

## A unifying framework for seed sensitivity and its application to subset seeds (Extended abstract)

Gregory Kucherov, Laurent Noé, Mikhail Roytberg

► **To cite this version:**

Gregory Kucherov, Laurent Noé, Mikhail Roytberg. A unifying framework for seed sensitivity and its application to subset seeds (Extended abstract). WABI, 2005, Majorca/Spain, Spain. inria-00001164

**HAL Id: inria-00001164**

**<https://hal.inria.fr/inria-00001164>**

Submitted on 24 Mar 2006

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A unifying framework for seed sensitivity and its application to subset seeds

## (Extended abstract)

Gregory Kucherov<sup>1</sup>, Laurent Noé<sup>1</sup>, and Mikhail Roytberg<sup>2</sup>

<sup>1</sup> INRIA/LORIA, 615, rue du Jardin Botanique, B.P. 101, 54602, Villers-lès-Nancy, France, {Gregory.Kucherov,Laurent.Noé}@loria.fr

<sup>2</sup> Institute of Mathematical Problems in Biology, Pushchino, Moscow Region, 142290, Russia, royberg@impb.psn.ru

**Abstract.** We propose a general approach to compute the seed sensitivity, that can be applied to different definitions of seeds. It treats separately three components of the seed sensitivity problem – a set of target alignments, an associated probability distribution, and a seed model – that are specified by distinct finite automata. The approach is then applied to a new concept of *subset seeds* for which we propose an efficient automaton construction. Experimental results confirm that sensitive subset seeds can be efficiently designed using our approach, and can then be used in similarity search producing better results than ordinary spaced seeds.

## 1 Introduction

In the framework of pattern matching and similarity search in biological sequences, seeds specify a class of short sequence motif which, if shared by two sequences, are assumed to witness a potential similarity. Spaced seeds have been introduced several years ago [1,2] and have been shown to improve significantly the efficiency of the search. One of the key problems associated with spaced seeds is a precise estimation of the sensitivity of the associated search method. This is important for comparing seeds and for choosing most appropriate seeds for a sequence comparison problem to solve.

The problem of seed sensitivity depends on several components. First, it depends on the *seed model* specifying the class of allowed seeds and the way that seeds match (*hit*) potential alignments. In the basic case, seeds are specified by binary words of certain length (*span*), possibly with a constraint on the number of 1's (*weight*). However, different extensions of this basic seed model have been proposed in the literature, such as multi-seed (or multi-hit) strategies [3,4,2], seed families [5,6,7,8,9,10], seeds over non-binary alphabets [11,12], vector seeds [13,10].

The second parameter is the class of *target alignments* that are alignment fragments that one aims to detect. Usually, these are *gapless* alignments of a given length. Gapless alignments are easy to model, in the simplest case they

are represented by binary sequences in the match/mismatch alphabet. This representation has been adopted by many authors [2,14,15,16,17,18]. The binary representation, however, cannot distinguish between different types of matches and mismatches, and is clearly insufficient in the case of protein sequences. In [13,10], an alignment is represented by a sequence of real numbers that are *scores* of matches or mismatches at corresponding positions. A related, but yet different approach is suggested in [12], where DNA alignments are represented by sequences on the ternary alphabet of match/transition/transversion. Finally, another generalization of simple binary sequences was considered in [19], where alignments are required to be *homogeneous*, i.e. to contain no sub-alignment with a score larger than the entire alignment.

The third necessary ingredient for seed sensitivity estimation is the probability distribution on the set of target alignments. Again, in the simplest case, alignment sequences are assumed to obey a Bernoulli model [2,16]. In more general settings, Markov or Hidden Markov models are considered [17,15]. A different way of defining probabilities on binary alignments has been taken in [19]: all homogeneous alignments of a given length are considered equiprobable.

Several algorithms for computing the seed sensitivity for different frameworks have been proposed in the above-mentioned papers. All of them, however, use a common dynamic programming (DP) approach, first brought up in [14].

In the present paper, we propose a general approach to computing the seed sensitivity. This approach subsumes the cases considered in the above-mentioned papers, and allows to deal with new combinations of the three seed sensitivity parameters. The underlying idea of our approach is to specify each of the three components – the seed, the set of target alignments, and the probability distribution – by a separate finite automaton.

A deterministic finite automaton (DFA) that recognizes all alignments matched by given seeds was already used in [17] for the case of ordinary spaced seeds. In this paper, we assume that the set of target alignments is also specified by a DFA and, more importantly, that the probabilistic model is specified by a *probability transducer* – a probability-generating finite automaton equivalent to HMM with respect to the class of generated probability distributions.

We show that once these three automata are set, the seed sensitivity can be computed by a unique general algorithm. This algorithm reduces the problem to a computation of the total weight over all paths in an acyclic graph corresponding to the automaton resulting from the product of the three automata. This computation can be done by a well-known dynamic programming algorithm [20,21] with the time complexity proportional to the number of transitions of the resulting automaton. Interestingly, all above-mentioned seed sensitivity algorithms considered by different authors can be reformulated as instances of this general algorithm.

In the second part of this work, we study a new concept of *subset seeds* – an extension of spaced seeds that allows to deal with a non-binary alignment alphabet and, on the other hand, still allows an efficient hashing method to locate seeds. For this definition of seeds, we define a DFA with a number of

states independent of the size of the alignment alphabet. Reduced to the case of ordinary spaced seeds, this DFA construction gives the same worst-case number of states as the Aho-Corasick DFA used in [17]. Moreover, our DFA has always no more states than the DFA of [17], and has substantially less states on average.

Together with the general approach proposed in the first part, our DFA gives an efficient algorithm for computing the sensitivity of subset seeds, for different classes of target alignments and different probability transducers. In the experimental part of this work, we confirm this by running an implementation of our algorithm in order to design efficient subset seeds for different probabilistic models, trained on real genomic data. We also show experimentally that designed subset seeds allow to find more significant alignments than ordinary spaced seeds of equivalent selectivity.

## 2 General Framework

Estimating the seed sensitivity amounts to compute the probability for a random word (target alignment), drawn according to a given probabilistic model, to belong to a given language, namely the language of all alignments matched by a given seed (or a set of seeds).

### 2.1 Target Alignments

Target alignments are represented by words over an alignment alphabet  $\mathcal{A}$ . In the simplest case, considered most often, the alphabet is binary and expresses a match or a mismatch occurring at each alignment column. However, it could be useful to consider larger alphabets, such as the ternary alphabet of match/transition/transversion for the case of DNA (see [12]). The importance of this extension is even more evident for the protein case ([10]), where different types of amino acid pairs are generally distinguished.

Usually, the set of target alignments is a finite set. In the case considered most often [2,14,15,16,17,18], target alignments are all words of a given length  $n$ . This set is trivially a regular language that can be specified by a deterministic automaton with  $(n + 1)$  states. However, more complex definitions of target alignments have been considered (see e.g. [19]) that aim to capture more adequately properties of biologically relevant alignments. In general, we assume that the set of target alignments is a finite regular language  $L_T \in \mathcal{A}^*$  and thus can be represented by an acyclic DFA  $T = \langle Q_T, q_T^0, q_T^F, \mathcal{A}, \psi_T \rangle$ .

### 2.2 Probability Assignment

Once an alignment language  $L_T$  has been set, we have to define a probability distribution on the words of  $L_T$ . We do this using probability transducers.

A probability transducer is a finite automaton without final states in which each transition outputs a *probability*.

**Definition 1.** A probability transducer  $G$  over an alphabet  $\mathcal{A}$  is a 4-tuple  $\langle Q_G, q_G^0, \mathcal{A}, \rho_G \rangle$ , where  $Q_G$  is a finite set of states,  $q_G^0 \in Q_G$  is an initial state, and  $\rho_G : Q_G \times \mathcal{A} \times Q_G \rightarrow [0, 1]$  is a real-valued probability function such that  $\forall q \in Q_G, \sum_{q' \in Q_G, a \in \mathcal{A}} \rho_G(q, a, q') = 1$ .

A transition of  $G$  is a triplet  $e = \langle q, a, q' \rangle$  such that  $\rho(q, a, q') > 0$ . Letter  $a$  is called the *label* of  $e$  and denoted  $\text{label}(e)$ . A probability transducer  $G$  is *deterministic* if for each  $q \in Q_G$  and each  $a \in \mathcal{A}$ , there is at most one transition  $\langle q, a, q' \rangle$ . For each path  $P = (e_1, \dots, e_n)$  in  $G$ , we define its *label* to be the word  $\text{label}(P) = \text{label}(e_1) \dots \text{label}(e_n)$ , and the associated probability to be the product  $\rho(P) = \prod_{i=1}^n \rho_G(e_i)$ . A path is *initial*, if its start state is the initial state  $q_G^0$  of the transducer  $G$ .

**Definition 2.** The probability of a word  $w \in \mathcal{A}^*$  according to a probability transducer  $G = \langle Q_G, q_G^0, \mathcal{A}, \rho_G \rangle$ , denoted  $\mathcal{P}_G(w)$ , is the sum of probabilities of all initial paths in  $G$  with the label  $w$ .  $\mathcal{P}_G(w) = 0$  if no such path exists. The probability  $\mathcal{P}_G(L)$  of a finite language  $L \subseteq \mathcal{A}^*$  according to a probability transducer  $G$  is defined by  $\mathcal{P}_G(L) = \sum_{w \in L} \mathcal{P}_G(w)$ .

Note that for any  $n$  and for  $L = A^n$  (all words of length  $n$ ),  $\mathcal{P}_G(L) = 1$ .

Probability transducers can express common probability distributions on words (alignments). Bernoulli sequences with independent probabilities of each symbol [2,16,18] can be specified with deterministic one-state probability transducers. In Markov sequences of order  $k$  [17,6], the probability of each symbol depends on  $k$  previous symbols. They can therefore be specified by a deterministic probability transducer with at most  $|\mathcal{A}|^k$  states.

A Hidden Markov model (HMM) [15] corresponds, in general, to a non-deterministic probability transducer. The states of this transducer correspond to the (hidden) states of the HMM, plus possibly an additional initial state. Inversely, for each probability transducer, one can construct an HMM generating the same probability distribution on words. Therefore, non-deterministic probability transducers and HMMs are equivalent with respect to the class of generated probability distributions. The proofs are straightforward and are omitted due to space limitations.

### 2.3 Seed automata and seed sensitivity

Since the advent of spaced seeds [1,2], different extensions of this idea have been proposed in the literature (see Introduction). For all of them, the set of possible alignment fragments matched by a seed (or by a set of seeds) is a finite set, and therefore the set of matched alignments is a regular language. For the original spaced seed model, this observation was used by Buhler et al. [17] who proposed an algorithm for computing the seed sensitivity based on a DFA defining the language of alignments matched by the seed. In this paper, we extend this approach to a general one that allows a uniform computation of seed sensitivity for a wide class of settings including different probability distributions on target alignments, as well as different seed definitions.

Consider a seed (or a set of seeds)  $\pi$  under a given seed model. We assume that the set of alignments  $L_\pi$  matched by  $\pi$  is a regular language recognized by a DFA  $S_\pi = \langle Q_S, q_S^0, Q_S^F, \mathcal{A}, \psi_S \rangle$ . Consider a finite set  $L_T$  of target alignments and a probability transducer  $G$ . Under this assumptions, the sensitivity of  $\pi$  is defined as the conditional probability

$$\frac{\mathcal{P}_G(L_T \cap L_\pi)}{\mathcal{P}_G(L_T)}. \quad (1)$$

An automaton recognizing  $L = L_T \cap L_\pi$  can be obtained as the product of automata  $T$  and  $S_\pi$  recognizing  $L_T$  and  $L_\pi$  respectively. Let  $K = \langle Q_K, q_K^0, Q_K^F, \mathcal{A}, \psi_K \rangle$  be this automaton. We now consider the product  $W$  of  $K$  and  $G$ , denoted  $K \times G$ , defined as follows.

**Definition 3.** *Given a DFA  $K = \langle Q_K, q_K^0, Q_K^F, \mathcal{A}, \psi_K \rangle$  and a probability transducer  $G = \langle Q_G, q_G^0, \mathcal{A}, \rho_G \rangle$ , the product of  $K$  and  $G$  is the probability-weighted automaton  $W = \langle Q_W, q_W^0, Q_W^F, \mathcal{A}, \rho_W \rangle$  (for short, PW-automaton) such that  $Q_W = Q_K \times Q_G$ ,  $q_W^0 = (q_K^0, q_G^0)$ ,  $Q_W^F = \{(q_K, q_G) | q_K \in Q_K^F\}$ ,  $\rho_W((q_K, q_G), a, (q'_K, q'_G)) = \rho_G(q_G, a, q'_G)$  if  $\psi_K(q_K, a) = q'_K$ , and 0 otherwise.*

$W$  can be viewed as a non-deterministic probability transducer with final states.  $\rho_W((q_K, q_G), a, (q'_K, q'_G))$  is the probability of the  $\langle (q_K, q_G), a, (q'_K, q'_G) \rangle$  transition. A path in  $W$  is called *full* if it goes from the initial to a final state.

**Lemma 1.** *Let  $G$  be a probability transducer. Let  $L$  be a finite language and  $K$  be a deterministic automaton recognizing  $L$ . Let  $W = G \times K$ . The probability  $\mathcal{P}_G(L)$  is equal to sum of probabilities of all full paths in  $W$ .*

*Proof.* Since  $K$  is a deterministic automaton, each word  $w \in L$  corresponds to a single accepting path in  $K$  and the paths in  $G$  labeled  $w$  (see Definition 1) are in one-to-one correspondence with the full path in  $W$  accepting  $w$ . By definition,  $\mathcal{P}_G(w)$  is equal to the sum of probabilities of all paths in  $G$  labeled  $w$ . Each such path corresponds to a unique path in  $W$ , with the same probability. Therefore, the probability of  $w$  is the sum of probabilities of corresponding paths in  $W$ . Each such path is a full path, and paths for distinct words  $w$  are disjoint. The lemma follows.

## 2.4 Computing Seed Sensitivity

Lemma 1 reduces the computation of seed sensitivity to a computation of the sum of probabilities of paths in a PW-automaton.

**Lemma 2.** *Consider an alignment alphabet  $\mathcal{A}$ , a finite set  $L_T \subseteq \mathcal{A}^*$  of target alignments, and a set  $L_\pi \subseteq \mathcal{A}^*$  of all alignments matched by a given seed  $\pi$ . Let  $K = \langle Q_K, q_K^0, Q_K^F, \mathcal{A}, \psi_K \rangle$  be an acyclic DFA recognizing the language  $L = L_T \cap L_\pi$ . Let further  $G = \langle Q_G, q_G^0, \mathcal{A}, \rho \rangle$  be a probability transducer defining a probability distribution on the set  $L_T$ . Then  $\mathcal{P}_G(L)$  can be computed in time  $\mathcal{O}(|Q_G|^2 \cdot |Q_K| \cdot |\mathcal{A}|)$  and space  $\mathcal{O}(|Q_G| \cdot |Q_K|)$ .*

*Proof.* By Lemma 1, the probability of  $L$  with respect to  $G$  can be computed as the sum of probabilities of all full paths in  $W$ . Since  $K$  is an acyclic automaton, so is  $W$ . Therefore, the sum of probabilities of all full paths in  $W$  leading to final states  $q_W^F$  can be computed by a classical DP algorithm [20] applied to acyclic directed graphs ([21] presents a survey of application of this technique to different bioinformatic problems). The time complexity of the algorithm is proportional to the number of transitions in  $W$ .  $W$  has  $|Q_G| \cdot |Q_K|$  states, and for each letter of  $\mathcal{A}$ , each state has at most  $|Q_G|$  outgoing transitions. The bounds follow.

Lemma 2 provides a general approach to compute the seed sensitivity. To apply the approach, one has to define three automata:

- a deterministic acyclic DFA  $T$  specifying a set of target alignments over an alphabet  $\mathcal{A}$  (e.g. all words of a given length, possibly verifying some additional properties),
- a (generally non-deterministic) probability transducer  $G$  specifying a probability distribution on target alignments (e.g. Bernoulli model, Markov sequence of order  $k$ , HMM),
- a deterministic DFA  $S_\pi$  specifying the seed model via a set of matched alignments.

As soon as these three automata are defined, Lemma 2 can be used to compute probabilities  $\mathcal{P}_G(L_T \cap L_\pi)$  and  $\mathcal{P}_G(L_T)$  in order to estimate the seed sensitivity according to equation (1).

Note that if the probability transducer  $G$  is deterministic (as it is the case for Bernoulli models or Markov sequences), then the time complexity is  $\mathcal{O}(|Q_G| \cdot |Q_K| \cdot |\mathcal{A}|)$ . In general, the complexity of the algorithm can be improved by reducing the involved automata. Buhler et al. [17] introduced the idea of using the Aho-Corasick automaton [22] as the seed automaton  $S_\pi$  for a spaced seed. The authors of [17] considered all binary alignments of a fixed length  $n$  distributed according to a Markov model of order  $k$ . In this setting, the obtained complexity was  $\mathcal{O}(w2^{s-w}2^kn)$ , where  $s$  and  $w$  are seed’s span and weight respectively. Given that the size of the Aho-Corasick automaton is  $\mathcal{O}(w2^{s-w})$ , this complexity is automatically implied by Lemma 2, as the size of the probability transducer is  $\mathcal{O}(2^k)$ , and that of the target alignment automaton is  $\mathcal{O}(n)$ . Compared to [17], our approach explicitly distinguishes the descriptions of matched alignments and their probabilities, which allows us to automatically extend the algorithm to more general cases.

Note that the idea of using the Aho-Corasick automaton can be applied to more general seed models than individual spaced seeds (e.g. to multiple spaced seeds, as pointed out in [17]). In fact, all currently proposed seed models can be described by a finite set of matched alignment fragments, for which the Aho-Corasick automaton can be constructed. We will use this remark in later sections.

The sensitivity of a spaced seed with respect to an HMM-specified probability distribution over binary target alignments of a given length  $n$  was studied by Brejova et al. [15]. The DP algorithm of [15] has a lot in common with the algorithm implied by Lemma 2. In particular, the states of the algorithm of [15]

are triples  $\langle w, q, m \rangle$ , where  $w$  is a prefix of the seed  $\pi$ ,  $q$  is a state of the HMM, and  $m \in [0..n]$ . The states therefore correspond to the construction implied by Lemma 2. However, the authors of [15] do not consider any automata, which does not allow to optimize the preprocessing step (counterpart of the automaton construction) and, on the other hand, does not allow to extend the algorithm to more general seed models and/or different sets of target alignments.

A key to an efficient solution of the sensitivity problem remains the definition of the seed. It should be expressive enough to be able to take into account properties of biological sequences. On the other hand, it should be simple enough to be able to locate seeds fast and to get an efficient algorithm for computing seed sensitivity. According to the approach presented in this section, the latter is directly related to the size of a DFA specifying the seed.

### 3 Subset seeds

#### 3.1 Definition

Ordinary spaced seeds use the simplest possible binary “match-mismatch” alignment model that allows an efficient implementation by hashing all occurring combinations of matching positions. A powerful generalization of spaced seeds, called *vector seeds*, has been introduced in [13]. Vector seeds allow one to use an arbitrary alignment alphabet and, on the other hand, provide a flexible definition of a hit based on a cooperative contribution of seed positions. A much higher expressiveness of vector seeds lead to more complicated algorithms and, in particular, prevents the application of direct hashing methods at the seed location stage.

In this section, we consider *subset seeds* that have an intermediate expressiveness between spaced and vector seeds. It allows an arbitrary alignment alphabet and, on the other hand, still allows using a direct hashing for locating seed, which maps each string to a unique entry of the hash table. We also propose a construction of a seed automaton for subset seeds, different from the Aho-Corasick automaton. The automaton has  $\mathcal{O}(w2^{s-w})$  states *regardless of the size of the alignment alphabet*, where  $s$  and  $w$  are respectively the span of the seed and the number of “must-match” positions. From the general algorithmic framework presented in the previous section (Lemma 2), this implies that the seed sensitivity can be computed for subset seeds with same complexity as for ordinary spaced seeds. Note also that for the binary alignment alphabet, this bound is the same as the one implied by the Aho-Corasick automaton. However, for larger alphabets, the Aho-Corasick construction leads to  $\mathcal{O}(w|\mathcal{A}|^{s-w})$  states. In the experimental part of this paper (section 4.1) we will show that even for the binary alphabet, our automaton construction yields a smaller number of states in practice.

Consider an alignment alphabet  $\mathcal{A}$ . We always assume that  $\mathcal{A}$  contains a symbol  $1$ , interpreted as “match”. A *subset seed* is defined as a word over a *seed alphabet*  $\mathcal{B}$ , such that



- letters of  $\mathcal{B}$  denote subsets of the alignment alphabet  $\mathcal{A}$  containing 1 ( $\mathcal{B} \subseteq \{1\} \cup 2^{\mathcal{A}}$ ),
- $\mathcal{B}$  contains a letter  $\#$  that denotes subset  $\{1\}$ ,
- a subset seed  $b_1 b_2 \dots b_m \in \mathcal{B}^m$  matches an alignment fragment  $a_1 a_2 \dots a_m \in \mathcal{A}^m$  if  $\forall i \in [1..m], a_i \in b_i$ .

The  $\#$ -weight of a subset seed  $\pi$  is the number of  $\#$  in  $\pi$  and the *span* of  $\pi$  is its length.

*Example 1.* [12] considered the alignment alphabet  $\mathcal{A} = \{1, \mathbf{h}, 0\}$  representing respectively a match, a transition mismatch, or a transversion mismatch in a DNA sequence alignment. The seed alphabet is  $\mathcal{B} = \{\#, @, _\}$  denoting respectively subsets  $\{1\}$ ,  $\{1, \mathbf{h}\}$ , and  $\{1, \mathbf{h}, 0\}$ . Thus, seed  $\pi = \# @ _ \#$  matches alignment  $s = 10\mathbf{h}1\mathbf{h}1101$  at positions 4 and 6. The span of  $\pi$  is 4, and the  $\#$ -weight of  $\pi$  is 2.

Note that unlike the weight of ordinary spaced seeds, the  $\#$ -weight cannot serve as a measure of seed selectivity. In the above example, symbol  $@$  should be assigned weight 0.5, so that the weight of  $\pi$  is equal to 2.5 (see [12]).

### 3.2 Subset Seed Automaton

Let us fix an alignment alphabet  $\mathcal{A}$ , a seed alphabet  $\mathcal{B}$ , and a seed  $\pi = \pi_1 \pi_2 \dots \pi_m \in \mathcal{B}^*$  of span  $m$  and  $\#$ -weight  $w$ . Let  $R_\pi$  be the set of all non- $\#$  positions in  $\pi$ ,  $|R_\pi| = r = m - w$ . We now define an automaton  $S_\pi = \langle Q, q_0, Q_f, \mathcal{A}, \psi : Q \times \mathcal{A} \rightarrow Q \rangle$  that recognizes the set of all alignments matched by  $\pi$ .

The states  $Q$  of  $S_\pi$  are pairs  $\langle X, t \rangle$  such that  $X \subseteq R_\pi, t \in [0, \dots, m]$ , with the following invariant condition. Suppose that  $S_\pi$  has read a prefix  $s_1 \dots s_p$  of an alignment  $s$  and has come to a state  $\langle X, t \rangle$ . Then  $t$  is the length of the longest suffix of  $s_1 \dots s_p$  of the form  $1^i$ ,  $i \leq m$ , and  $X$  contains all positions  $x_i \in R_\pi$  such that prefix  $\pi_1 \dots \pi_{x_i}$  of  $\pi$  matches a suffix of  $s_1 \dots s_{p-t}$ .

$$\begin{array}{ll}
 (a) \ \pi = \# \# \# \_ \# \# \_ \# \# \# & (c) \quad \begin{array}{c} 111\mathbf{h}1011\mathbf{h}\overline{11} \dots \\ \phantom{111\mathbf{h}1011\mathbf{h}} \overset{s_9 \ t}{\# \# \# \_ \# \# \_} \\ \phantom{111\mathbf{h}1011\mathbf{h}} \overset{\pi_{1..7}}{\# \# \# \_} \\ \phantom{111\mathbf{h}1011\mathbf{h}} \overset{\pi_{1..4}}{\# \# \# \_} \\ \phantom{111\mathbf{h}1011\mathbf{h}} \overset{\pi_{1..2}}{\# \#} \end{array} \\
 (b) \ s = 111\mathbf{h}1011\mathbf{h}11\dots &
 \end{array}$$

**Fig. 1.** Illustration to Example 2

*Example 2.* In the framework of Example 1, consider a seed  $\pi$  and an alignment prefix  $s$  of length  $p = 11$  given on Figure 1(a) and 1(b) respectively. The length  $t$  of the last run of 1's of  $s$  is 2. The last mismatch position of  $s$  is  $s_9 = \mathbf{h}$ . The set  $R_\pi$  of non- $\#$  positions of  $\pi$  is  $\{2, 4, 7\}$  and  $\pi$  has 3 prefixes ending at positions of  $R_\pi$  (Figure 1(c)). Prefixes  $\pi_{1..2}$  and  $\pi_{1..7}$  do match suffixes of  $s_1 s_2 \dots s_9$ , and prefix  $\pi_{1..4}$  does not. Thus, the state of the automaton after reading  $s_1 s_2 \dots s_{11}$  is  $\langle \{2, 7\}, 2 \rangle$ .

The initial state  $q_0$  of  $S_\pi$  is the state  $\langle \emptyset, 0 \rangle$ . The final states  $Q_f$  of  $S_\pi$  are all states  $q = \langle X, t \rangle$ , where  $\max\{X\} + t = m$ . All final states are merged into one state.

The transition function  $\psi(q, a)$  is defined as follows: If  $q$  is a final state, then  $\forall a \in \mathcal{A}$ ,  $\psi(q, a) = q$ . If  $q = \langle X, t \rangle$  is a non-final state, then

- if  $a = 1$  then  $\psi(q, a) = \langle X, t + 1 \rangle$ ,
- otherwise  $\psi(q, a) = \langle X_U \cup X_V, 0 \rangle$  with
  - $X_U = \{x \mid x \leq t + 1 \text{ and } a \in \pi_x\}$
  - $X_V = \{x + t + 1 \mid x \in X \text{ and } a \in \pi_{x+t+1}\}$

**Lemma 3.** *The automaton  $S_\pi$  accepts the set of all alignments matched by  $\pi$ .*

*Proof.* It can be verified by induction that the invariant condition on the states  $\langle X, t \rangle \in Q$  is preserved by the transition function  $\psi$ . The final states verify  $\max\{X\} + t = m$ , which implies that  $\pi$  matches a suffix of  $s_1 \dots s_p$ .

**Lemma 4.** *The number of states of the automaton  $S_\pi$  is no more than  $(w+1)2^r$ .*

*Proof.* Assume that  $R_\pi = \{x_1, x_2, \dots, x_r\}$  and  $x_1 < x_2 < \dots < x_r$ . Let  $Q_i$  be the set of non-final states  $\langle X, t \rangle$  with  $\max\{X\} = x_i$ ,  $i \in [1..r]$ . For states  $q = \langle X, t \rangle \in Q_i$  there are  $2^{i-1}$  possible values of  $X$  and  $m - x_i$  possible values of  $t$ , as  $\max\{X\} + t \leq m - 1$ . Thus,  $|Q_i| \leq 2^{i-1}(m - x_i) \leq 2^{i-1}(m - i)$  and  $\sum_{i=1}^r |Q_i| \leq \sum_{i=1}^r 2^{i-1}(m - i) = (m - r + 1)2^r - m - 1$ . Besides states  $Q_i$ ,  $Q$  contains  $m$  states  $\langle \emptyset, t \rangle$  ( $t \in [0..m - 1]$ ) and one final state. Thus,  $|Q| \leq (m - r + 1)2^r = (w + 1)2^r$ .

Note that if  $\pi$  starts with  $\#$ , which is always the case for ordinary spaced seeds, then  $X_i \geq i + 1$ ,  $i \in [1..r]$ , and previous bound rewrites to  $2^{i-1}(m - i - 1)$ . This results in the same number of states  $w2^r$  as for the Aho-Corasick automaton [17]. The construction of automaton  $S_\pi$  is optimal, in the sense that no two states can be merged in general. A straightforward generation of the transition table of the automaton  $S_\pi$  can be performed in time  $\mathcal{O}(r \cdot w \cdot 2^r \cdot |\mathcal{A}|)$ . A more complicated algorithm allows one to reduce the bound to  $\mathcal{O}(w \cdot 2^r \cdot |\mathcal{A}|)$ . In the next section, we demonstrate experimentally that on average, our construction yields a very compact automaton, close to the minimal one. Together with the general approach of section 2, this provides a fast algorithm for computing the sensitivity of subset seeds and, in turn, allows to perform an efficient design of spaced seeds well-adapted to the similarity search problem under interest.

## 4 Experiments

Several types of experiments have been performed to test the practical applicability of the results of sections 2,3. We focused on DNA similarity search, and set the alignment alphabet  $\mathcal{A}$  to  $\{1, \mathbf{h}, 0\}$  (match, transition, transversion). For subset seeds, the seed alphabet  $\mathcal{B}$  was set to  $\{\#, @, -\}$ , where  $\# = \{1\}$ ,  $@ = \{1, \mathbf{h}\}$ ,  $- = \{1, \mathbf{h}, 0\}$  (see Example 1). The weight of a subset seed is computed by assigning weights 1, 0.5 and 0 to symbols  $\#$ ,  $@$  and  $-$  respectively.

#### 4.1 Size of the automaton

We compared the size of the automaton  $S_\pi$  defined in section 3 and the Aho-Corasick automaton [22], both for ordinary spaced seeds (binary seed alphabet) and for subset seeds (ternary seed alphabet). The Aho-Corasick automaton for spaced seeds was constructed as defined in [17]. For subset seeds, a straightforward generalization was considered: the Aho-Corasick construction was applied to the set of alignment fragments matched by the seed.

Tables 1(a) and 1(b) present the results for spaced seeds and subset seeds respectively. For each seed weight  $w$ , we computed the average number of states (*avg. s.*) of the Aho-Corasick automaton and our automaton  $S_\pi$ , and reported the corresponding ratio ( $\delta$ ) with respect to the average number of states of the minimized automaton. The average was computed over all seeds of span up to  $w + 8$  for spaced seeds and all seeds of span up to  $w + 5$  with two @'s for subset seeds. Interestingly, our automaton turns out to be more compact than the Aho-

Spaced $w$	Aho-Corasick		$S_\pi$		Minimized	Subset $w$	Aho-Corasick		$S_\pi$		Minimized
	<i>avg. s.</i>	$\delta$	<i>avg. s.</i>	$\delta$	<i>avg. s.</i>		<i>avg. s.</i>	$\delta$	<i>avg. s.</i>	$\delta$	<i>avg. s.</i>
9	345.94	3.06	146.28	1.29	113.21	9	1900.65	15.97	167.63	1.41	119.00
10	380.90	3.16	155.11	1.29	120.61	10	2103.99	16.50	177.92	1.40	127.49
11	415.37	3.25	163.81	1.28	127.62	11	2306.32	16.96	188.05	1.38	135.95
12	449.47	3.33	172.38	1.28	134.91	12	2507.85	17.42	198.12	1.38	144.00
13	483.27	3.41	180.89	1.28	141.84	13	2709.01	17.78	208.10	1.37	152.29

(a)

(b)

**Table 1.** Comparison of the average number of states of Aho-Corasick automaton, automaton  $S_\pi$  of section 3 and minimized automaton

Corasick automaton not only on non-binary alphabets (which was expected), but also on the binary alphabet (cf Table 1(a)). Note that for a given seed, one can define a surjective mapping from the states of the Aho-Corasick automaton onto the states of our automaton. This implies that our automaton has *always* no more states than the Aho-Corasick automaton.

#### 4.2 Seed Design

In this part, we considered several probability transducers to design spaced or subset seeds. The target alignments included all alignments of length 64 on alphabet  $\{1, h, 0\}$ . Four probability transducers have been studied (analogous to those introduced in [23]):

- *B*: Bernoulli model
- *DT1*: deterministic probability transducer specifying probabilities of  $\{1, h, 0\}$  at each codon position (extension of the  $M^{(3)}$  model of [23] to the three-letter alphabet),
- *DT2*: deterministic probability transducer specifying probabilities of each of the 27 codon instances  $\{1, h, 0\}^3$  (extension of the  $M^{(8)}$  model of [23]),
- *NT*: non-deterministic probability transducer combining four copies of *DT2* specifying four distinct codon conservation levels (called HMM model in [23]).

Models *DT1*, *DT2* and *NT* have been trained on alignments resulting from a pairwise comparison of 40 bacteria genomes. For each of the four probability

$w$	spaced seeds	Sens.	subset seeds, two @	Sens.
9	###_#_#_#_#	0.4183	###_#_#@#_#@#	0.4443
10	##_###_###_#_###	0.2876	###_#@#_#@#_###	0.3077
11	###_###_#_#_###	0.1906	##@#_###_#_#@###	0.2056
12	###_#_###_#_###	0.1375	##@#_#_###_#@#####	0.1481

**Table 2.** *Best seeds and their sensitivity for probability transducer B*

$w$	spaced seeds	Sens.	subset seeds, two @	Sens.
9	###_###_###_#	0.4350	##@_###_###_###@	0.4456
10	##_###_###_###_###	0.3106	##_###_#@#_###@#	0.3173
11	##_###_###_###_###	0.2126	##@#@#_###_###_###	0.2173
12	##_###_###_###_###	0.1418	##_@###_###_###@###	0.1477

**Table 3.** *Best seeds and their sensitivity for probability transducer DT1*

$w$	spaced seeds	Sens.	subset seeds, two @	Sens.
9	_###_###_###_#	0.5121	_#@_###_#@_###_###	0.5323
10	##_###_###_###_###	0.3847	##_#@#_###_#@_###_###	0.4011
11	##_###_#_#_###_###	0.2813	##_###_#@#_#_#@_###_###	0.2931
12	##_###_###_#_###_###	0.1972	##_###_#@#_###_#@_###_###	0.2047

**Table 4.** *Best seeds and their sensitivity for probability transducer DT2*

transducers, we computed the best seed of weight  $w$  ( $w = 9, 10, 11, 12$ ) among two categories: ordinary spaced seeds of weight  $w$  and subset seeds of weight  $w$  with two @. Ordinary spaced seeds were enumerated exhaustively up to a given span, and for each seed, the sensitivity was computed using the algorithmic approach of section 2 and the seed automaton construction of section 3. Each such computation took between 10 and 500ms on a Pentium IV 2.4GHz computer depending on the seed weight/span and the model used. In each experiment, the most sensitive seed found has been kept. The results are presented in Tables 2-5.

In all cases, subset seeds yield a better sensitivity than ordinary spaced seeds. The sensitivity increment varies up to 0.04 which is a notable increase. As shown in [12], the gain in using subset seeds increases substantially when the transition probability is greater than the transversion probability, which is very often the case in related genomes.

$w$	spaced seeds	Sens.	subset seeds, two @	Sens.
9	##_##_#_#_#_#_#	0.5253	##_@#_#_#_#_#	0.5420
10	##_##_#_#_#_#_#	0.4123	##_##_#_#_#_#_#	0.4190
11	##_##_#_#_#_#_#	0.3112	##_##_#_#_#_#_#	0.3219
12	##_##_#_#_#_#_#	0.2349	##_##_#_#_#_#_#	0.2412

**Table 5.** *Best seeds and their sensitivity for probability transducer NT*

## 5 Discussion

We introduced a general framework for computing the seed sensitivity for various similarity search settings. The approach can be seen as a generalization of methods of [17,15] in that it allows to obtain algorithms with the same worst-case complexity bounds as those proposed in these papers, but also allows to obtain efficient algorithms for new formulations of the seed sensitivity problem. This versatility is achieved by distinguishing and treating separately the three ingredients of the seed sensitivity problem: a set of target alignments, an associated probability distributions, and a seed model.

We then studied a new concept of *subset seeds* which represents an interesting compromise between the efficiency of spaced seeds and the flexibility of vector seeds. For this type of seeds, we defined an automaton with  $\mathcal{O}(w2^r)$  states regardless of the size of the alignment alphabet, and showed that its transition table can be constructed in time  $\mathcal{O}(w2^r|\mathcal{A}|)$ . Projected to the case of spaced seeds, this construction gives the same worst-case bound as the Aho-Corasick automaton of [17], but results in a smaller number of states in practice. Different experiments we have done confirm the practical efficiency of the whole method, both at the level of computing sensitivity for designing good seeds, as well as using those seeds for DNA similarity search.

As far as the future work is concerned, it would be interesting to study the design of efficient spaced seeds for protein sequence search (see [10]), as well as to combine spaced seeds with other techniques such as seed families [5,6,8] or the group hit criterion [12].

*Acknowledgments* G. Kucherov and L. Noé have been supported by the *ACI IMPBio* of the French Ministry of Research. M. Roytberg has been also supported by the Russian Foundation for Basic Research (projects 03-04-49469, 02-07-90412) and by grants from the RF Ministry of Industry, Science and Technology (20/2002, 5/2003) and NWO (Netherlands Science Foundation).

## References

1. Burkhardt, S., Kärkkäinen, J.: Better filtering with gapped  $q$ -grams. *Fundamenta Informaticae* **56** (2003) 51–70 Preliminary version in *Combinatorial Pattern Matching* 2001.
2. Ma, B., Tromp, J., Li, M.: PatternHunter: Faster and more sensitive homology search. *Bioinformatics* **18** (2002) 440–445

3. Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25** (1997) 3389–3402
4. Kent, W.J.: BLAT—the BLAST-like alignment tool. *Genome Research* **12** (2002) 656–664
5. Li, M., Ma, B., Kisman, D., Tromp, J.: PatternHunter II: Highly sensitive and fast homology search. *Journal of Bioinformatics and Computational Biology* (2004) Earlier version in GIW 2003 (International Conference on Genome Informatics).
6. Sun, Y., Buhler, J.: Designing multiple simultaneous seeds for DNA similarity search. In: *Proceedings of the 8th Annual International Conference on Computational Molecular Biology*, ACM Press (2004)
7. Yang, I.H., Wang, S.H., Chen, Y.H., Huang, P.H., Ye, L., Huang, X., Chao, K.M.: Efficient methods for generating optimal single and multiple spaced seeds. In: *Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering*, IEEE Computer Society Press (2004) 411–416
8. Kucherov, G., Noé, L., Roytberg, M.: Multiseed lossless filtration. *IEEE Transactions on Computational Biology and Bioinformatics* **2** (2005) 51 – 61
9. Xu, J., Brown, D., Li, M., Ma, B.: Optimizing multiple spaced seeds for homology search. In: *Proceedings of the 15th Symposium on Combinatorial Pattern Matching*. Volume 3109 of *Lecture Notes in Computer Science.*, Springer (2004)
10. Brown, D.: Optimizing multiple seeds for protein homology search. *IEEE Transactions on Computational Biology and Bioinformatics* **2** (2005) 29 – 38
11. Chen, W., Sung, W.K.: On half gapped seed. *Genome Informatics* **14** (2003) 176–185 preliminary version in the 14th International Conference on Genome Informatics (GIW).
12. Noé, L., Kucherov, G.: Improved hit criteria for DNA local alignment. *BMC Bioinformatics* **5** (2004)
13. Brejova, B., Brown, D., Vinar, T.: Vector seeds: an extension to spaced seeds allows substantial improvements in sensitivity and specificity. In Benson, G., Page, R., eds.: *Proceedings of the 3rd International Workshop in Algorithms in Bioinformatics*. Volume 2812 of *Lecture Notes in Computer Science.*, Springer (2003)
14. Keich, U., Li, M., Ma, B., Tromp, J.: On spaced seeds for similarity search. to appear in *Discrete Applied Mathematics* (2002)
15. Brejova, B., Brown, D., Vinar, T.: Optimal spaced seeds for homologous coding regions. *Journal of Bioinformatics and Computational Biology* **1** (2004) 595–610
16. Choi, K., Zhang, L.: Sensitivity analysis and efficient method for identifying optimal spaced seeds. *Journal of Computer and System Sciences* **68** (2004) 22–40
17. Buhler, J., Keich, U., Sun, Y.: Designing seeds for similarity search in genomic DNA. In: *Proceedings of the 7th Annual International Conference on Computational Molecular Biology*, ACM Press (2003) 67–75
18. Choi, K.P., Zeng, F., Zhang, L.: Good Spaced Seeds For Homology Search. *Bioinformatics* **20** (2004) 1053–1059
19. Kucherov, G., Noé, L., Ponty, Y.: Estimating seed sensitivity on homogeneous alignments. In: *Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering*, IEEE Computer Society Press (2004) 387–394
20. Ullman, J.D., Aho, A.V., Hopcroft, J.E.: *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading (1974)
21. Finkelstein, A., Roytberg, M.: Computation of biopolymers: A general approach to different problems. *BioSystems* **30** (1993) 1–19
22. Aho, A.V., Corasick, M.J.: Efficient string matching: An aid to bibliographic search. *Communications of the ACM* **18** (1975) 333–340

23. Brejova, B., Brown, D., Vinar, T.: Optimal spaced seeds for Hidden Markov Models, with application to homologous coding regions. In R. Baeza-Yates, E. Chavez, M.C., ed.: Proceedings of the 14th Symposium on Combinatorial Pattern Matching. Volume 2676 of Lecture Notes in Computer Science., Springer (2003) 42–54