Etudiants

CHEIKH MOHAMED Yarbana
TAKOUGOUM Steeve Rodrigue
YATABARE Cheikna Amala

Sujet 27 : TRANSPORT OPTIMAL GENERALISE POUR L'ANALYSE SINGLE-CELL AVEC QUANTIFICATION D'INCERTITUDES

Tuteur

STEPHANE CHRETIEN

Contexte

Le séquençage scRNA-seq permet d'étudier l'expression génique à l'échelle cellulaire, mais fait face aux défis de la haute dimension et du bruit. Les distances classiques étant limitées, ce projet exploite le transport optimal pour définir des mesures de similarité adaptées. Ces métriques permettent un clustering précis et une représentation fidèle des structures biologiques. L'étude utilise la distance de Wasserstein régularisée (analyse intra-individu) et la distance de Gromov-Wasserstein (analyse inter-individus).

Objectifs

- **Implémenter l'OT** : Utilisation de Sinkhorn (rapidité) et Gromov-Wasserstein (espaces distincts).
- **Analyser la biologie** : Étude des graphes de voisinage, de la qualité des clusters et des trajectoires cellulaires.
- **Quantifier l'incertitude** : Application de la prédiction conforme pour évaluer la fiabilité des distances.

Données

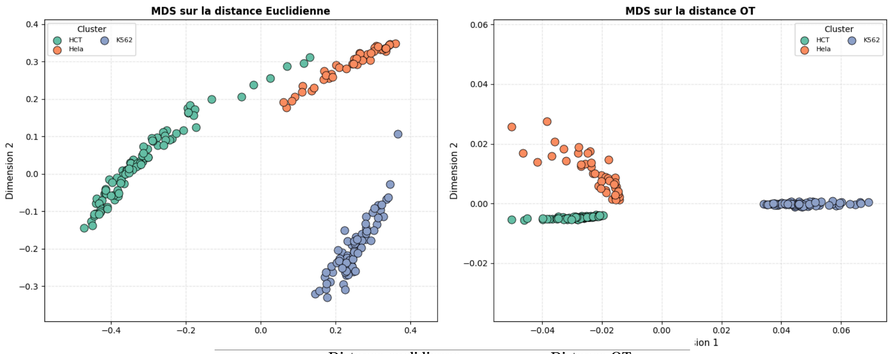
Base 1: Gabriel Peyré et al.

- **Échantillon** : 206 cellules issues de 3 lignées cancéreuses.
- **Structure** : Matrice fixe $X \in \mathbb{R}^{n \times m}$ (10,094 gènes et 206 cellules).
- **Mesure** : Logarithme du nombre de transcrits par gène.

Base 2: Données du Centre de Recherche en Cancérologie de Lyon (CRCL)

- **Échantillon** : 19 individus, chacun représenté par un ensemble de cellules.
- **Structure** : Matrices $X^{(i)} \in \mathbb{R}^{n_i \times p_i}$ propres à chaque patient.
- **Variabilité** : Dimensions (ni, pi) hétérogènes d'un individu à l'autre.

Données Intra-Individus : Wasserstein



	Distance euclidienne		Distance OT	
Indicateur	Clustering hiérarchique	Leiden	Clustering hiérarchique	Leiden
ARI	0.9216	0.9217	1.0000	1.0000
NMI	0.9084	0.8963	1.0000	1.0000

La distance de l'OT obtenue avec la version entropique de Sinkhorn nous permet de découvrir les clusters

L'ARI et le NMI sont deux indicateurs qui mesurent la qualité d'un clustering en le comparant aux vrais groupes connus. Ils valent 1 lorsque le clustering est parfait et 0 lorsqu'il n'est pas meilleur qu'un regroupement aléatoire.

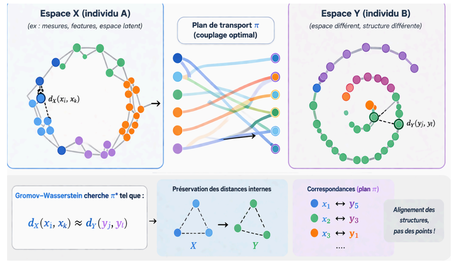
Effet de la variation d'epsilon

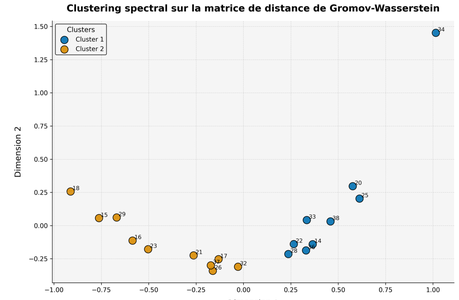
$$W_{C,\varepsilon}(a,\, b) := \min_{P \in \mathbb{R}_+^{n \times n}} << C, P >> + \varepsilon \sum_{i,j} P_{ij}(\log(P_{ij}) - 1),$$

La qualité des clusters s'améliore pour des valeurs de epsilon supérieures à 0.1

Données Inter Individus : Gromov-Wasserstein

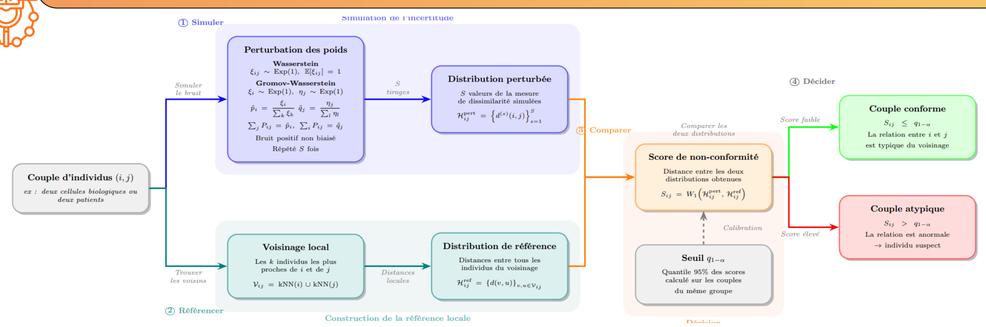
$$GW(\mu, \nu) = \sum_{i,j,k,l} (d_X(x_i, x_k) - d_Y(y_j, y_l))^2 P^*_{ij} P^*_{kl}$$



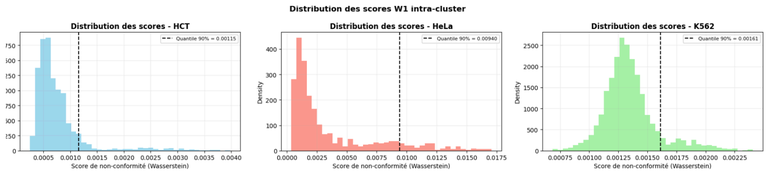


Les résultats montrent que la distance de Gromov-Wasserstein permet de capturer des similarités structurelles entre individus, même en présence d'espaces de représentation différents.

Méthodologie de perturbations et d'evaluation du score de non conformité

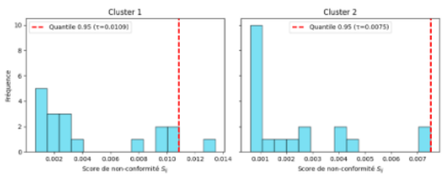


Base 1: Gabriel Peyré et al.



Le cluster HeLa se distingue par des scores nettement plus élevés et plus variables, suggérant une plus grande hétérogénéité interne.

Base2:Données du Centre de Recherche en Cancérologie de Lyon (CRCL)



Les patients du cluster 2 présentent un comportement nettement plus homogène après perturbation que ceux du cluster 1.

Discussion

- **Biais de Régularisation** : Un epsilon élevé lisse les distances et augmente artificiellement la compacité des clusters, au risque de perdre la fidélité à la structure réelle.
- **Limites de l'ARI** : L'obtention d'un ARI de 1 sur des données simples (lignées distinctes) ne garantit pas la supériorité de la méthode sur des tissus plus complexes.
- **Sensibilité aux Données** : La distance de Gromov-Wasserstein reste sensible à l'hétérogénéité et à la faible taille des échantillons (atypisme de certains patients).
- **Apport de la quantification de l'incertitude** : L'approche par perturbation montre que certains clusters sont fragiles, rendant l'analyse de stabilité indispensable.

Conclusion

- **Performance de l'OT :** Le Transport Optimal surpasse les métriques usuelles en capturant mieux la géométrie intrinsèque des données single-cell.
- **Flexibilité Multi-Espaces :** La variante Gromov-Wasserstein permet d'aligner avec succès des patients ayant des dimensions de gènes/cellules différentes.
- **Robustesse des Méthodes :** Les indicateurs (Silhouette, ARI, NMI) confirment la pertinence de l'OT pour identifier des similarités biologiques fines.
- **Perspectives** : Nécessité d'étendre ces outils à des jeux de données plus hétérogènes.