Passage à l'échelle de la méthode de « Multidimensional Scaling » pour l'étude de la biodiversité

Encadrants: Olivier Coulaud (Hiepacs) & Alain Franc (Pleiade), INRIA SO & Labri

Téléphones : OC : 05 24 57 40 80 ; AF : 05 35 38 53 53

Courriels : olivier.coulaud@inria.fr ; alain.franc@inra.fr

Présentation du sujet :

L'étude de la biodiversité, classiquement fondée sur une approche naturaliste et prudente de reconnaissances d'espèces et production d'inventaires, est entrée brutalement dans le monde de la valorisation de données massives, produites par des séquenceurs actuellement de troisième génération (domaine du métabarcoding, voir https://en.wikipedia.org/wiki/DNA_barcoding). On connaît des marqueurs qui permettent d'identifier les espèces, et le séquençage massif permet actuellement la production de centaines d'inventaires (qui sont à la base des études en écologie) en quelque semaines uniquement (alors qu'il aurait fallu plusieurs décennies au siècle dernier).

La question est posée de traiter ces données, avec classiquement des méthodes d'analyses multivariées, donc du calcul matriciel (voir [5] pour les grandes dimensions). Les besoins concernent *in fine* des décompositions spectrales, soit en valeurs singulières, de très grandes matrices pleines (de l'ordre de 10^5 à 10^6 comme dimension). Les algorithmes « standards » sont cubiques selon la dimension n, et ne passent pas facilement à cette échelle. Dans ce cadre, une collaboration entre les équipes Pleiade et Hiepacs (dans le cadre d'une thèse, voir [1]) a permis de connecter le besoin de ces outils pour les études de biodiversité avec un champs récent pour le passage à l'échelle en calcul matriciel dense : les méthodes de projection aléatoires [3,4]. Cela s'est traduit par le développement d'une librairie (`fmr`, voir [1]), qui met en œuvre ces méthodes, et plus largement des sélections de colonnes, avec un parallélisme basé sur OpenMP. Elle est écrite en C++ et fait appel aux bibliothèques standards de calcul matriciel intensif (BLAS, Lapack, Arpack). Dans le cadre de l'approche MDS (Multidimensional Scaling, voir [2]), elle permet de traiter des matrices de dimension 10^5. Les opérations « élémentaires » sont des produits matrice-matrice, matrice-vecteur, et une décomposition QR. Ces opérations peuvent se réaliser par blocs, ce qui permet une distribution des calculs sur plusieurs nœuds de calcul.

Travail :

L'objectif du stage est de paralléliser l'algorithme MDS en utilisant le paradigme MPI/OpenMP afin d'offrir la possibilité de traiter des matrices pleines de dimension 10^6. Pour cela, les opérations élémentaires produits matrice-matrice, matrice-vecteur, et une décomposition QR seront parallélisées en se basant sur les bibliothèques Scalapack, mkl, ou chameleon. Les développements seront intégrés au sein de `fmr`. Dans la deuxième partie du stage, on étudiera l'influence de la

décomposition sur différents jeux de données. Nous disposons pour cela de dix jeux de données, d'environ 10^5 séquences chacun, concernant des cortèges de diatomées du lac Léman. L'objectif de ce stage est de permettre le traitement de l'ensemble des ces échantillons conjointement, afin de réaliser une intercalibration entre échantillons (unités taxonomiques partagées, ou spécifiques). La taille du jeu de données à traiter pour cela est de 10^6 séquences.

La validation de l'approche se fera faite sur les machines de PlaFRIM et du Genci pour les gros cas tests.

Ce sujet contribue à

- un rapprochement entre le domaine de la biodiversité et du calcul intensif (biodiversité computationnelle)

- une connexion entre des méthodes issues du calcul HPC (algèbre linéaire, paradigme MPI) et des besoins actuels du calcul intensif dans le cadre du traitement de données massives (HTC, paradigme map-reduce).

Mot-clés : Décomposition en valeurs singulières ; données massives ; algèbre linéaire dense ; données massives ; parallélisation MPI/OpenMP ; projections aléatoires.

Commentaires : Stage effectué sur Bordeaux, au sein de l'équipe Pleiade, en partenariat avec l'équipe Hiepacs.

Références :

[1] Blanchard, P. Fast hierarchical methods for the low-rank approximation of matrices. *PhD thesis, Université de Bordeaux, 2016*

[2] T.F. Cox and M. A. A. Cox. *Multidimensional Scaling* - Second edition, volume 88 of Monographs on Statistics and Applied Probability. Chapman & al., 2001.

[3] N . Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review,* **53(2):217-288,** 2011.

[4] S. Vempala. *The Random Projection Method,* volume 65 of DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences. American Mathematical Society, 2004.

[5] J. Wang. *Geometric structure in high-dimensional data and dimensionality reduction.* Springer & Higher Education Press, 2012.