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

Gregory Kucherov, Laurent Noé, Mikhail Roytberg

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

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

Gregory Kucherov — Laurent Noé — Mikhail Roytberg

N° 5374

Novembre 2004

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--5374--FR+ENG

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

Gregory Kucherov , Laurent Noé , Mikhail Roytberg

Thème BIO —Systèmes biologiques
Projet Adage

Rapport de recherche n° 5374 —Novembre 2004 —21 pages

Résumé : Nous proposons une approche générale pour calculer la sensibilité des graines espacées, qui peut être appliquée à différents modèles de graines. Cette approche considère séparément trois composants – l'ensemble des alignements cibles, la distribution de probabilité associée, et le modèle de graines – qui sont respectivement spécifiés par des automates finis. L'approche est expérimentée sur un nouveau modèle de graines de type *subset seeds* pour lequel nous proposons une construction efficace de l'automate. Les résultats expérimentaux montrent que des graines efficaces peuvent être conçues sur ce modèle, et donnent de meilleurs résultats que des graines espacées ordinaires en pratique.

Mots-clés : alignement local, ADN , graines espacées, graines sous-ensemble, graines à transitions, automates finis, sensibilité de la graine, programmation dynamique

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 —Télécopie : +33 3 83 27 83 19

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

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.

Key-words: local alignment, DNA , spaced seeds, subset seeds, transition-constrained seeds, finite automata, seed sensitivity, dynamic programming

Table des matières

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 [8, 18] 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 [2, 14, 18], seed families [17, 20, 23, 16, 22, 6], seeds over non-binary alphabets [9, 19], vector seeds [4, 6].

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 [18, 13, 5, 10, 7, 11]. The binary representation, however, cannot distinguish between different types of matches and mismatches, and is clearly insufficient in the case of protein sequences. In [4, 6], 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 [19], 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 [15], 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 [18, 10]. In more general settings, Markov or Hidden Markov models are considered [7, 5]. A different way of defining probabilities on binary alignments has been taken in [15] : 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 [13].

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 [7] 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 [21, 12] 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 [7]. Moreover, our DFA has always no more states than the DFA of [7], 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 [19]). The importance of this extension is even more evident for the protein case ([6]), 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 [18, 13, 5, 10, 7, 11], 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. [15]) 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 $T =< Q_T, q_T^0, q_T^F, \mathcal{A}, \psi_T >$.

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 $< Q_G, q_G^0, \mathcal{A}, \rho_G >$, where Q_G is a finite set of states, $q_G^0 \in Q_G$ is an initial state, and $\rho_G : Q_G \times \mathcal{A} \times Q_G \to [0,1]$ 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 $e =< q, a, q' >$ such that $\rho(q,a,q') > 0$. Letter a is called the *label* of e and denoted $label(e)$. 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 $< q, a, q' >$. For each path $P = (e_1, ..., e_n)$ in G, we define its *label* to be the word $label(P) = label(e_1)...label(e_n)$, 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 $G =< Q_G, q_G^0, \mathcal{A}, \rho_G >$, 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 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 [18, 10, 11] can be specified with deterministic one-state probability transducers. In Markov sequences of order k [7, 20], 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) [5] 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 [8, 18], 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. [7] 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) π under a given seed model. We assume that the set of alignments L_π matched by π is a regular language recognized by a DFA $S_\pi =< Q_S, q_S^0, Q_S^F, \mathcal{A}, \psi_S >$. Consider a finite set L_T of target alignments and a probability transducer G. Under this assumptions, the sensitivity of π is defined as the conditional probability

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

An automaton recognizing $L = L_T \cap L_\pi$ can be obtained as the product of automata T and S_π recognizing L_T and L_π respectively. Let $K =< Q_K, q_K^0, Q_K^F, \mathcal{A}, \psi_K >$ 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 $K =< Q_K, q_K^0, Q_K^F, \mathcal{A}, \psi_K >$ and a probability transducer $G =< Q_G, q_G^0, \mathcal{A}, \rho_G >$, the product of K and G is the *probability-weighted automaton* (for short, *PW-automaton*) $W =< Q_W, q_W^0, Q_W^F, \mathcal{A}, \rho_W >$ such that

- $Q_W = Q_K \times Q_G$,
- $q_W^0 = (q_K^0, q_G^0)$,
- $q_W^F = \{(q_K, q_G) | q_K \in Q_K^F\}$,
- $\rho_W((q_K, q_G), a, (q'_K, q'_G)) = \begin{cases} \rho_G(q_G, a, q'_G) & \text{if } \psi_K(q_K, a) = q'_K, \\ 0 & \text{otherwise.} \end{cases}$

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 transition $< (q_K, q_G), a, (q'_K, q'_G) >$. A path in W is called *full* if it goes from the initial to a final state.

Lemma 4. *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.*

Démonstration. 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 4 reduces the computation of seed sensitivity to a computation of the sum of probabilities of paths in a PW-automaton.

Lemma 5. *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 π. Let $K =< Q_K, q_t^0, Q_K^F, \mathcal{A}, \psi_Q >$ be an acyclic DFA recognizing the language $L = L_T \cap L_\pi$. Let further $G =< Q_G, q_G^0, \mathcal{A}, \rho >$ be a probability transducer defining a probability distribution on the set L_T. Then $\mathcal{P}_G(L)$ can be computed in time*

$$\mathcal{O}(|Q_G|^2 \cdot |Q_K| \cdot |\mathcal{A}|) \tag{2}$$

and space

$$\mathcal{O}(|Q_G| \cdot |Q_K|). \tag{3}$$

Démonstration. By Lemma 4, 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 [21] applied to acyclic directed graphs ([12] 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 5 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_π specifying the seed model via a set of matched alignments.

As soon as these three automata are defined, Lemma 5 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 (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 (2) 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. [7] introduced the idea of using the Aho-Corasick automaton [1] as the seed automaton S_π for a spaced seed. The authors of [7] 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 $\mathcal{O}(w2^{s-w}2^k n)$, 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 5, 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 [7], 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 [7]). 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. [5]. The DP algorithm

of [5] has a lot in common with the algorithm implied by Lemma 5. In particular, the states of the algorithm of [5] are triples $< w, q, m >$, where w is a prefix of the seed π, q is a state of the HMM, and $m \in [0..n]$. The states therefore correspond to the construction implied by Lemma 5. However, the authors of [5] 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 [4]. 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 5), 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` ($\mathcal{B} \in \{\mathtt{1}\} \cup 2^{\mathcal{A}\setminus\{\mathtt{1}\}}$),
- $\mathcal{B}$ contains a letter $\#$ that denotes subset $\{\mathtt{1}\}$,
- 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 π is the number of $\#$ in π and the *span* of π is its length.

Example 1. [19] considered the alignment alphabet $\mathcal{A} = \{1, h, 0\}$ 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 $\{1\}$, $\{1, h\}$, and $\{1, h, 0\}$. Thus, seed $\pi = \#@\_\#$ matches alignment $s = 10h1h1101$ at positions 4 and 6. The span of π is 4, and the $\#$-weight of π 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 π is equal to 2.5 (see [19]).

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_π be the set of all non-$\#$ positions in π, $|R_\pi| = r = m - w$. We now define an automaton $S_\pi =< Q, q_0, Q_f, \mathcal{A}, \psi : Q \times \mathcal{A} \to Q >$ that recognizes the set of all alignments matched by π.

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

FIG. 1 – Illustration to Example 2

Example 2. In the framework of Example 1, consider a seed π and an alignment prefix s of length $n = 11$ given on Figure 1(a) and (b) respectively. The length t of the last run of 1's of s is 2. The last mismatch position of s is $s_9 = h$. The set R_π of non-$\#$ positions of π is $\{2, 4, 7\}$ and π has 3 prefixes ending at positions of R_π (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 $< \{2, 7\}, 2 >$.

The initial state q_0 of S_π is the state $< \emptyset, 0 >$. The final states Q_f of S_π are all states $q =< X, t >$, 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 $q =< X, t >$ is a non-final state, then

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

Lemma 6. *The automaton S_π accepts the set of all alignments matched by π.*

Démonstration. It can be verified by induction that the invariant condition on the states $< X, t > \in Q$ is preserved by the transition function ψ. The final states verify $max\{X\} + t = m$, which implies that π matches a suffix of $s_1 \dots s_n$. □

Lemma 7. *The number of states of the automaton S_π is no more than $(w+1)2^r$.*

Démonstration. Assume that $R_\pi = \{x_1, x_2, \dots, x_r\}$ and $x_1 < x_2 \cdots < x_r$. Let Q_i be the set of non-final states $< X, t >$ with $max\{X\} = x_i$, $i \in [1..r]$. For states $q =< X, t > \in Q_i$ 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,

$$|Q_i| \;\leq\; 2^{i-1}(m - x_i) \quad\leq\quad 2^{i-1}(m-i), \text{ and} \tag{4}$$

$$\sum_{i=1}^{r} |Q_i| \;\leq\; \sum_{i=1}^{r} 2^{i-1}(m-i) \quad=\quad (m-r+1)2^r - m - 1. \tag{5}$$

Besides states Q_i, Q contains m states $< \emptyset, t >$ ($t \in [0..m-1]$) and one final state. Thus, $|Q| \leq (m-r+1)2^r = (w+1)2^r$. □

Note that if π starts with $\#$, which is always the case for ordinary spaced seeds, then $X_i \geq i+1$, $i \in [1..r]$, and the bound of (4) rewrites to $2^{i-1}(m-i-1)$. This results in the same number of states $w2^r$ as for the Aho-Corasick automaton [7]. The construction of automaton S_π is optimal, in the sense that no two states can be merged in general, as the following Lemma states.

Lemma 8. *Consider a spaced seed π which consists of two "must-match" symbols $\#$ separated by r jokers. Then the automaton S_π is reduced, that is any non-final state is reachable from the initial state q_0, and any two non-final states q, q' are non-equivalent.*

Démonstration. See appendix A. □

A straightforward generation of the transition table of the automaton S_π 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 $\mathcal{O}(w \cdot 2^r \cdot |\mathcal{A}|)$. This algorithm is described in full details in Appendix B. Here we summarize it in the following Lemma.

Lemma 9. *The transition table of automaton S_π can be constructed in time proportional to its size, which is $\mathcal{O}(w \cdot 2^r \cdot |\mathcal{A}|)$.*

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 $\{1, h, 0\}$ (match, transition, transversion). For subset seeds, the seed alphabet $\mathcal{B}$ was set to $\{\#, @, \_\}$, where $\# = \{1\}, @ = \{1, h\}, \_ = \{1, h, 0\}$ (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_π defined in section 3 and the Aho-Corasick automaton [1], both for ordinary spaced seeds (binary seed alphabet) and for subset seeds. The Aho-Corasick automaton for spaced seeds was constructed as defined in [7]. 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, we computed the average number of states ($avg.\#$) of the Aho-Corasick automaton AC_π and our automaton S_π, and reported the corresponding ratio (δ) with respect to the average number of states of the minimized automaton. The average was computed over all seeds of span $w + 8$ for spaced seeds and all seeds of span $w + 5$ with two $@$ elements for subset seeds. Interestingly, our automaton turns out to be more compact than the

(a) Spaced-seeds

	AC_π		S_π		Minimized
w	$avg.\#$	δ	$avg.\#$	δ	$avg.\#$
9	345.94	3.06	146.28	1.29	113.21
10	380.90	3.16	155.11	1.29	120.61
11	415.37	3.25	163.81	1.28	127.62
12	449.47	3.33	172.38	1.28	134.91
13	483,27	3.41	180.89	1.28	141.84

(b) Subset-seeds

	AC_π		S_π		Minimized
w	$avg.\#$	δ	$avg.\#$	δ	$avg.\#$
9	1900.65	15.97	167.63	1.41	119,00
10	2103.99	16.50	177.92	1.40	127.49
11	2306.32	16.96	188.05	1.38	135.95
12	2507.85	17.42	198.12	1.38	144.00
13	2709.01	17.78	208.10	1.37	152.29

TAB. 1 – *Comparison of the average number of states of Aho-Corasick automaton, automaton S_π of section 3 and minimized automaton*

Aho-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 has been studied (analogous to those introduced in [3]) :

- B : Bernoulli model with probabilities $Pr(1) = 0.7, Pr(h) = Pr(0) = 0.15$,
- $DT1$: deterministic probability transducer specifying probabilities of $\{1, h, 0\}$ at each codon position (extension of the $M^{(3)}$ model of [3] 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 [3] to the three-letter alphabet),
– NT : non-deterministic probability transducer combining four copies of $DT2$ specifying four distinct codon conservation levels (called HMM model in [3]).

Models $DT1$, $DT2$ and NT have been trained on alignments resulting from a pairwise comparison of 40 bacteria genomes. Details of the training procedure as well as the resulting parameter values are given in Appendix C.

For each of the four probability transducers, we computed the best seed of weight w ($w = 9, 10, 11, 12$) among three categories : ordinary spaced seeds of weight w, subset seeds of weight w with two @, subset seeds of weight w with four @. All seeds were processed exhaustively, 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 50ms on a Pentium IV 2.4GHz computer. In each experiment, the most sensitive seed has been kept. The results are presented in Tables 2-5.

w	spaced seeds	Sens.	subset seeds, two @	Sens.	subset seeds, four @	Sens.
9	###___#_#_##_##	0.7292	###@_#_#__#@_##	0.7375	###_@_@#__#@_#@#	0.7381
10	##_##___##_#_###	0.5957	##_#__##_@_#_@###	0.6042	#@#_#@_#_@#__@###	0.6069
11	###_#__#_#__##_###	0.4671	###_@#__#_@#_#_###	0.4761	##@@#__#@_#_#_@###	0.4802
12	###_#_##_#__##_###	0.3564	###_#@_##__#_#_@###	0.3637	##@#_@_#_#@_##_@###	0.3669

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

w	spaced seeds	Sens.	subset seeds, two @	Sens.	subset seeds, four @	Sens.
9	##_##__#_#__###	0.4696	##_@#__#_#_@###	0.4696	##@_#_#@__#_@#@#	0.4655
10	###_#_##__#__###	0.3305	###_#__#_@#@_###	0.3329	##@#__@#_#_@#_@##	0.3316
11	###_#__#__##_#_###	0.2262	###_#__#@_#_##_@##	0.2283	##@#_@@#_#__#@_###	0.2268
12	###_#_##_#__##_###	0.1511	###@#__##_#_@#_###	0.1521	##@#@_##_#_@#@_###	0.1513

TAB. 3 – *Best seeds and their sensitivity for probability transducer $DT1$*

w	spaced seeds	Sens.	subset seeds, two @	Sens.	subset seeds, four @	Sens.
9	###____##_##_##	0.4961	##_##_____##_@#@#	0.5011	##_##____@#@_@#@#	0.4977
10	##_##_____##_##_##	0.3589	##@#@_##_____##_##	0.3650	##@#@_##____@#@_##	0.3648
11	##_###____##_##_##	0.2487	##@##____@##_##_##	0.2540	##@#@_##@____##_##@	0.2530
12	##_##_____##_##_####	0.1658	##_##_##@____##_@###	0.1705	##@#@_##@#__##_@##	0.1709

TAB. 4 – *Best seeds and their sensitivity for probability transducer $DT2$*

w	spaced seeds	Sens.	subset seeds, two @	Sens.	subset seeds, four @	Sens.
9	###___##_##_##	0.4397	##@___##_##_##@	0.4460	##@_@_##_##_@#@	0.4442
10	##_##_____##_##_##	0.3145	##_##____@##_##@#	0.3172	##_@#@#@_##_@##	0.3156
11	##_##_____##_##_###	0.2162	##@#@_##_##__###	0.2181	##@#@_#@_##_@###	0.2186
12	##_##_____##_##_####	0.1446	##_@###__##_##@##	0.1485	##@#@_##_##@@###	0.1469

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

In all cases, subset seeds yield a better sensitivity than ordinary spaced seeds. The sensitivity increment varies up to 0.013 which is a notable increase. It is important to note that the parameters of the transducers are such that the probability of a transition (h) is generally slightly smaller than the probability of a transversion (0) (see Appendix C). As shown in [19], the gain in using subset seeds increases substantially when the transition probability is greater than the inversion probability, which is very often the case in related genomes.

4.3 Comparative performance of spaced and subset seeds

We performed a series of whole genome comparisons in order to compare the performance of designed spaced and subset seeds. Eight complete bacterial genomes[1] have been processed against each other using the YASS software [19]. Each comparison was done twice : with a spaced seed and a subset seed of the same weight.

The threshold E-value for the output alignments was set to 10, and for each comparison, the number of alignments with E-value smaller than 10^{-3} found by each seed, and the number of exclusive alignments were reported. By "exclusive alignment" we mean any alignment of E-value less than 10^{-3} that does not share a common part (do not overlap in both compared sequences) with any alignment found by another seed. To take into account a possible bias caused by splitting alignments into smaller ones (X-drop effect), we also computed the total length of exclusive alignments. Table 6 summarizes these experiments for weights 9,10 and the $DT2$ and NT probabilistic models. Each line corresponds to a seed given in Table 4 or Table 5, depending on the indicated probabilistic model. The table implies that best subset seeds detect 2 to 5 percent more of significant alignments

$seed$	$time$	$\#align$	$\#ex.align$	$ex.\ align\ length$
$DT2$, $w = 9$, spaced seed	12 :45	21060	1452	158667
$DT2$, $w = 9$, subset seed, two @	13 :05	21590	1772	257123
$DT2$, $w = 10$, spaced seed	8 :36	19787	1027	131054
$DT2$, $w = 10$, subset seed, two @	9 :02	21586	1421	217362
NT, $w = 9$, spaced seed	38 :10	20625	1057	134095
NT, $w = 9$, subset seed, two @	36 :13	21385	1910	270655
NT, $w = 10$, spaced seed	13 :09	19867	1008	91319
NT, $w = 10$, subset seed, two @	12 :53	21590	1259	215904

TAB. 6 – *Comparative test of subset seeds vs spaced seeds. Reported execution times (min :sec) were obtained on a Pentium IV 2.4GHz computer.*

missed by best spaced seeds of the same weight.

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 [7, 5] in that it allows to obtain algorithms with the same worst-case complexity bounds as those proposed in these papers,

[1] NC_000907.fna, NC_002662.fna, NC_003317.fna, NC_003454.fna, NC_004113.fna, NC_001263.fna, NC_003112.fna obtained from NCBI

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 [7], 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 [6]), as well as to combine spaced seeds with other techniques such as seed families [17, 20, 16] or the group hit criterion [19].

Acknowledgements G. Kucherov and L. Noé have been supported by the *ACI IMPBio* of the French Ministry of Research. A part of this work has been done during a stay of M. Roytberg at LORIA, Nancy, supported by INRIA. 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).

Références

[1] AHO, A. V., AND CORASICK, M. J. Effi cient string matching : An aid to bibliographic search. *Communications of the ACM 18*, 6 (1975), 333–340.

[2] ALTSCHUL, S., MADDEN, T., SCHÄFFER, A., ZHANG, J., ZHANG, Z., MILLER, W., AND LIPMAN, D. Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Research 25*, 17 (1997), 3389–3402.

[3] BREJOVA, B., BROWN, D., AND VINAR, T. Optimal spaced seeds for Hidden Markov Models, with application to homologous coding regions. In *Proceedings of the 14th Symposium on Combinatorial Pattern Matching, Morelia (Mexico)* (June 2003), M. C. R. Baeza-Yates, E. Chavez, Ed., vol. 2676 of *Lecture Notes in Computer Science*, Springer, pp. 42–54.

[4] BREJOVA, B., BROWN, D., AND VINAR, T. Vector seeds : an extension to spaced seeds allows substantial improvements in sensitivity and specifi city. In *Proceedings of the 3rd International Workshop in Algorithms in Bioinformatics (WABI), Budapest (Hungary)* (September 2003), G. Benson and R. Page, Eds., vol. 2812 of *Lecture Notes in Computer Science*, Springer.

[5] BREJOVA, B., BROWN, D., AND VINAR, T. Optimal spaced seeds for homologous coding regions. *Journal of Bioinformatics and Computational Biology 1*, 4 (Jan 2004), 595–610.

[6] BROWN, D. Multiple vector seeds for protein alignment. In *Proceedings of the 4th International Workshop in Algorithms in Bioinformatics (WABI), Bergen (Norway)* (September 2004), I. Jonassen and J. Kim, Eds., Lecture Notes in Computer Science, Springer. to appear.

[7] BUHLER, J., KEICH, U., AND SUN, Y. Designing seeds for similarity search in genomic DNA. In *Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB03), Berlin (Germany)* (April 2003), ACM Press, pp. 67–75.

[8] BURKHARDT, S., AND KÄRKKÄINEN, J. Better filtering with gapped q-grams. *Fundamenta Informaticae 56*, 1-2 (2003), 51–70. Preliminary version in Combinatorial Pattern Matching 2001.

[9] CHEN, W., AND SUNG, W.-K. On half gapped seed. *Genome Informatics 14* (2003), 176–185. preliminary version in the 14th International Conference on Genome Informatics (GIW).

[10] CHOI, K., AND ZHANG, L. Sensitivity analysis and efficient method for identifying optimal spaced seeds. *Journal of Computer and System Sciences 68* (2004), 22–40.

[11] CHOI, K. P., ZENG, F., AND ZHANG, L. Good Spaced Seeds For Homology Search. *Bioinformatics 20* (2004), 1053–1059.

[12] FINKELSTEIN, A., AND ROYTBERG, M. Computation of biopolymers : A general approach to different problems. *BioSystems 30*, 1-3 (1993), 1–19.

[13] KEICH, U., LI, M., MA, B., AND TROMP, J. On spaced seeds for similarity search. to appear in Discrete Applied Mathematics, 2002.

[14] KENT, W. J. BLAT–the BLAST-like alignment tool. *Genome Research 12* (2002), 656–664.

[15] KUCHEROV, G., NOÉ, L., AND PONTY, Y. Estimating seed sensitivity on homogeneous alignments. In *Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE 2004), May 19-21, 2004, Taichung (Taiwan)* (2004), IEEE Computer Society Press, pp. 387–394.

[16] KUCHEROV, G., NOÉ, L., AND ROYTBERG, M. Multi-seed lossless filtration. In *Proceedings of the 15th Annual Combinatorial Pattern Matching Symposium (CPM), Istanbul (Turkey)* (July 2004), vol. 3109 of *Lecture Notes in Computer Science*, Springer Verlag. to appear.

[17] LI, M., MA, B., KISMAN, D., AND 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).

[18] MA, B., TROMP, J., AND LI, M. PatternHunter : Faster and more sensitive homology search. *Bioinformatics 18*, 3 (2002), 440–445.

[19] NOÉ, L.AND KUCHEROV, G. Improved hit criteria for DNA local alignment. *BMC Bioinformatics 5*, 149 (14 October 2004).

[20] SUN, Y., AND BUHLER, J. Designing multiple simultaneous seeds for DNA similarity search. In *Proceedings of the 8th Annual International Conference on Computational Molecular Biology (RECOMB04), San Diego (California)* (March 2004), ACM Press.

[21] ULLMAN, J. D., AHO, A. V., AND HOPCROFT, J. E. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, 1974.

[22] XU, J., BROWN, D., LI, M., AND MA, B. Optimizing multiple spaced seeds for homology search. In *Proceedings of the 15th Symposium on Combinatorial Pattern Matching, Istambul (Turkey)* (July 2004), vol. 3109 of *Lecture Notes in Computer Science*, Springer.

[23] YANG, I.-H., WANG, S.-H., CHEN, Y.-H., HUANG, P.-H., YE, L., HUANG, X., AND CHAO, K.-M. Efficient methods for generating optimal single and multiple spaced seeds. In *Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE 2004), May 19-21, 2004, Taichung (Taiwan)* (2004), IEEE Computer Society Press, pp. 411–416.

A Proof of Lemma 8

Let $\pi = \#\ -^r\ \#$ be a spaced seed of span $r+2$ and weight 2. We prove that the automaton S_π (see Lemma 6) is reduced, i.e.

(i) all its non-final states are reachable from the initial state $<\emptyset, 0>$;

(ii) any two non-final states q, q' are non-equivalent, i.e. there is a word $w = w(q,q')$ such that exactly one of the states $\psi(q,w), \psi(q',w)$ is a final state.

(i) Let $q =< X, t >$ be a state of the automaton S_π, and let $X = \{x_1, \ldots, x_k\}$ and $x_1 < \cdots < x_k$. Obviously, $x_k + t < r+2$. Let $s \in \{0,1\}^*$ be an alignment word of length x_k such that for all $i \in [1, x_k]$, $s_i = 1$ iff $\exists j \in [1,k], \quad i = x_k - x_j + 1$. Note, that, $\pi_1 = \#$, therefore $1 \notin X$ and $s_{x_k} = 0$. Finally, $\psi(<\phi, 0>, s \cdot 1^t) = q$.

(ii) Let $q_1 =< X_1, t_1 >$ and $q_2 =< X_2, t_2 >$ be non-final states of S_π. Let $X_1 = \{y_1, \ldots, y_a\}$, $X_2 = \{z_1, \ldots, z_b\}$, and $y_1 < \cdots < y_a$, $z_1 < \cdots < z_b$.

Assume that $max\{X_1\}+t_1 > max\{X_2\}+t_2$ and let $d = (r+2)-(max\{X_1\}+t_1)$. Obviously, $\psi(q_1, 1^d)$ is a final state, and $\psi(q_2, 1^d)$ is not. Now assume that $max\{X_1\} + t_1 = max\{X_2\} + t_2$. For a set $X \subseteq \{1, \ldots, r+1\}$ and a number t, define a set $X\{t\}$ by $X\{t\} = \{v+t | v \in X$ and $v+t < r+2\}$.

Let $g = max\{v | (v + t_1 \in X_1$ and $v + t_2 \notin X_2)$ or $(v + t_2 \in X_2$ and $v + t_1 \notin X_1)\}$ and let $d = r+1-g$. Then $\psi(q_1, 0^d \cdot 1)$ is a final state and $\psi(q_2, 0^d \cdot 1)$ is not or vice versa. This completes the proof.

B Subset seed automaton

Let π be a subset seed of #-weight w and span s, and $r = s - w$ be the number of non-# positions. We define a DFA S_π recognizing all words of $\mathcal{A}^*$ matched by π (see definition of section 3.1). The transition table of S_π is stored in an array such that each element describes a state $< X, t >$ of S_π. Now we define

1. how to compute the array index $Ind(q)$ of a state $q =< X, t >$,
2. how to compute values $\psi(q, a)$ given a state q and a letter $a \in \mathcal{A}$.

B.1 Encoding state indexes

We will need some notation. Let $L = \{l_1, \ldots, l_r\}$ be a set of all non-# positions in π ($l_1 < l_2 < \cdots < l_r$). For a subset $X \subseteq L$, let $v(X) = v_1 \ldots v_r \in \{0,1\}^r$ be a binary vector such that $v_i = 1$ iff $l_i \in X$. Let further $n(X)$ be the integer corresponding to the binary representation $v(X)$ (read from left to right) :

$$n(X) = \sum_{j=1}^{r} 2^{j-1} \cdot v_j.$$

Define $p(t) = max\{p \mid l_p < m - t\}$. Informally, for a given non-final state $< X, t >$, X can only be a subset of $\{l_1, \ldots, l_{p(t)}\}$. This implies that $n(X) < 2^{p(t)}$. Then, the index of a given state

$\{< X, t >\}$ in the array is defined by

$$Ind(< X, t >) = n(X) + 2^{p(t)}.$$

This implies that the worst-case size of the array is no more than $w2^r$ (the proof is similar to the proof of Lemma 7).

B.2 Computing transition function $\psi(q, a)$

We compute values $\psi(< X, t >, a)$ based on already computed values $\psi(< X', t >, a)$. Let $q =< X, t >$ be a non-final and reachable state of S_π, where $X = \{l_1, \ldots, l_k\}$ with $l_1 < l_2 \cdots < l_k$ and $k \leq r$. Let $X' = X \setminus \{l_k\} = \{l_1, \ldots, l_{k-1}\}$ and $q' =< X', t >$. Then the following lemma holds.

Lemma 10. *If $q =< X, t >$ is a reachable state, then $q' =< X', t >$ is reachable and has been processed before.*

Démonstration. First prove that $< X', t >$ is reachable. If $< X, t >$ is reachable, then $< X, 0 >$ is reachable due to the definition of transition function for $t > 0$. Thus, one can find at least one sequence $S \in \mathcal{A}^{l_k}$ such that $\forall i \in [1..r]$, $l_i \in X$ iff $\pi_1 \cdots \pi_{l_i}$ matches $S_{l_k - l_i + 1} \cdots S_{l_k}$. For such a sequence S, one can find a word $S' = S_{l_k - l_{k-1} + 1} \cdots S_{l_k}$ which reaches state $< X', 0 >$. To conclude, if there exists a word $S \cdot 1^t$ that reaches the state $< X, t >$, there also exists a word $S' \cdot 1^t$ that reaches $< X', t >$.

Note that as $|S' \cdot 1^t| < |S \cdot 1^t|$, then a breadth-first computation of states of S_π always processes state $< X', t >$ before $< X, t >$. □

Now we present how to compute values $\psi(< X, t >, a)$ from values $\psi(< X', t >, a)$. This is done by Algorithm B.2 shown below, that we comment on now.

Due to implementation choices, we represent a state q as triple $q = \langle X, k_X, t \rangle$, where $k_X = max\{i | l_i \in X\}$. Note first that if $a = 1$, the transition function $\psi(q, a)$ can be computed in constant time due to its definition (part a. of Algorithm B.2). If $a \neq 1$, we have to

1. retrieve the index of q' given $q = \langle X, k_X, t \rangle$ (part c. of Algorithm B.2),
2. compute $\psi(\langle X, k_X, t \rangle, a \neq 1')$ given $\psi(\langle X', k_{X'}, t \rangle, a \neq 1')$ value. (part d. of Algorithm B.2)

1. Note first that $Ind(\langle X, k_X, t \rangle) = Ind(\langle X', k_{X'}, t \rangle) - 2^{k_X}$, which can be computed in constant time since k_X is explicitly stored in the current state.

2. Let

$$V_X(k, t, a \neq 1) \quad = \quad \begin{cases} l_i & \text{if } l_i = l_k + t + 1 \text{ and } a \text{ matches } \pi_{l_i} \\ \emptyset & \text{otherwise} \end{cases}$$

and

$$V_k(k,t,a \neq \mathtt{1}) \quad = \quad \begin{cases} i & \text{if } l_i = l_k + t + 1 \text{ and } \ a \text{ matches } \pi_{l_i} \\ 0 & \text{otherwise} \end{cases}$$

Tables $V_X(k,t,a)$ and $V_k(k,t,a)$ can be precomputed in time and space $\mathcal{O}(|\mathcal{A}| \cdot m^2)$
Let $\psi(\langle X, k_X, t\rangle, a) = \langle Y, k_Y, 0\rangle$ and $\psi(\langle X', k_{X'}, t\rangle, a) = \langle Y', k_{Y'}, 0\rangle$. The set Y differs from Y' at most with one element. This element can be computed in constant time using tables V_X, V_k. Namely $Y = Y' \cup V_X(k_X, t, a)$ and $k_Y = max(k_{Y'}, V_k(k_X, t, a))$.

Note that a final situation arises when $X = \emptyset$. (part b. of Algorithm B.2). One also has to compute two tables U_X, U_k defined as :

$$\begin{aligned} U_X(t, a \neq \mathtt{1}) &= \cup\{x | x \leq t+1 \text{ and } a \text{ matches } \pi_x\} \\ U_k(t, a \neq \mathtt{1}) &= max\{x | x \leq t+1 \text{ and } a \text{ matches } \pi_x\} \end{aligned}$$

Lemma 11. *The transition function $\psi(q, a)$ can be computed in constant time for every reachable state q and every $a \in \mathcal{A}$.*

C Training probability transducers

We selected 40 bacterial complete genomes from NCBI :

NC_000117.fna, NC_000907.fna, NC_000909.fna, NC_000922.fna, NC_000962.fna, NC_001263.fna, NC_001318.fna, NC_002162.fna, NC_002488.fna, NC_002505.fna, NC_002516.fna, NC_002662.fna, NC_002678.fna, NC_002696.fna, NC_002737.fna, NC_002927.fna, NC_003037.fna, NC_003062.fna, NC_003112.fna, NC_003210.fna, NC_003295.fna, NC_003317.fna, NC_003454.fna, NC_003551.fna, NC_003869.fna, NC_003995.fna, NC_004113.fna, NC_004307.fna, NC_004342.fna, NC_004551.fna, NC_004631.fna, NC_004668.fna, NC_004757.fna, NC_005027.fna, NC_005061.fna, NC_005085.fna, NC_005125.fna, NC_005213.fna, NC_005303.fna, NC_005363.fna

YASS [19] has been run on each pair of genomes to detect alignments with E-value at most 10^{-3}. Resulting ungapped regions of length 64 or more have been used to train models $DT1$, $DT2$ and NT by the maximal likelihood criterion.

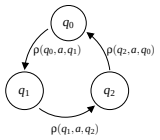
Table 7 gives the ρ function of the probability transducer $DT1$, that specifies the probabilities of match (1), transition (h) and transversion (0) at each codon position. Table 8 specifies the probability of each codon instance $a_1a_2a_3 \in \mathcal{A}^3$, used to define the probability transducer $DT2$. Finally, Table 9 specifies the probability transducer NT by specifying the four $DT2$ models together with transition probabilities between the initial states of each of these models.

Algorithm 1: S_π computation

Data : a seed π of span m, $'\#'$-weight w, and number of jokers $r = m - w$
Result : an automaton $S_\pi =< Q, q_0, q_F, \mathcal{A}, \psi >$

$Q.add(q_F)$;
$q_0 \leftarrow \langle X = \emptyset, k = 0, t = 0\rangle$;
$Q.add(q_0)$;
$queue.push(q_0)$;
while $queue \neq \emptyset$ **do**
  $\langle X, k_X, t_X\rangle = queue.pop()$;
  for $a \in \mathcal{A}$ **do**
    /* compute $\psi(< X, t_X >, a) = \langle Y, k_Y, t_Y\rangle$ */
    if $a =' \#'$ **then**
      $t_Y \leftarrow t_X + 1$;
a $k_Y \leftarrow k_X$;
      $Y \leftarrow X$;
    else
      if $X = \emptyset$ **then**
b $Y \leftarrow U_X(t_X, a)$;
        $k_Y \leftarrow U_k(t_X, a)$;
      else
        /* use already processed $\psi(< X', t_{X'} >, a) \ldots$ */
c $X' \leftarrow X \setminus \{l_{k_X}\}$;
        $\langle Y', k_{Y'}, t_{Y'}\rangle \leftarrow \psi(< X', t >, a)$;
        /* ... to compute $\psi(< X, t_X >, a)$ */
d $k_Y \leftarrow max\big(k_{Y'}, V_k(k_X, t_X, a)\big)$;
        $Y \leftarrow Y' \cup V_X(k_X, t_X, a)$;
      $t_Y \leftarrow 0$;
    if $L[k_Y] + t_Y \geq m$ **then**
      /* $< Y, t_Y >$ is a final state */
      $\psi(< X, t_X >, a) \leftarrow q_F$;
    else
      if $\langle Y, k_Y, t_Y\rangle \notin Q$ **then**
        $Q.add(\langle Y, k_Y, t_Y\rangle)$;
        $queue.push(\langle Y, k_Y, t_Y\rangle)$;
      $\psi(< X, t_X >, a) \leftarrow \langle Y, k_Y, t_Y\rangle$;

$a:$	0	h	1
$\rho(q_0, a, q_1)$	0.2151	0.1923	0.5926
$\rho(q_1, a, q_2)$	0.1742	0.1372	0.6886
$\rho(q_2, a, q_0)$	0.2033	0.1673	0.6294



TAB. 7 – *Parameters of the $DT1$ model*

$a_1 \backslash a_2 a_3:$	00	0h	01	h0	hh	h1	10	1h	11
0	0.01528	0.01051	0.02346	0.01092	0.00589	0.01838	0.02321	0.01566	0.09701
h	0.01073	0.00680	0.01765	0.00604	0.00424	0.01386	0.01607	0.01257	0.10026
1	0.02200	0.01609	0.05257	0.01625	0.01174	0.05017	0.08430	0.08483	0.25349

TAB. 8 – *Probability of each codon instance specified by the $DT2$ model*

$Pr(q_i \to q_j)$	$j=0$	1	2	3
$i=0$	0.88560	0.11440	0	0
1	0.15471	0.71550	0.12979	0
2	0	0.19805	0.72473	0.07722
3	0.03720	0.04031	0.22548	0.69701

$a_1 \backslash a_2 a_3:$		00	0h	01	h0	hh	h1	10	1h	11
0		0.01962	0.01404	0.02872	0.01431	0.00829	0.02288	0.03002	0.02040	0.09582
q_0 : h		0.01397	0.00928	0.02117	0.00820	0.00606	0.01680	0.02171	0.01773	0.09349
1		0.02418	0.01826	0.05088	0.01824	0.01373	0.04902	0.08167	0.08565	0.19589
0		0.01436	0.01013	0.02312	0.01011	0.00539	0.01783	0.02181	0.01450	0.09870
q_1 : h		0.01020	0.00620	0.01677	0.00533	0.00365	0.01323	0.01436	0.01149	0.10103
1		0.02234	0.01587	0.05349	0.01624	0.01096	0.05114	0.08792	0.09174	0.25207
0		0.01032	0.00614	0.01740	0.00714	0.00357	0.01334	0.01621	0.01150	0.09941
q_2 : h		0.00740	0.00441	0.01361	0.00371	0.00260	0.01032	0.01074	0.00812	0.10737
1		0.01844	0.01327	0.05469	0.01300	0.00985	0.05384	0.08305	0.08628	0.31426
0		0.00750	0.00488	0.01286	0.00565	0.00265	0.01016	0.01321	0.00900	0.09021
q_3 : h		0.00474	0.00305	0.00988	0.00248	0.00195	0.00824	0.00823	0.00755	0.10824
1		0.01364	0.01022	0.04729	0.01034	0.00763	0.04667	0.07385	0.07238	0.40751

TAB. 9 – *Probabilities specified by the NT model*

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399