تنفيذ موازٍ لتحليل تنميط الحمض النووي مكون من خليط كبير من المساهمين

عماد محمد أحمد العمودي

المستخلص

تحديد عدد المساهمين في الملف الحمض النووي هو ممارسة شائعة في مختبرات الطب الشرعي. على سبيل المثال، فإنه يساعد في حالات الاعتداء الجنسي عندما يحتوي خليط الحمض النووي على افراد مختلفين مثل الضحية، والمجرم، بالإضافة إلى شريك الضحية. هذه المشكلة (المعروف أيضا باسم تنميط الحمض النووي) هي واحدة من أصعب المشاكل في مجال العلوم الجنائية. وسيزداد التعقيد مع زيادة عدد المساهمين غير المعروفين. وقد وضعت بعض الأساليب وتطبيقاتها البرمجية لمعالجة هذه المشكلة. مع ذلك، فإن التعقيدات الحسابية هي عامل الردع الرئيسي الذي يمنع تقدم المجال وتطبيقاته. هناك بعض التقدم الذي تم باستخدام التنفيذ الموازي باستخدام OpenMP لتنميط الحمض النووي ولكنه تم على مشاكل صغيرة غير قابلة للتطوير.

الهدف من هذه الرسالة هو تحسين أداء أدوات تنميط الحمض النووي باستخدام الحوسبة المتوازية، مما يتيح تفسير المخاليط مع عدد أكبر من المجاهيل ضمن إطار زمني أقصر. قمنا بتطوير اثنين من التطبيقات لتنميط الحمض النووي مع التركيز على احتمالات Likelihood computation. يستند التنفيذ الأول على نموذج OpenMP. وتمكنا من تحقيق تنفيذ أسرع بعامل يصل إلى ثلاث مرات مقارنة بأفضل أداة متاحة (NOCIt). ويستند التنفيذ الثاني إلى تنفيذ هجين من OpenMP/ MPI ويعتبر هذا هو أول تطبيق هجين (OpenMP/MPI) لتنميط الحمض النووي والذي سيمكن حسابات النسبة الاحتمالية (Likelihood ration) لتوسيع نطاقها إلى أي عدد تقريبا من المجاهيل. قمنا بعرض النتائج لتنفيذ الهجين لعدد يصل إلى ١٠ مجاهيل وحصلنا على اداء أفضل بـ x٥٢ من اداء OpenMP، وذلك باستخدام ما يصل إلى ٣٠٧٢ معالج، وتخفيض المدة المتطلبة لحساب النسبة الاحتمالية من ٥,٨ ايام إلى ٢,٧ ساعة. في السنوات المقبلة، من المتوقع أن تكون تقنيات تسلسل الجينوم الكامل متاحة لخلية واحدة أو عدد قليل من الخلايا، ومن المرجح أن تغيير هذه التقنيات مشهد تنميط الحمض النووي. ومن المتوقع أن تفتح تفسيرات أسرع لمخاليط الحمض النووي مع عدد كبير من المجاهيل ودرجات دقة اعلى حدودا جديدة لهذا المجال.

ساهمت هذه الأطروحة بنشر ورقة علمية في Web of Science مفهرسة من قبل Springer.

Parallel Analysis of DNA Profile Mixtures with a Large Number of Contributors

Emad Mohammed Alamoudi

ABSTRACT

Determining the number of contributors in a DNA profile is a popular practice in various forensic laboratories. For example, it helps in cases of sexual assault when the source of DNA mixture can combine different individuals such as the victim, the criminal, and the victim partner. This problem (also known as DNA profiling) is one of the hardest in the forensic science domain. The complexity would increase as the number of unknown contributors would increase. A few methods and their software implementations have been developed to address this problem. However, its (exponential) computational complexity has been the major deterring factor holding its advancements and applications. Some work on OpenMP based parallel implementation of DNA profiling exists but it has only been applied to small problems and is not scalable.

The aim of this thesis is to improve the performance of DNA profiling tools using parallel computing, enabling the interpretation of mixtures with a larger number of unknowns within a shorter time frame. We developed two different implementations of DNA profiling focusing on the maximum likelihood computations. The first implementation is based on the OpenMP paradigm. We were able to achieve faster multicore implementations by a factor of up to three compared to the best software tool available (i.e. NOCIt) globally. The second implementation is based on the hybrid OpenMP/MPI implementation. This is the first ever hybrid implementation of DNA profiling enabling the likelihood ratio computations to scale to virtually any number of unknowns. We report results of the hybrid implementation for up to 10 unknowns delivering over 52x performance over OpenMP, using up to 3072 cores, reducing the likelihood computation time from 5.8 days to 2.7 hours. In the coming years, the complete genome sequencing technologies are expected to be available in a single or only a few cells, and this is likely to change the DNA profiling landscape. Faster interpretations of DNA mixtures with a large number of unknowns and higher accuracies are expected to open up new frontiers for this area.

This thesis has contributed one paper in Web of Science indexed proceedings by Springer.