{Q} := \Sigma \cup \{q_0, \bullet, \circ\}$ , where  $q_0$  is the start state. For each state  $q \in \Sigma$ , the emission  $\eta_q$  gives the mass of amino acid  $q$ ; the set of possible values is  $\mathcal{V} := \mathbb{R}^+$ , where  $v_0 := 0$  is the start value. Note that, despite the infinity of  $\mathcal{V}$ , the number of values reachable in  $n$  steps ( $\vartheta_n$ ) is finite for all  $n$  (see Section 3). For states  $q' \in \{q_0, \bullet, \circ\}$  no mass is emitted, that means,  $\eta_{q'} = 0$ . To sum amino acid masses, we define operations by  $\theta_q : (v, e) \mapsto v + e$  for all  $q \in \mathcal{Q}$ . The transition function  $\delta : \mathcal{Q} \times \Sigma \rightarrow \mathcal{Q}$  reflects the cleavage rules:

$$\delta(q, \sigma) = \begin{cases} \sigma & \text{if } q \in (\Sigma \setminus \Gamma) \cup \{q_0\}, \\ \sigma & \text{if } q \in \Gamma, \sigma \in \Pi, \\ \bullet & \text{if } q \in \Gamma, \sigma \notin \Pi, \\ \circ & \text{if } q \in \{\bullet, \circ\}. \end{cases}$$

As can be easily verified, this DAA has the property that  $\delta(q_0, s[..i+1]) = \bullet$  if and only if the first fragment in  $s$  has length  $i$ . Then,  $\text{value}_{\mathcal{D}}(s[..i+1]) = m$ , where  $m$  is the mass of the first fragment.

From the DAA and a text model, a PAA is obtained by using Lemma 5.4. To take the isotopic mass distribution of each amino acid into account, each state's emission distribution can be changed accordingly. For practical computations, it is sufficient to perform calculations up to a limited mass accuracy as the accuracy of MS instruments is limited as well. One could, for example, consider masses up to one decimal position. A sketch of a PAA for first fragments and a first-order Markovian text model is shown in Figure 5 (including isotopic distributions). From now on, we assume that masses have been scaled and then rounded to integers to achieve the desired precision.

The difference between the first fragment  $F_1$  and the following fragments  $F_2, F_3, \dots$  lies solely in the initial state distribution. The transition probabilities from the start state to other states need to be adjusted in such a way that (1) prohibition characters are forbidden at the beginning of a following fragment and (2) the transitions reflect the distribution of text model states at the end of the previous fragment. This distribution can be obtained from the PAA for the previous fragment. Note again that statistics for all fragments  $F_k$  with  $k > 1$  are i.i.d. if the underlying text model is i.i.d. as well.

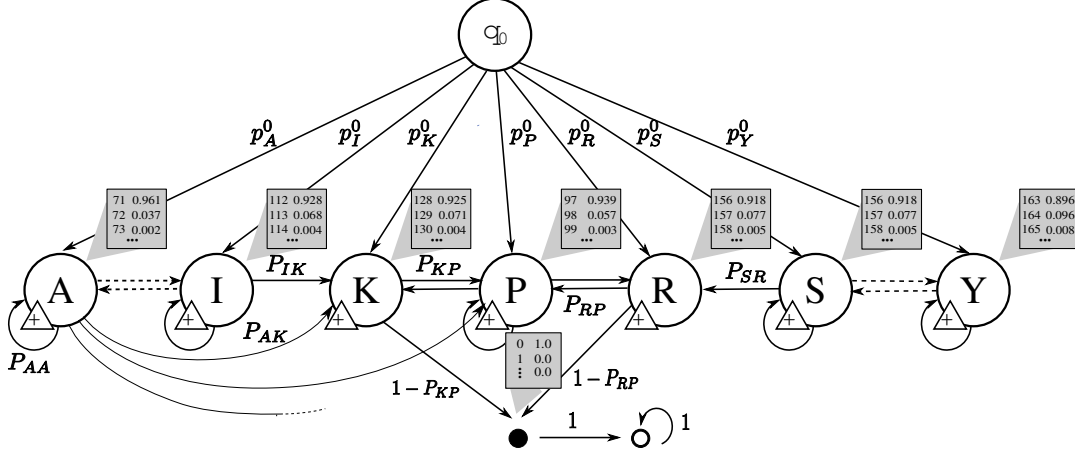


Figure 5: Sketch of the PAA measuring the mass of the first fragment resulting from tryptic cleavage of a random protein. Following fragments are handled by a modified PAA where the initial distribution is normalized by  $1 - p_P$  and start in  $P$  is prohibited since following fragments cannot start with a prohibition character. For the sake of simplicity, not all transitions are shown. The states' weight distributions correspond to the respective isotopic distributions. Here, integer masses are displayed. The operation associated to each state is “+”.

## 9.2 Mass and Length Distributions

By means of the constructed PAAs, we compute fragment statistics as the length distribution and the joint length-mass distribution. We denote the length of a fragment by  $L(F_\star)$  and the (integer) mass of a fragment by  $M(F_\star)$ . The length of a random fragment generated by the PAA corresponds to the waiting time for reaching state  $\bullet$  minus one (as the last processed character does not belong to the fragment). When only the distribution of fragment lengths (and not the mass distribution) is required, we can use Lemma 4.4 to compute the distribution of this waiting time.

Generally, the joint state-value distribution  $f_n(q, v) = \mathbb{P}(Q_n = q, V_n = v)$  discussed in Section 3 yields the joint length-mass distribution  $\nu_\star$  of  $F_\star$ :

$$\nu_\star(n, m) := \mathbb{P}(L(F_\star) = n, M(F_\star) = m) = f_{n+1}(\bullet, m) = \mathbb{P}(Q_{n+1} = \bullet, V_{n+1} = m).$$

The computation of  $f_{n+1}$  takes  $\mathcal{O}(n \cdot |\mathcal{Q}|^2 \cdot \vartheta_n \cdot |\mathcal{E}|)$  time and  $\mathcal{O}(|\mathcal{Q}| \cdot \vartheta_n)$  space (see Lemma 3.2). Since both  $|\mathcal{Q}|$  and  $|\mathcal{E}|$  are constants, and the range of possible masses  $\vartheta_n$  grows linearly with both  $n$  and the mass scaling factor (or mass precision)  $\lambda$ , we need  $\mathcal{O}(\lambda n^2)$  time and  $\mathcal{O}(\lambda n)$  space.

## 9.3 Mass Occurrence Probabilities

For the interpretation of mass spectra, the most important quantity is the mass occurrence probability of a measured mass  $m$ , i.e. the probability that the fragmentation of a random protein sequence contains at least one fragment of mass  $m$ . To account

for possibly inaccurate measurements, it is advisable to compute the probability that a fragment in a certain mass range  $[m - \Delta, m + \Delta]$  occurs. We modify the PAA such that it does not move into the absorbing state  $\circ$  once the first fragment has ended. We therefore remove state  $\circ$  and replace  $\bullet$  by  $|\Sigma|$  states named  $\bullet_\sigma$  for each  $\sigma \in \Sigma$ . That means,  $\mathcal{Q} := \{q_0\} \cup \Sigma \cup \{\bullet_\sigma : \sigma \in \Sigma\}$ . Being in state  $\bullet_\sigma$  means that a fragment has ended and the first character of the next fragment is  $\sigma$ . Therefore, the emission distribution of  $\bullet_\sigma$  is the same as for state  $\sigma$ . In terms of the transition function  $\delta$ , each state  $\bullet_\sigma$  acts like state  $\sigma$ . Formally,

$$\delta(q, \sigma) := \begin{cases} \sigma & \text{if } q \in (\Sigma \setminus \Gamma) \cup \{q_0\}, \\ \sigma & \text{if } q \in \Gamma, \sigma \in \Pi, \\ \bullet_\sigma & \text{if } q \in \Gamma, \sigma \notin \Pi, \\ \delta(\sigma', \sigma) & \text{if } q = \bullet_{\sigma'}. \end{cases}$$

To keep track of whether a fragment in the given mass range has already been observed, we introduce an absorbing value  $\diamond$ ; that means, all operations are adapted such that once the value  $\diamond$  has been attained, it is not changed. The new states  $\bullet_\sigma$  get the following operations:

$$\theta_{\bullet_\sigma}(v, e) := \begin{cases} \diamond & \text{if } v \in [m - \Delta, m + \Delta], \\ e & \text{otherwise.} \end{cases}$$

The probability that a protein sequence of length  $n$  contains a segment in the mass range  $[m - \Delta, m + \Delta]$  is given by  $\mathbb{P}(V_n = \diamond) + \mathbb{P}(V_n \in [m - \Delta, m + \Delta])$ , where the second summand accounts for the probability that the last fragment has the sought mass.

## 9.4 Missed Cleavages and Post-Translational Modifications

Protein identification by MS is complicated by the occurrence of partial enzymatic cleavage which results in peptides with internal missed cleavage sites. Even for Trypsin, which is reported to have a high cleavage specificity, incomplete digestion is not uncommon [63, 51]. The presented PAA can be modified to account for missed cleavages by adjusting the transition probabilities outgoing from the cleavage characters: The probability to end a fragment after reading a cleavage character is reduced, while transitions from a cleavage character to non-prohibition characters occur with a small probability. Moreover, in a Markovian text model, one can include information about missed cleavage patterns, i.e. the probabilities can be weighted according to the amino acid propensities in proximity to missed cleavage sites.

Another complicating issue is that many proteins are modified after translation. A *post-translational modification* (PTM) is a chemical process that changes the properties of a protein. It includes the addition of functional groups such as acetate or phosphate, the modification of amino acids, or structural changes such as the formation of disulfide bridges. Putative modifications can themselves be discovered by means of MS, comparing the mass measured to the mass expected for the identified protein or peptide [4, 45, 68]. PTMs result in a change in the molecular mass of the protein. Examples

of common modifications include phosphorylation (+80 Da), acetylation (+42 Da), and methylation (+14 Da or +28 Da). An overview of important PTMs along with the resulting mass shifts in Da is provided in the review of Mann and Jensen [45]. The PAA presented above can be modified to incorporate PTMs as follows: “Global” modifications of the protein are incorporated into the mass distribution of either the start state  $q_0$  or of an end state  $\circ$  or  $\bullet$ . Amino acid-specific modifications are incorporated by modifying the mass distribution of that particular amino acid state, possibly splitting the state into several copies accounting for different modification probabilities. Reasonable frequencies of PTMs associated with particular amino acids are provided in a study of Tsur et al. [68].

## 9.5 Characteristic Masses of Protein Families

PAAs provide the possibility to combine any probabilistic protein model with a cleavage scheme and, hence, to obtain a probabilistic PMF for the model. This allows to compute statistics of “random” peptide fragments, as shown above, but also to compute characteristic masses of protein families, which are often represented by probabilistic models, such as HMMs in the Pfam database [22]. As character-emitting HMMs are special cases of finite-memory text models (see Section 5.1), we can combine a Pfam model with a DAA to compute fragment statistics specific to the protein family.

Mass  $m$  is said to be specific for a protein family (for a particular cleavage scheme, with respect to a set of protein families) if the occurrence probability of fragments with mass  $m$  is greater than  $1 - \varepsilon$  for some small  $\varepsilon > 0$  and if fragments of mass  $m$  are not expected in any other protein family. We are not aware of any work done in this area; hence it is unknown if family-specific fragment masses exist for any cleavage scheme.

## 10 High-Throughput Sequencing

With the 454 sequencing technology, up to a million DNA reads of lengths up to 400 nucleotides can be determined in parallel by pyrosequencing. Nucleotides are successively synthesized to single-stranded DNA templates evoking an enzymatic cascade, resulting in detectable flashes of light. These signals are reported in a so-called *pyrogram* from which the sequence of nucleotides (the *read*) is determined. In 454 sequencing, the four different nucleotides are sequentially flowed over the reaction plate which holds the DNA templates. The particular order of nucleotide flows is called *dispensation order* and repeated for several (say, 100) cycles. The 454 systems GS20 and GS-FLX use the standard dispensation order TACG. Whenever the next nucleotide in the template sequence matches the currently dispensed nucleotide (in reality, its complement, but this is a technical detail), the dispensed nucleotide extends the complementary strand of the template. In case of a homopolymer run (the same type of nucleotide appearing several times consecutively), the strand is extended by the corresponding number of nucleotides at once. Such an event is detected by a correspondingly more intense flash of light. The more recent IonTorrent technology has the same characteristics, but uses a different

Table 3: Depending on the dispensation order, different numbers of nucleotides can be recognized during three cycles. The dispensation orders **TACG** and **GTCA** are compared for the sequence **GTCGTATCCC**. Note the homopolymer run of three Cs, which is sequenced with a single flow.

Cycle	Dispensation order <b>TACG</b>		Dispensation order <b>GTCA</b>	
	Nucleotide	Recognized?	Nucleotide	Recognized?
1	T	-	G	✓
	A	-	T	✓
	C	-	C	✓
	G	✓	A	-
2	T	✓	G	✓
	A	-	T	✓
	C	✓	C	-
	G	✓	A	✓
3	T	✓	G	-
	A	✓	T	✓
	C	-	C	✓✓✓
	G	-	A	-
Read length:		6	10	

underlying technology (change in pH instead of light detection). Table 3 shows how the sequence of a template fragment is determined with two different dispensation orders.

We investigate the exact length distribution of 454 sequencing reads, assuming a fixed number of nucleotide flows, a given dispensation order, and an arbitrary finite-memory text model. Previously, Rahmann [57] studied the combinatorics of sequences that can be reliably sequenced by the 454 technology. Kong [34] obtained generating functions for the length distribution, but the dispensation order was restricted to a permutation of the nucleotides, and the text model was restricted to an i.i.d. model.

Here we design a PAA that allows for arbitrary finite-memory text models and arbitrary dispensation orders: Every ordering of nucleotides, where each nucleotide occurs at least once and no nucleotide is flowed twice consecutively, is reasonable, e.g., TCGACG.

## 10.1 PAA for the Length Distribution of 454 Reads

Instead of attacking the length distribution problem directly, we first construct a DAA that computes the number of nucleotide flows needed to read (a finite prefix of) an input sequence with a given dispensation order  $d = d[0] \dots d[\ell - 1]$ . In other words, in each step, the DAA reads one nucleotide to be sequenced, and the emitted value models the waiting time (number of flows) since the previous sequenced nucleotide. We cast the length distribution problem as a waiting time problem by waiting for a fixed number  $f$

of flows (Section 4; waiting for a value, Lemma 4.2). As a consequence, we obtain the exact length distribution  $\mathcal{L}(L) = \mathcal{L}(L(f, d))$  of reads sequenced after  $f$  nucleotide flows according to a dispensation order  $d$ . In practice,  $f$  is usually a multiple of  $\ell = |d|$ .

We define a DAA with state set  $\mathcal{Q}^{\mathcal{D}} := (\Sigma \times \{0, \dots, \ell - 1\}) \cup \bigcup_{i=-1}^{\ell-1} \{(\epsilon, i)\}$  with start state  $q_0 := (\epsilon, -1)$  and attach the following semantics: Being in state  $(\sigma, j)$  means that the last two sequenced nucleotides were  $\sigma$  and  $d[j]$ .

The (deterministic) emission of state  $(\sigma, j)$  tells us how many flows it took to reach nucleotide  $d[j]$  after  $\sigma$ : If  $d[j] = \sigma$ , they are part of the same homopolymer run and the emission is zero; otherwise, it is the smallest forward distance between these nucleotides in the (cyclically interpreted) dispensation order. For example, if  $d = \text{TCGACG}$ ,  $\sigma = \text{G}$ , and  $j = 1$ , so  $d[j] = \text{C}$ , then the emission is 2 (skipping over  $d[0] = \text{T}$ ). Formally,

$$\eta_{(\sigma, j)} := \min \{i \in \{0, \dots, \ell - 1\} \mid d[j - i \bmod \ell] = \sigma\}. \quad (25)$$

The transition target from state  $(\sigma, j)$  when reading character  $\sigma'$  is consequently

$$\delta((\sigma, j), \sigma') := \left( d[j], j + \min \{i \in \{0, \dots, \ell - 1\} \mid d[j + i \bmod \ell] = \sigma'\} \bmod \ell \right);$$

this is also valid for  $\sigma = \epsilon$ . Intuitively, the “old”  $j$  becomes the “new”  $\sigma = d[j]$ , and the “new”  $j$  is obtained by cyclically searching forward in the dispensation order. The initial transition targets and emissions are given by

$$\begin{aligned} \delta((\epsilon, -1), \sigma') &:= (\epsilon, \min \{i \in \{0, \dots, \ell - 1\} \mid d[i] = \sigma'\}); \\ \eta_{(\epsilon, j)} &:= j + 1 \quad \text{for } j \in \{-1, \dots, \ell - 1\}. \end{aligned}$$

The sequencing protocol specifies the number  $f$  of nucleotide flows; typically  $f = 400$  for the standard dispensation order TACG, i.e. 100 cycles. The operation in each state adds the emitted number of flows, truncating at  $f + 1$ . This completes the construction of a DAA.

We are interested in the sequence length distribution up to a given length  $n$ . We define the target set  $\mathcal{T} := \{f + 1\}$  (cf. Definition 4.1) and obtain the read length as one less than the waiting time

$$W_{\mathcal{T}} = \min \{t \in \mathbb{N}_0 \mid V_t = f + 1\}.$$

Invoking Lemma 5.4 yields the corresponding PAA for a general finite-memory text model, from which we obtain the read length distribution up to a given  $n$ . Applying Lemma 5.6’s statement about waiting times with  $\Sigma = \mathcal{O}(1)$ ,  $|\mathcal{C}| = \mathcal{O}(1)$ ,  $|\mathcal{Q}^{\mathcal{D}}| = \mathcal{O}(\ell)$  and  $\vartheta_n = \mathcal{O}(\ell n)$  results in a running time of  $\mathcal{O}(n^2 \ell^2)$  and a space requirement of  $\mathcal{O}(n \ell^2)$ .

As an application, we compared the expected read lengths under an estimated first-order Markov model for all reasonable dispensation orders of length 4 to 10 on two datasets of 454 reads of yet unfinished strains of the GC-rich bacteria *S. meliloti* and *R. Lupinii*, provided by the Genetics Department of Bielefeld University. We observed that the standard machine settings could be improved to yield approximately 10% longer reads on average, namely 282 nt instead of 257 nt with the GS-FLX instrument during  $f = 400$  nucleotide flows.

## 11 Discussion

We have presented the concept of probabilistic arithmetic automata that blends well into the landscape of existing stochastic models like Markov chains and hidden Markov models. In fact, PAAs can be seen as Markov chains over a larger state space. The benefit of PAAs lies in their utility as a modelling technique, specifically (1) the required state space is often small, (2) the connection between states and values becomes evident, and (3) an elegant notation and well-arranged recurrences are obtained. As shown in the second part of this article, many applications can conveniently be approached using the PAA framework. The method is especially suited for applications involving the deterministic processing of random texts. In these cases, a PAA can be constructed from a deterministic arithmetic automaton and a finite-memory text model, a quite general class of text models that covers simple i.i.d. models as well as arbitrary-order Markov models and character-emitting HMMs.

## Acknowledgments

The authors wish to thank Jens Stoye and Sebastian Böcker for their valuable comments. Timo Stöcker carefully read the manuscript. IH was supported by the NRW Graduate School in Bioinformatics and Genome Research, Bielefeld.

## References

- [1] A. V. Aho and M. J. Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.
- [2] S. F. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.
- [3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997.
- [4] R. S. Annan and S. A. Carr. The essential role of mass spectrometry in characterizing protein structure: Mapping posttranslational modifications. *J. Protein Chem.*, 16(5):391–402, 1997.
- [5] R. A. Baeza-Yates, G. H. Gonnet, and M. Régnier. Analysis of Boyer-Moore-type string searching algorithms. In *SODA '90: Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, pages 328–343. SIAM, 1990.
- [6] R. A. Baeza-Yates and M. Régnier. Average running time of the Boyer-Moore-Horspool algorithm. *Theor. Comput. Sci.*, 92(1):19–31, 1992.



- [7] A. Bairoch and B. Boeckmann. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res*, 19 Suppl:2247–2249, 1991.
- [8] B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31(1):365–370, 2003.
- [9] V. Boeva, J. Clément, M. Régnier, M. A. Roytberg, and V. J. Makeev. Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms for Molecular Biology*, 2:13, October 2007.
- [10] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Communications of the ACM*, 20(10):762–772, 1977.
- [11] B. Brejová, D. G. Brown, and T. Vinar. Optimal spaced seeds for homologous coding regions. *J. Bioinform. Comput. Biol.*, 1(4):595–610, 2004.
- [12] B. Brejová, D. G. Brown, and T. Vinar. Vector seeds: an extension to spaced seeds. *J. Computer System Sci.*, 70(3):364–380, 2005.
- [13] P. Brémaud. *Markov chains, Gibbs fields, Monte Carlo simulation, and queues*. Springer, 1999.
- [14] D. G. Brown. Optimizing multiple seeds for protein homology search. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2(1):29–38, 2005.
- [15] J. Buhler, U. Keich, and Y. Sun. Designing seeds for similarity search in genomic dna. In *Proc. of the 7th annual international conference on Research in computational molecular biology (RECOMB)*, pages 67–75, 2003.
- [16] K. P. Choi, F. Zeng, and L. Zhang. Good spaced seeds for homology search. *Bioinformatics*, 20(7):1053–1059, 2004.
- [17] K. P. Choi and L. Zhang. Sensitivity analysis and efficient method for identifying optimal spaced seeds. *J. Computer System Sci.*, 68:22–40, 2004.
- [18] Erhan Cinlar. Markov additive processes I. *Z. Wahrscheinl. verw. Geb.*, 24:85–93, 1972.
- [19] Erhan Cinlar. Markov additive processes II. *Z. Wahrscheinl. verw. Geb.*, 24:95–121, 1972.
- [20] S. Dori and G. M. Landau. Construction of Aho Corasick automaton in linear time for integer alphabets. *Information Processing Letters*, 98(2):66–72, 2006.
- [21] W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley & sons, 1968.

- [22] Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O. Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–D222, 2010.
- [23] I. Herms and S. Rahmann. Computing alignment seed sensitivity with probabilistic arithmetic automata. In *Proc. of the 8th Workshop on Algorithms in Bioinformatics (WABI)*, volume 5251 of *Lecture Notes in Computer Science*, pages 318–329. Springer, 2008.
- [24] J. Hopcroft. An  $n \log n$  algorithm for minimizing the states in a finite automaton. In Z. Kohavi and A. Paz, editors, *The theory of machines and computations*, pages 189–196. Academic Press, New York, 1971.
- [25] R. N. Horspool. Practical fast searching in strings. *Software-Practice and Experience*, 10:501–506, 1980.
- [26] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. A. Sigrist. The PROSITE database. *Nucleic Acids Research*, 34(S1):D227–230, 2006.
- [27] L. Ilie and S. Ilie. Multiple spaced seeds for homology search. *Bioinformatics*, 23(22):2969–2977, 2007.
- [28] H.-M. Kaltenbach. *Statistics and Algorithms for Peptide Mass Fingerprinting*. PhD thesis, Bielefeld University, 2007.
- [29] H.-M. Kaltenbach, S. Böcker, and S. Rahmann. Markov additive chains and applications to fragment statistics for peptide mass fingerprinting. In Trey Ideker and Vineet Bafna, editors, *Systems Biology and Computational Proteomics*, volume 4532 of *LNCS*, pages 29–41. Springer, 2006.
- [30] W. J. Kent. BLAT—the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664, 2002.
- [31] D. E. Knuth, Jr. Morris, and V. R. Pratt. Fast pattern matching in strings. *SIAM Journal on Computing*, 6(2):323–350, 1977.
- [32] T. Knuutila. Re-describing an algorithm by Hopcroft. *Theoretical Computer Science*, 250:333–363, 2001.
- [33] Y. Kong. Generalized correlation functions and their applications in selection of optimal multiple spaced seeds for homology search. *J. Comput. Biol.*, 14(2):238–254, 2007.
- [34] Y. Kong. Statistical distributions of pyrosequencing. *J Comput Biol*, 16(1):31–42, 2009.

- [35] G. Kucherov, L. Noé, and M. Roytberg. Multiseed lossless filtration. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 02(1):51–61, 2005.
- [36] G. Kucherov, L. Noé, and M. Roytberg. A unifying framework for seed sensitivity and its application to subset seeds. *J. Bioinform. Comput. Biol.*, 4(2):553–569, 2006.
- [37] M. Li, B. Ma, D. Kisman, and J. Tromp. PatternHunter II: Highly sensitive and fast homology search. *J. Bioinform. Comput. Biol.*, 2(3):417–439, 2004.
- [38] M. Li, B. Ma, and L. Zhang. Superiority and complexity of the spaced seeds. In *Proc. of the 17th annual ACM-SIAM symposium on discrete algorithms (SIAM)*, pages 444–453, 2006.
- [39] M. Lladser, M. D. Betterton, and R. Knight. Multiple pattern matching: A Markov chain approach. *Journal of Mathematical Biology*, 56(1-2):51–92, 2008.
- [40] M. Lothaire. *Applied Combinatorics on Words (Encyclopedia of Mathematics and its Applications)*. Cambridge University Press, 2005.
- [41] B. Ma, J. Tromp, and M. Li. PatternHunter - faster and more sensitive homology search. *Bioinformatics*, 18:440–445, 2002.
- [42] H. M. Mahmoud, R. T. Smythe, and M. Régnier. Analysis of Boyer-Moore-Horspool string-matching heuristic. *Random Structures and Algorithms*, 10(1-2):169–186, 1997.
- [43] D. Mak, Y. Gelfand, and G. Benson. Indel seeds for homology search. *Bioinformatics*, 22(14):e341–e349, 2006.
- [44] D. Y. F. Mak and G. Benson. All hits all the time: Parameter free calculation of seed sensitivity. In *Proc. of the 5th asia-pacific APBC bioinformatics conference (APBC)*, volume 5 of *Advances in Bioinformatics and Computational Biology*, pages 327–340. Imperial College Press, 2007.
- [45] M. Mann and O. N. Jensen. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, 21(3):255–261, 2003.
- [46] T. Marschall. Construction of minimal DFAs from biological motifs. *arXiv*, 1004.1298, 2010. <http://arxiv.org/abs/1004.1298>.
- [47] T. Marschall and S. Rahmann. Probabilistic arithmetic automata and their application to pattern matching statistics. In Paolo Ferragina and Gad M. Landau, editors, *Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching (CPM)*, volume 5029 of *LNCS*, pages 95–106. Springer-Verlag, 2008.
- [48] T. Marschall and S. Rahmann. Efficient exact motif discovery. *Proceedings of ISMB (Bioinformatics)*, 25(12):i356–364, June 2009.

- [49] T. Marschall and S. Rahmann. Exact analysis of pattern matching algorithms with probabilistic arithmetic automata. *arXiv*, 1009.6114, 2010.
- [50] T. Marschall and S. Rahmann. Exact analysis of Horspool’s and Sunday’s pattern matching algorithms with probabilistic arithmetic automata. In *Proceedings of the 4th International Conference on Language and Automata Theory and Applications (LATA 2010)*, to appear.
- [51] R. Matthiesen. *Mass Spectrometry Data Analysis in Proteomics*. Humana Press, 2007.
- [52] G. Navarro and M. Raffinot. *Flexible pattern matching in strings*. Cambridge University Press, 2002.
- [53] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287:593–617, 2002.
- [54] L. Noé and G. Kucherov. Improved hit criteria for dna local alignment. *BMC Bioinformatics*, 5:149, 2004.
- [55] G. Nuel. Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *Journal of Applied Probability*, 45:226–243, 2008.
- [56] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.
- [57] S. Rahmann. Subsequence combinatorics and applications to microarray production, DNA sequencing and chaining algorithms. In *Combinatorial Pattern Matching (CPM)*, volume 4009 of *LNCS*, pages 153–164, 2006.
- [58] M. Régnier. A unified approach to word occurrence probabilities. *Discrete Applied Mathematics*, 104:259–280, 2000.
- [59] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words: An overview. *Journal of Computational Biology*, 7(1-2):1–46, 2000.
- [60] E. Roquain and S. Schbath. Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain. *Advances in Applied Probability*, 39(1):128–140, 2007.
- [61] S. Schbath. Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*, 1:1–16, 1995.
- [62] M. Schulz, D. Weese, T. Rausch, A. Döring, K. Reinert, and M. Vingron. Fast and adaptive variable order markov chain construction. In Keith A. Crandall and Jens Lagergren, editors, *Proceedings of the 8th International Workshop on Algorithms in Bioinformatics (WABI), Karlsruhe, Germany*, volume 5251 of *LNCS*, pages 306–317, 2008.

- [63] J. A. Siepen, E. Keevil, D. Knight, and S. J. Hubbard. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J. Proteome Res.*, 6(1):399–408, 2007.
- [64] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, 1981.
- [65] R. T. Smythe. The Boyer-Moore-Horspool heuristic with Markovian input. *Random Structures and Algorithms*, 18(2):153–163, 2001.
- [66] Y. Sun and J. Buhler. Designing multiple simultaneous seeds for DNA similarity search. *J. Comput. Biol.*, 12(6):847–861, 2005.
- [67] D. M. Sunday. A very fast substring search algorithm. *Communications of the ACM*, 33(8):132–142, 1990.
- [68] D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. A. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. In *Proc. of the IEEE Computer Society Bioinformatics Conference (CSB)*, pages 157–166, 2005.
- [69] I.-J. Wang, C. P. Diehl, and F. J. Pineda. A statistical model of proteolytic digestion. In *Proc. of the IEEE Computer Society Bioinformatics Conference (CSB)*, pages 506–508, 2003.
- [70] M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall/CRC, 1995.
- [71] J. Zhang, B. Jiang, M. Li, J. Tromp, X. Zhang, and M. Q. Zhang. Computing exact p-values for DNA motifs. *Bioinformatics*, 23(5):531–537, 2007.