Sujet : Recherche de motifs approchés dans des séquences: de la biologie à la musique.

Responsable : Pascal FERRARO, Pierre HANNA et Julien ALLALI (LaBRI)
Téléphone : 05 40 00 35 07
e-mail : ferraro@labri.fr
Site Web : www.labri.fr/~ferraro
École doctorale de mathématiques et informatique de Bordeaux, ED-39
Laboratoire : Laboratoire Bordelais de Recherche en Informatique
Equipe : Modèles et Algorithmes pour la BioInformatique et la Visualisation et Modélisation du Son et de la Musique - Image et Son (PROJET SIMBALS)
Localisation : Bordeaux

Contexte. La recherche sur l'extraction de données musicales (Music Information Retrieval) consiste à développer des systèmes d'analyse, de comparaison ou d'indexation de pièces musicales. Le projet SIMBALS, développé au LaBRI, propose dans ce contexte des méthodes d'évaluation de la similarité entre morceaux de musique.

Description du travail demandé et objectif.

L'analyse de données séquentielles est une opération classique pour laquelle il existe un grand nombre d'applications. Plus particulièrement, les séquences de caractères sont des structures largement utilisées pour codées des données en biologie et ont été très largement étudiée dan ce contexte. Nous proposons dans ce projet d'adapter les méthodes développées pour analyser les structures séquentielles biologiques au contexte d'analyse musicale. Ce travail sera réalisé dans le cadre du projet transdisciplinaire SIMBALS regroupant des chercheurs de l'équipe Algorithmique pour l'analyse de structures biologiques et de l'équipe Modélisation du Son et de la Musique.

L'intérêt de cette étude repose sur la croissance importante des grandes bases de données musicales disponibles en ligne (mobile, internet). La taille de ces bases de données impose des techniques rapides de navigation. L'indexation des morceaux de musique représente donc un enjeu important.

Nous proposons ainsi d'étudier la recherche de motifs approchés dans une séquence musicale, en nous basant sur des recherches effectuées dans le domaine de l'algorithmique du texte et/ou de la bioinformatique. Un travail important a déjà été réalisé pour rechercher des motifs ou pour identifier des répétitions exactes dans des séquences. Nous nous intéresserons dans un premier temps à la généralisation de ces méthodes dans le contexte musical.

Une méthode d'indexation efficace (et donc de recherche de motifs) en Bioinformatique est basée sur la notion d'arbre des suffixes. Un arbre de suffixes est une structure en arbre où chaque nœud de l'arbre est une chaîne de caractères où, à partir de la racine et jusqu'à chaque « feuille », on énumère l'ensemble des suffixes ; chaque nœud de l'arbre qui n'est pas une « feuille » possède au moins deux « descendants » (voir l'exemple). On peut construire un arbre de suffixes en temps O(n) où n est la longueur de la chaîne de caractères. Les méthodes que nous développerons dans le contexte musical s'appuieront dans un premier temps sur l'utilisation de cette structure d'indexation particulière.

En revanche, dans les applications réelles, il semble plus pertinent de s'intéresser à la détermination de répétitions ou de motifs approchés. Or il n'existe pour l'heure que peu de travaux dans la littérature s'intéressant à ce type d'algorithmes de recherche. Nous nous appuierons sur la mesure de la similarité entre séquences pour traiter ce problème.

La famille d'algorithmes que nous étudierons consiste à reconstruire une séquence cible en appliquant à une séquence initiale différentes opérations structurelles élémentaires, appelées opérations d'édition. On peut montrer que la définition d'une telle distance se ramène à la recherche d'une solution optimale dans un espace dont la taille croît exponentiellement avec la taille des objets à comparer. Aussi, pour définir des algorithmes efficaces à la fois en temps et en espace, la solution optimale est construite de

façon incrémentale en utilisant des techniques de programmation dynamique. Dans les 40 dernières années, plusieurs équipes se sont attaquées à différents aspects du problème posé par l'édition des séquences biologiques. Ces premières méthodes d'analyse ont ouvert un nombre important de voies de recherches algorithmiques et applicatives. En nous appuyant sur ce type d'algorithmes et sur leur adaptation au contexte musicale, l'objectif de ce travail de thèse consistera à adapter les méthodes de recherche de motifs exacts à l'utilisation de ces méthodes de mesure de similarité entre séquences.

Les applications sont très nombreuses et concernent aussi bien les méthodes rapides de navigation dans les bases de données que la génération automatique de résumés musicaux. Ce type d'algorithmes pourra permettre entre autres de détecter toutes sortes de répétitions dans un morceau de musique. Ces motifs peuvent être très représentatifs du morceau analysé : un thème, un refrain, un couplet, etc. L'une des principales applications alors induites pourra être la génération automatique de résumés. Au lieu de représenter un morceau de musique par ses premières secondes, celui-ci pourra être résumé par quelques secondes du refrain et du couplet et ainsi être plus représentatif du morceau original.

Références :

Ukkonen E. On-line construction of suffix trees. Algorithmica, 14:249-60, 1995

Smith TS and Waterman MS. Identification of common molecular subsequences, J. Mol. Biol., 147:195-97, 1981