Mésochallenge 2014: Biodiversiton

Y. Laizet, A. Moreau, J.-M. Frigerio, Ph. Chaumeil, P. Gay, P. Ramet, D. Sherman, A. Franc

Equipe Pleiade, INRIA/INRA, Bordeaux

JSCIA, 24 mars 2015

Contexte général

1. Contexte de biodiversité : en quoi les objets biologiques sont-ils différents ?
2. Quels sont les patterns et motifs, locaux et gloabaux ?
3. Domaine de l'évolution et l'écologie des communautés
4. Travaux sur les arbres de la forêts guyanaise
5. Génomes comme empreinte moléculaire de l'évolution
6. Pas de processus fonctionnels

Projet

1. Taxonomie moléculaire
2. Chaque individu a un *attribut* : une séquence (un mot de 500 lettres, alphabet de 4 lettres $w \in \{A, T, C, G\}^n$
3. Histoire inférée par des phylogénies moléculaire (Tree of life)
4. Modèles statistiques atteignent leurs limites si $10^4 \sim 10^3$ individus (ML, bayésien)
5. Travail sur des distances
6. Distance d'édition (Levenstein, 1965)

Projet

1. Taxonomie moléculaire
2. Chaque individu a un *attribut* : une séquence (un mot de 500 lettres, alphabet de 4 lettres $w \in \{A, T, C, G\}^n$
3. Histoire inférée par des phylogénies moléculaire (Tree of life)
4. Modèles statistiques atteignent leurs limites si $10^4 \sim 10^3$ individus (ML, bayésien)
5. Travail sur des distances
6. Distance d'édition (Levenstein, 1965)

Question

Quelle est la forme d'un nuage de points ?
Distance geometry, machine learning, manifold learning, ...

Multidimensional Scaling : le problème

- Soit un ensemble $V = \{1, n\} \subset \mathbb{N}$ de n objets
- Les distances deux à deux sont donnés : $d(i,j) := d_{ij}$
- On se donne un entier $r < n$
- On cherche n points $x_i \in \mathbb{R}^r$ tels que

$$d_x(i,j) := \|x_i - x_j\| \simeq d_{ij} \tag{1}$$

Résultat

Il existe une solution exacte : méthode spectrale

voir Izenman, 2007, Modern Multivariate Statistical Techniques

Multidimensional Scaling : la solution

Remarque

Il est possible de reconstruire $\langle x_i, x_j \rangle$ *à partir des* $d_x(i,j) = \|x_i - x_j\|$.

Algorithm 1 pseudocode for Multidimensional Scaling

1: compute $C_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2), \quad 1 \leq i,j \leq n$
2: compute $x_k \ : \ Cx_k = \lambda_k x_k$
3: $X = (x_k)_k$
4: $L = \mathrm{diag}\,(\sqrt{\lambda_k})_k$
5: $Y = XL$

Nonlinear mapping

- MDS : minimisation de la valeur absolue des écarts
- NLM (Sammon, 1969) : poids faible sur les distances grandes : minimisation sur les valeurs relatives
- Fonction à minimiser :

$$\phi = \frac{1}{c} \sum_{i<j} \frac{[d_{ij} - d_x(i,j)]^2}{d_{ij}} \tag{2}$$

où c est une constante de normalisation

Nonlinear mapping : formalisation

- On se donne un tableau de distances

$$D = [d_{ij}]_{i,j=1\dots n} \tag{3}$$

et une fonction objectif

$$\phi = \sum_{i<j} \omega(d_{ij})[d_{ij} - d_x(i,j)]^2 \tag{4}$$

Problème

$$\left|\begin{array}{ll} D & \textit{étant donné} \\ p & \textit{étant donné} \\ \omega & \textit{étant donné} \\ \textit{trouver} & X = \{x_1, \dots, x_n\} \subset \mathbb{R}^p \\ \textit{tel que} & \phi \quad \textit{minimal} \end{array}\right. \tag{5}$$

Verrou pour la parallélisation

- On peut définir

$$\begin{aligned}\phi(X) &= 2\sum_{i<j}\omega(d_{ij})[d_{ij} - \|x_i - x_j\|]^2 \\ &= \sum_i \phi_i(x_i) \quad \phi_i(x) = \sum_{j\neq i}\omega(d_{ij})[d_{ij} - \|x - x_j\|]^2\end{aligned} \tag{6}$$

- Mais ...ce diagramme n'est pas commutatif

$$\begin{tikzcd}
(\ldots, x_i, \ldots, x_j \ldots) \arrow[r, "t+1"'] \arrow[d, "t+1"] \arrow[dr] & (\ldots, x_i, \ldots, x_j' \ldots) \arrow[d, "t+2"] \\
(\ldots, x_i', \ldots, x_j \ldots) \arrow[r, "t+2"] & (\ldots, x_i'', \ldots, x_j'' \ldots)
\end{tikzcd}$$

Un algorithme séquenciel

Algorithm 2 pseudocode for sequential optimisation

Sequential optimization
for $t = 1$ to s **do**
    for $i = 1$ to n **do**
        select x_i
        compute $\nabla_i = \nabla\phi|_{x_i}$
        find α such that $\phi_i(x_i + \alpha\nabla_i)$ minimum
        update $x_i \quad \leftarrow \quad x_i + \alpha\nabla_i$
    end for
end for

Approximation

- remplacer

$$\phi_i(x) = \sum_{j \neq i} \omega(d_{ij})[d_{ij} - \|x - x_j\|]^2 \tag{7}$$

par

$$\phi_i(x) = \sum_{0 < d_{ij} < \theta} \omega(d_{ij})[d_{ij} - \|x - x_j\|]^2 \tag{8}$$

- Alors, si $d_{ij} > \theta$, le diagramme est commutatif, et (i, j) peuvent être optimisés en parallèle (si $d_x(i, j) > \theta$)

Problèmes à aborder pour la parallélisation

Un exemple de parallélisation MPI avec les mathématiques discrètes

1. partition : choisir quels points envoyer en parallèle ou non
2. distribution : allouer les points aux cœurs
3. équilibre des charges

Parallélisation

- On définit un seuil ϵ et on en déduit θ tel que

$$d_{ij} > \theta \quad \Longrightarrow \quad \omega(d_{ij})[d_{ij} - d_x(i,j)]^2 < \epsilon \tag{9}$$

- On construit deux graphe $G = (V, E)$ et $G' = (V, E')$ où
  1. V est l'ensemble des points $V = \{1, n\}$
  2. $i \sim j$ dès que $d_{ij} > \theta$ pour G, $d_{ij} < \theta$ pour G',

- Alors, si $i \sim j$ dans E, on peut envoyer i et j en parallèle sur deux cœurs différents. De même un ensemble formé de un point par composante connexe de G'.

Résultat

S'il existe une clique de m points dans G, alors, on peut envoyer m points en parallèle sur m cœurs. De même s'il existe m composantes connexes dans G'

Partition par coloriage

- Cette partition est non optimale (cliques trop restrictives, et G' peut être connexe par linkage avec des points très élignés).

Remarque

Il existe un lien avec le coloriage de graphes : trouver le nombre minimal de couleurs telles que deux sommets reliés par un lien sont de couleurs différentes

- Si deux points sont de même couleur, alors ils peuvent être envoyés en parallèle

Procédure

1. *Identifier les composantes connexes de G'*
2. *Colorier chaque composante connexe*
3. *Envoyer en parallèle une couleur complète de chaque composante*

Coloration de graphe

Definition

Une *coloration* d'un graphe est une partition des sommets en ensembles deux à deux non adjacents. Le *nombre chromatique d'un graphe* est le nombre minimal d'éléments d'une coloration. Il est noté en général $\chi(G)$.

Remarque

Trouver une coloration d'un graphe est un problème NP-complet.

- Il existe en revanche des algorithmes approchés quadratiques. Un algorithme glouton permet de trouver une coloration avec $d+1$ composantes si d est le degré maximal dans G.

Algorithme de parallélisation

Algorithm 3 pseudocode for parallel optimisation

Have $D = (d_{ij})_{i,j}$
Build $G = (V, E) : i \sim j \implies d(i,j) < \theta$
Compute the connected components $= (C_1, \ldots, C_m)$
Compute a minimal coloration $C_k = \bigsqcup_{j \leq K} K_{kj}$
set $X = (x_1^0, \ldots, x_n^0)$ (Initialize by MDS)
for $t = 1$ to s **do**
    $\texttt{distrib} = \emptyset$
    for $k = 1$ to K **do**
        for $a = 1$ to m **do**
            $X_{ak} = \{x_i \in C_a \ : \ \kappa(x_i) = k\}$
            $\texttt{distrib} = \texttt{distrib} \cup X_{ak}$
        end for
    end for
    `mpiexec -np N optimize [param]`
end for

Etat des lieux

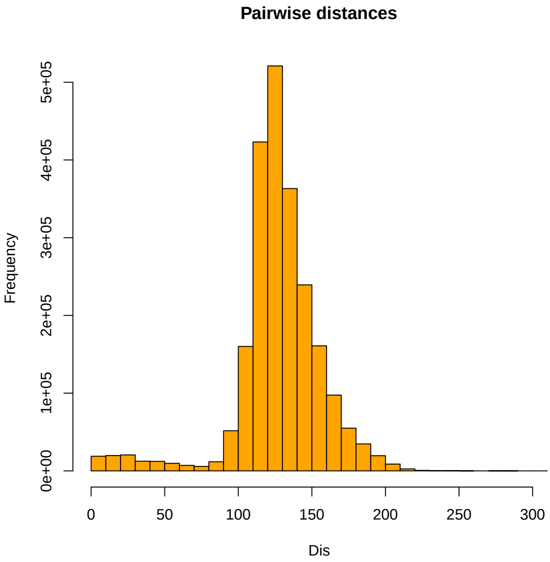
- Ce qui a été fait

1. Poser le problème
2. Résoudre l'optimisation pour un point
3. Algorithme d'optimisation en séquenciel (moindres carrés alternés)
4. Paralléliser le code sur les points
    1. partition des points selon une coloration
    2. construire une distribution
    3. écrire un code en MPI-python

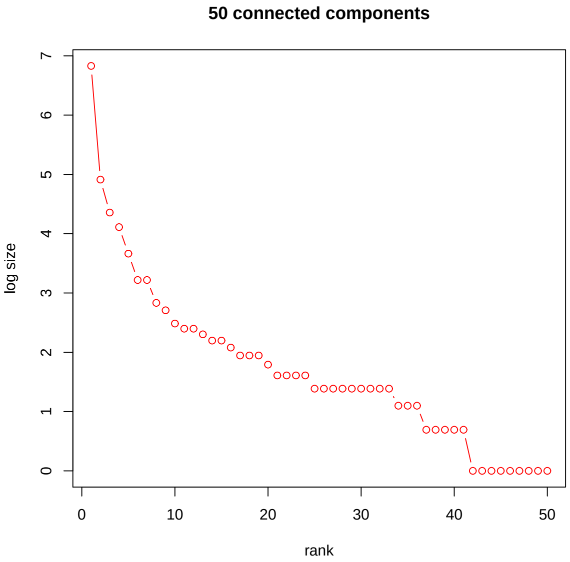
- Ce qui reste à faire

1. Equilibrer la distribution
2. Accélérer l'optimisation pour un point
3. Calculer la partition à chaque étape sur $d_x(i,j)$ et non une fois pour toute sur $d(i,j)$
4. Utiliser la hiérarchie cœurs / nœuds pour écrire une partie en open-MP (mémoire partagée)

Histogramme des distances



Taille des composantes connexes



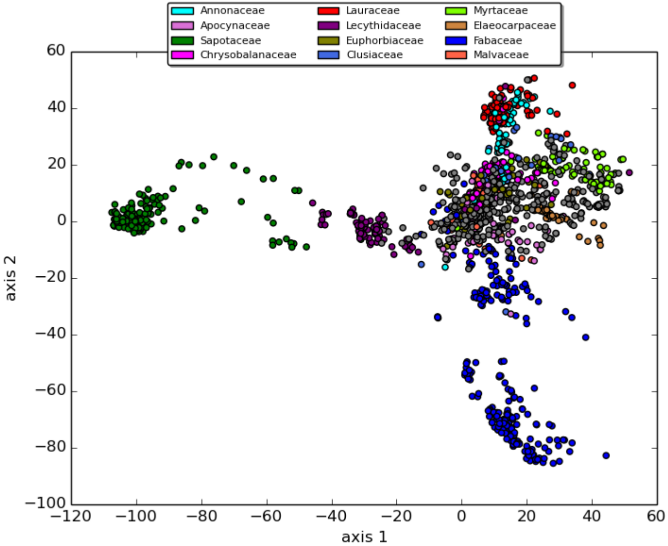
Apport du mésochallenge

Apport : développement

Mise au point du code plus que calcul intensif

1. Pas de calcul très intensif actuellement (reste du niveau de la queue habituelle Tier 2 en nombre de CPU et *wall clock*)
2. Mais accès prioritaire d'une partie de la machine pour la mise au point des codes
3. Perspective de passage à l'échelle sur Tier 1 (plus grands jeux de données)

Taxonomie : MDS sur 1500 arbres de Guyane



Taxonomie : NLM sur 1500 arbres de Guyane