

Quelques aspects de l'algorithmique du texte

Thierry Lecroq

*Habilitation soutenue le 8 décembre 2000
à l'Université de Rouen*

Membres du jury :

Michel Habib (président, rapporteur), Serge Dulucq (rapporteur), William F. Smyth (rapporteur), Maxime Crochemore, Jean-Pierre Duval, Jean Frédéric Myoupo

Bien que les données peuvent être mémorisées à l'aide de divers supports, le texte demeure la forme principale pour échanger l'information. C'est particulièrement évident en littérature ou en linguistique où les données sont constituées de corpus et de dictionnaires énormes. Ceci s'applique aussi à l'informatique où une grande quantité de données est enregistrée dans des fichiers linéaires. Et c'est également le cas en biologie moléculaire où les molécules biologiques peuvent souvent être représentées par des suites de nucléotides ou d'acides aminés. En outre, la quantité de données disponible dans ce domaine tend à doubler tous les dix-huit mois.

C'est pour toutes ces raisons que les algorithmes de recherche d'informations doivent être efficaces même si la vitesse et la capacité de mémoire des ordinateurs augmentent régulièrement.

Le problème de la recherche de motif consiste à localiser une ou plus généralement toutes les occurrences d'un motif dans un texte. Un motif peut être un mot, un ensemble fini de mots ou un ensemble infini de mots exprimé sous la forme d'une expression rationnelle. Ce problème important apparaît dans beaucoup de domaines de l'informatique notamment dans les traitements de textes et en analyse lexicale. En biologie cela intervient dans l'analyse de l'ADN et des séquences de protéines. Pour un certain nombre de programmes les techniques utilisées en recherche de motif constituent un gros pourcentage du travail effectué. Améliorer ces techniques augmente considérablement le rendement de ces programmes.

On s'intéressera ici à la recherche d'un mot dans un texte. Les techniques utilisées dans ce type de problème servent généralement de bases aux autres types de recherche.

Le problème de la recherche d'un mot x de longueur m dans un texte y de longueur n possède deux variantes. Lorsque le texte y est connu à l'avance, un prétraitement

est alors autorisé sur celui-ci, pour construire une structure d'index en temps et espace $O(n)$ et la localisation de toutes les occurrences de x dans y peut alors s'effectuer en temps et espace $O(m)$. Lorsque le mot x est connu à l'avance, un prétraitement est alors autorisé sur celui-ci, pour construire une structure de données en temps et espace $O(m)$ et la localisation de toutes les occurrences de x dans y peut alors s'effectuer en temps $O(n)$. On s'intéresse ici au second cas.

Les algorithmes de recherche d'un mot fixe dans un texte utilisent un mécanisme dit « de *fenêtre glissante* ». Une fenêtre de la même longueur que le mot x est utilisée pour accéder au texte y . L'extrémité gauche de cette fenêtre est d'abord alignée avec l'extrémité gauche du texte. Un travail spécifique appelé *tentative* permet de déterminer si le contenu de la fenêtre est identique à celui du mot. Ensuite l'algorithme *décale* la fenêtre vers la droite et le même processus est appliqué jusqu'à ce que l'extrémité droite de la fenêtre dépasse l'extrémité droite du texte. Les algorithmes de recherche de mot diffèrent par les méthodes employées lors des tentatives et des décalages. Un « bon » algorithme de recherche de mot cherche à minimiser le travail effectué lors de chaque tentative et à maximiser la longueur des décalages.

Nous présentons la famille des algorithmes de type « Boyer-Moore » qui effectue les comparaisons entre lettres du mot et lettres du texte de la droite vers la gauche. Ces algorithmes sont très performants dans la pratique : des résultats expérimentaux sont exhibés.

La comparaison entre deux mots est un autre champ d'investigation important. Le problème est connu sous le nom d'alignement de séquences. Le problème de base consiste à déterminer les ressemblances, ou de manière duale, les différences entre deux mots x et y . Outre le fait qu'il existe différentes manières d'exprimer ces ressemblances ou ces différences, les solutions peuvent être visualisées sous forme d'alignements. Ce problème intervient de manière cruciale en biologie moléculaire où le degré de similarité entre deux séquences est une indication d'une homologie éventuelle entre ces deux séquences. On présente les méthodes classiques de résolution de ce problème qui utilise des techniques de programmation dynamique. On en donne une solution systolique dans le cas restreint de la recherche d'un plus long sous-mot commun à deux mots.

Nous présentons deux sites Web dynamiques illustrant les deux notions précédentes. Ces notions sont effectivement difficiles et il est important de disposer d'outils permettant de les mieux appréhender.

Le premier site, intitulé « Exact String Matching Algorithms » [CHA 96] s'adresse à des informaticiens théoriciens et praticiens et présente une trentaine d'algorithmes de recherche exacte de mots. Le second s'appelle « Sequence Comparison » [CHA 98] et peut aider des biologistes et des informaticiens à comprendre les techniques de bases d'alignement de séquences.

Nous présentons ensuite des applications de techniques algorithmiques du texte aux séquences musicales et biologiques respectivement. Ces deux domaines demandent en effet d'adapter des méthodes générales à des situations très spécialisées.

À un niveau très rudimentaire, une *séquence musicale* peut être représentée par un mot. L'alphabet peut être l'ensemble des notes ou l'ensemble des intervalles entre les notes (les hauteurs des notes peuvent être données par des nombres MIDI (Musical Instrument Digital Interface) et les intervalles par des nombres de demi-tons). Il faut bien sûr ajouter à cette représentation un codage pour les durées des notes. La recherche exacte de mot dans ce contexte ne présente guère d'intérêt. Nous présentons des méthodes de recherche approchée appropriées aux séquences musicales ainsi représentées.

Nous présentons ensuite une méthode rapide pour détecter de longues répétitions dans les mots ainsi que des applications aux *séquences génomiques*. Les séquences génomiques peuvent être représentées par des mots sur un alphabet à quatre lettres pour l'ADN et l'ARN (alphabet des *acides nucléiques*) et sur un alphabet à vingt lettres pour les *protéines* (alphabet des *acides aminés*). Les séquences d'acides nucléiques peuvent dépasser les trois milliards de lettres comme c'est le cas pour le génome humain. Il est donc fondamental de trouver des structures de données qui soient à la fois compactes et rapides à manipuler pour les traiter. Nous utilisons pour ce faire l'oracle des facteurs.

Quelques-uns de ces travaux ont été publiés dans les ouvrages suivants : [CRO 96], [CRO 98] et [CRO 01].

Enfin nous concluons en décrivant les champs d'activité où un travail important reste à effectuer.

Bibliographie

- [CHA 96] CHARRAS C., LECROQ T., « Exact string matching algorithms », URL:<http://www-igm.univ-mlv.fr/~lecroq/string/>, 1996.
- [CHA 98] CHARRAS C., LECROQ T., « Sequence comparison », URL:<http://www-igm.univ-mlv.fr/~lecroq/seqcomp/>, 1998.
- [CRO 96] CROCHEMORE M., LECROQ T., « Pattern matching and text data compression algorithms », TUCKER, JR A. B., Ed., *The Computer Science and Engineering Handbook*, Chapitre 8, p. 162–202, CRC Press Inc., Boca Raton, FL, 1996.
- [CRO 98] CROCHEMORE M., LECROQ T., « Text data compression algorithms », ATALLAH M. J., Ed., *Algorithms and Theory of Computation Handbook*, Chapitre 12, CRC Press Inc., Boca Raton, FL, 1998.
- [CRO 01] CROCHEMORE M., HANCART C., LECROQ T., *Algorithmique du texte*, Vuibert, 2001, À paraître.

LIFAR – ABISS
Faculté des Sciences et Techniques
Université de Rouen
76821 Mont-Saint-Aignan cedex
Thierry.Lecroq@univ-rouen.fr