Bioinformatics, 38, 2022, ii148–ii154
https://doi.org/10.1093/bioinformatics/btac481
Proceedings of ECCB2022

Linking cells across single-cell modalities by synergistic matching of neighborhood structure

Borislav H. Hristov[1], Jeffrey A. Bilmes[2,3] and William Stafford Noble[1,3,*]

[1]Department of Genome Sciences, [2]Department of Electrical Engineering and [3]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA

*To whom correspondence should be addressed.

Abstract

Motivation: A wide variety of experimental methods are available to characterize different properties of single cells in a complex biosample. However, because these measurement techniques are typically destructive, researchers are often presented with complementary measurements from disjoint subsets of cells, providing a fragmented view of the cell's biological processes. This creates a need for computational tools capable of integrating disjoint multi-omics data. Because different measurements typically do not share any features, the problem requires the integration to be done in unsupervised fashion. Recently, several methods have been proposed that project the cell measurements into a common latent space and attempt to align the corresponding low-dimensional manifolds.

Results: In this study, we present an approach, Synmatch, which produces a direct matching of the cells between modalities by exploiting information about neighborhood structure in each modality. Synmatch relies on the intuition that cells which are close in one measurement space should be close in the other as well. This allows us to formulate the matching problem as a constrained supermodular optimization problem over neighborhood structures that can be solved efficiently. We show that our approach successfully matches cells in small real multi-omics datasets and performs favorably when compared with recently published state-of-the-art methods. Further, we demonstrate that Synmatch is capable of scaling to large datasets of thousands of cells.

Availability and implementation: The Synmatch code and data used in this manuscript are available at https://github.com/Noble-Lab/synmatch.

Contact: william-noble@uw.edu

1 Introduction

Recent developments in single-cell high-throughput sequencing technologies have led to the emergence of a myriad of experimental methods that are capable of characterizing different properties of single cells in a complex biosample. For example, high-throughput sequencing methods can measure RNA expression using single-cell RNA-seq (scRNA-seq), chromatin accessibility using scATAC-seq, chromatin 3D architecture using scHi-C and methylation profiles using scMethyl-seq. Ideally, researchers would like to be able to measure all of these properties in the same single cell in order to better understand the molecular underpinnings of the biological processes behind cell development and disease. However, because these measurement techniques are typically destructive, frequently only complementary measurements from disjoint subsets of a given population of cells are available, providing a patchwork view of the cell's biological processes. In such a situation, integration of the disjoint single-cell multi-omics data is critical. This has led to the rapid development of a variety of computational methods for single-cell multi-omics integration (Adossa *et al.*, 2021; Argelaguet *et al.*, 2020; Johansen and Quon, 2019). What makes the problem particularly challenging, however, is the fact that the different measurements, or modalities, typically do not share any features, and further, identifying correspondences between features in the domains may not be possible. Accordingly, existing methods which rely on either common cells or features across the data types cannot be applied in the fully unsupervised setting where correspondence information is absent.

This multi-modal integration problem can be generally framed in two distinct ways: (i) finding a discrete mapping between cells in the two modalities or (ii) embedding the disjoint measurements into a continuous shared latent space representing the intrinsic cellular structures across cellular modalities. The generalized unsupervised manifold alignment (GUMA) algorithm (Cui *et al.*, 2014), which uses a local geometry matching term, and MAGAN (Amodio and Krishnaswamy, 2018), which uses two generative adversarial networks, are examples of tools that find matchings of the cells between the two datasets. Recent methods that embed the two modalities into a common latent space and then attempt to align the embedded low-dimensional manifolds are SCOT (Demetci *et al.*, 2022), Pomona (Cao *et al.*, 2022a) and uniPort (Cao *et al.*, 2022b) all of which employ Gromov–Wasserstein optimal transport for alignment, and MMD-MA (Singh *et al.*, 2020), which aims to minimize the maximum mean discrepancy between the datasets in the latent space. Several methods rely on deep neural architectures to solve the manifold alignment task (Stark *et al.*, 2020; Zhang *et al.*, 2021; Zuo and Chen, 2020). These methods typically use variational autoencoders as building blocks to project the data into low-dimensional manifolds and adversarial discriminators (Stark *et al.*, 2020) to align the manifolds. Recently, GLUE (Cao and Gao, 2021) expanded the deep neural framework by incorporating prior knowledge about regulatory interactions to connect the feature spaces. LIGER (Welch *et al.*, 2019) differs from the above methods in that it employs an integrative non-negative matrix factorization approach to find the shared and dataset-specific factors across datasets in the embedded space. Finally, the UnionCom algorithm (Cao *et al.*, 2020) solves both problems: it first finds a matching between distance matrices from the two modalities and then uses that matching to induce an embedding. Table 1 summarizes the recent methods based on some of their key properties. For a good review, see Stanojevic *et al.* (2022).

Here, we present Synmatch, a discrete optimization algorithm that exploits neighborhood structure and uses supermodular optimization to find a matching of the cells from two different multi-omics datasets that do not have any features in common. The key idea behind Synmatch is that the same cell, when measured in two different modalities, is likely to have similar sets of neighboring cells in the two spaces. We use this intuition to formulate the matching problem as a supermodular optimization over the neighborhood structure of the two modalities and we solve the problem using a fast greedy heuristic that offers good theoretical guarantees. We demonstrate that Synmatch offers excellent performance in finding matchings of cells in several small single-cell multi-omics datasets, outperforming several state-of-the-art methods. We also propose an iterative procedure to allow our algorithm to scale up to datasets of thousands of cells while maintaining its excellent performance. Our work stands out from recently developed algorithms for modality integration in that it seeks direct mapping between the cells based in their shared combinatorial properties in, respectively, their own spaces rather than to find a common latent space within which affinity may be defined. Although Synmatch was designed to integrate single-cell multi-omics data, it is applicable to problems in other areas where matching of observations from different modalities is needed, as long as the main assumption—observations that are close (and have combinatorial properties) in one modality should be close (and thus have similar combinatorial properties) in the other modality—holds.

2 Materials and methods

2.1 The Synmatch algorithm

Synmatch takes as input two matrices of single-cell profiles measuring different cellular properties, such as gene expression and chromatin accessibility, and outputs a matching of the cells across the datasets. Figure 1 illustrates the key concept behind Synmatch. Two similar cells that are close in one modality likely share the same biological (and thus combinatorial) properties. [In this work, 'combinatorial properties' includes things such as neighborhood structure in graphs, but could, in general, include any graph property such as triadic closure, neighborhood reciprocity and so on (Newman, 2018).] Hence, it is likely that they are close in the other modality, reflecting the similarity of their biological properties.

We denote the two sets of measurements as $U = [u_1, u_2, \dots, u_m]^T \in \mathbb{R}^{m\times d_u}$ and $V = [v_1, v_2, \dots, v_n]^T \in \mathbb{R}^{n\times d_v}$, where $u_i \in \mathbb{R}^{d_u}$ and $v_j \in R^{d_v}$ are column vectors describing, respectively, cell i and cell j. Our goal is to find a matching between these two sets of cells, where we describe a matching as a set E' of edges in a bipartite graph between U and V: $E' \subseteq E = U \times V$. In the resulting matching, we require that no node in U or V has incident edge degree greater than 1. Hence, if $m \neq n$ then some cells in one of the two sets will be left unmatched. The Synmatch algorithm proceeds in two phases.

In the first phase, we compute a diffusion-based similarity between the cells in U and, separately, among the cells in V (so there is no direct similarity computed at this stage between any $u \in U$ and $v \in V$). This measure captures both the local and global relationships among the cells in the each modality. For the moment, we discuss only U. We use the cosine distance between any two cells u_i and u_j in U to assign a weight to the edge $e_U(i,j)$ in the complete graph $G_U = (U, U \times U)$ induced by the cells in U. We choose cosine

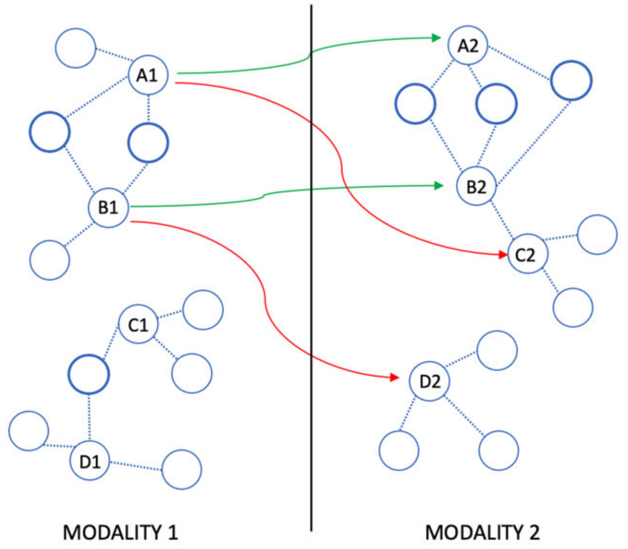

Fig. 1. Synergistic matching of neighborhood structure. Synmatch aims to match cells that share common neighbors in each data modality. In the figure, each labeled cell is connected to its three nearest neighbors by dotted edges. The two cells A1 and B1, which have neighbors in common (indicated by thicker circles) in modality 1, should be matched (green arrows) to the two cells A2 and B2, which also share neighbors in modality 2. Conversely, A1 and B1 should not be matched (red arrows) to cells C2 and D2, which do not share neighbors. In the first step of the algorithm, these common neighbors help diffusion propagate between A1–B1 and A2–B2. This in turn facilitates the optimization, which operates cooperatively on pairs of edges and aims to match pairs of cells with shared local structure across the modalities

similarity instead of Euclidean because it has been shown to be a considerably more robust measure of cell-to-cell similarity (Korsunsky *et al.*, 2019). Next, we employ a diffusion kernel (Kondor and Lafferty, 2002) to spread activation across the graph G_U. Briefly, the Laplacian of a graph G_U shifted by λ is defined as $L_U = D_U + \lambda I - A_U$, where I is the identity matrix, D_U is the diagonal matrix $d_{ii} = \sum_j e_U(i,j)$, A_U is the adjacency matrix of the graph and λ is a parameter controlling how far the activation spreads across the graph G_U. As shown in Qi *et al.* (2008), the amount of activation at equilibrium can be efficiently computed as $S_U = L_U^{-1} b$, where b is the elementary unit vector with 1 for the nodes introducing the flow and 0 for the rest. We note that this diffusion kernel has been successfully utilized in a variety of computational problems ranging from protein function prediction (Tsuda and Noble, 2004) to cancer gene identification (Hristov *et al.*, 2020). We use $S_U(u_i, u_j)$ as a measure of the *similarity* between cells u_i and u_j. We analogously compute S_V for the cells in V.

In the second phase, we construct a mapping between the cells in U and V based on S_U and S_V. We consider all pairs of cells $(u_i, u_j) \in U$ and all pairs of cells $(v_l, v_k) \in V$. Intuitively, if cells u_i and u_j are close to one another in U, then the corresponding cells in V should also be close to one another. That is, a good matching is one in which a large $S_U(u_i, u_j)$ implies a large $S_V(v_l, v_k)$ and vice versa, a property well expressed by the square root of the product, that is, $\sqrt{S_U(u_i,u_j)S_V(v_l,v_k)}$. A large value of this product thereby provides evidence not only that the cells (u_i, v_l) should be matched but also that the cells (u_j, v_k) should be matched. The second part of our objective, in fact, expresses a form of complementarity between matched edges. Any given matched pairs of cells, in the form of an edge, say $(i, l) \in E$, should offer benefit to all other cell pairs $(j, k) \in E'$ commensurate with the tendency of the corresponding cells (j, k) to be close whenever (i, l) is close. This property is expressed precisely using an objective $g(E') = \sum_{e(i,l)\in E', e(j,k)\in E'} \sqrt{S_U(u_i,u_j)S_V(v_l,v_k)}$ that judges the quality of the set of edges E' being considered in a match. We note that $g(E')$ is a set function objective that scores any $E' \subseteq E$ and in fact is a well-known supermodular objective (Bilmes, 2022). Of course not all subsets $E' \subseteq E$ are valid matchings, so this leads us to a constrained optimization problem of the form:

Table 1. Methods for unsupervised multi-model data integration

Method	Manifold alignment	Cell matching	Optimal transport	Neural net model	Matrix factorization	Discrete optimization	Reference
MAGAN		✓		✓			(Amodio and Krishnaswamy, 2018)
LIGER	✓				✓		(Welch *et al.*, 2019)
MMD-MA	✓						(Singh *et al.*, 2020)
UnionCom	✓	✓				✓	(Cao *et al.*, 2020)
SCIM	✓			✓			(Stark *et al.*, 2020)
scMVAE	✓			✓			(Zuo and Chen, 2020)
SCOT	✓		✓				(Demetci *et al.*, 2022)
Pomona	✓		✓				(Cao *et al.*, 2022a)
uniPort	✓		✓				(Cao *et al.*, 2022b)
scDART	✓			✓			(Zhang *et al.*, 2021)
GLUE	✓			✓			(Cao and Gao, 2021)
Synmatch		✓				✓	

$$\max_{E' \subseteq E: E' \in \mathcal{C}} \sum_{e(i,l) \in E', e(j,k) \in E'} \sqrt{S_U(u_i, u_j) S_V(v_l, v_k)},$$

where $\mathcal{C}$ is the constraint that the edges E' must form a bipartite matching. Since the function being optimized is supermodular, and as long as the diagonal is not zero, we can efficiently maximize it using a greedy heuristic which iteratively adds to E' the edge that improves the objective function the most while maintaining the bipartite constraints. While supermodular maximization subject to matroid constraints is normally hard, this algorithm has a theoretical approximation guarantee (Bai and Bilmes, 2018) depending on the diagonal component of the implicitly expressed $|E| \times |E|$ matrix and depending on the curvature of the supermodular function.

2.2 Scaling Synmatch to large numbers of cells

Because our approach needs to examine all $O(m^2n^2)$ possible pairs of edges, it does not immediately scale to thousands of cells due to memory constraints. In practice, Synmatch can easily run on a personal computer if $|U| \leq 300$ and $|V| \leq 300$ or $|U||V| \leq 10\,000$. For larger datasets, we aggregate the cells in each modality into a small number $c < 100$ of clusters, which we refer to as a 'meta-cells', using equal size k-means clustering. Then, we compute pairwise similarity matrix S_{MU} between the meta-cells in modality U. Specifically, for two meta-cells m_1 and m_2 in U, $S_{MU}(m_1, m_2) = \sum_{u_i \in m_1, u_j \in m_2} S_U(u_i, u_j)/|m_1||m_2|$. We analogously compute S_{MV}. We match the meta-cells using the Synmatch algorithm as described above (using S_{MU} and S_{MV} instead of S_U and S_V) and then recursively match individual cells within each pair of matched meta-cells. If a given pair of matched meta-cells contains more than a total of 10 000 cells and hence cannot be matched directly, then we repeat the procedure of aggregating these cells into sub-meta-cells that we proceed to match and so on.

2.3 Datasets

We use real single-cell multi-omics datasets in our analysis. All datasets are generated by co-assays; hence, we know the correct cell-to-cell correspondence for benchmarking.

The first dataset comes from the SNARE-seq assay (Chen *et al.*, 2019) (accession number GSE126074) and consists of a mixture of human cell lines (BJ, H1, K562 and GM12878). The gene expression information is stored in a cell × gene counts matrix with dimensionality $1047 \times 18\ 666$ while the chromatin accessibility information is stored in a Boolean cell × peak matrix of size $1047 \times 136\,771$. We reduce the dimensionality of the datasets in the same way as in Singh *et al.* (2020): we apply Principal Component Analysis (PCA) to the gene expression data and select the top 10 components, resulting in a 1047×10 matrix. We reduce the sparsity and noise of chromatin accessibility data by using the cisTopic (González-Blas *et al.*, 2019) framework, resulting in a 1047×19 matrix.

The second dataset is generated by the scGEM assay (Cheow *et al.*, 2016) (accession SRP077853) and simultaneously profiles gene expression and DNA methylation of human somatic cells undergoing conversion to induced pluripotent stem cells. This dataset consists of 177 cells and has dimensions 177×34 for the gene expression data and 177×27 for the chromatin accessibility data.

The third dataset is derived from the recently developed SHARE-seq assay (Ma *et al.*, 2020) (accession GSE140203). It jointly profiles chromatin accessibility and gene expression in 34 774 mouse skin cells. The unprocessed data matrices have dimensionally $34\ 774 \times 164\ 105$ and $34\ 774 \times 20\ 085$, respectively. We reduce each data matrix using PCA to two matrices of sizes $34\ 774 \times 10$ each.

2.4 Evaluation metrics

To assess the performance of each algorithm, we employ three evaluation metrics.

The FOSCTTM score is the fraction of samples closer than the true match (Liu *et al.*, 2019). For a given cell c in one modality, we identify its correct match $m(c)$ in the other modality. We then calculate the Euclidean distance between all of the cells in the other modality to $m(c)$ and we compute the fraction of them that are closer to $m(c)$ than the predicted match $p(c)$. The final score is the average of this fraction across all data points in both domains. Lower scores are better, with a score of 0 reflecting a perfect matching.

The neighborhood overlap score quantifies the percentage of all cells whose correct match lies within a given size neighborhood of the cell they have been matched to (Stanley *et al.*, 2020). Specifically, if a cell c is matched to cell $p(c)$, then a neighborhood of fixed size $k = 0, 1, 2, \ldots, n$ around $p(c)$ is examined to ascertain whether it contains the correct match $m(c)$. For each k, the average of all cells from each modality for which this condition is true is reported. The score ranges from 0 to 100%, with a higher percentage being indicative of a better recovery of the cell-to-cell relationship between the two datasets.

Unlike the previous two scores, the third score, label transfer accuracy, does not require knowing the correspondence between cells in the two domains. Instead, label transfer accuracy makes use of cell type label information. This score aims to assess the ability to correctly transfer cell labels from one domain to another based on the predicted matching. As in Cao *et al.* (2020), we train a k-nearest neighbor classifier (with $k = 5$) on one of the modalities and we use it to predict the cell labels in the other modality. The label transfer accuracy is the percentage of cells with correctly predicted labels. This score ranges from 0 to 100%, with a higher percentage being indicative of better performance.

2.5 Hyperparameter tuning

In our analysis, we compare Synmatch with three state-of-the-art single-cell alignment methods, none of which uses any correspondence information: SCOT (Demetci *et al.*, 2022), UnionCom (Cao *et al.*, 2020) and MMD-MA (Singh *et al.*, 2020). We run each competing method over grid of its hyperparameters (trying to keep the grids about the same size of 120 points) selecting the hyperparameters that yield the lowest average FOSCTTM score. SCOT has two hyperparameters: regularization weight $\epsilon \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$ and number of neighbors $k \in \{10, 20, 30, 40, 60, 80, 100, 200, 500, 1000\}$; MMD-MA 3: weights $\lambda_1, \lambda_2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ and dimensionally $p \in \{4, 5, 6, 16, 32\}$; UnionCom 4: trade-off $\beta \in \{0.1, 1, 10, 20\}$, regularization coefficient $\rho \in \{0, 5, 10, 15, 20\}$, dimensionally $p \in \{4, 6, 16, 32\}$ and $k_{\max} \in \{40, 100\}$. We note that the performance of these methods greatly depends on the choice of parameters and the ones provided by default achieve significantly worse performance than optimal.

3 Results

3.1 Synmatch improves cell matching on small single-cell multi-omics datasets

First, we investigate performance of our method when it constructs a matching between cells in small datasets, when no clustering into meta-cells is necessary. We run Synmatch on scGEM co-assay data, which profile gene expression and DNA methylation in 177 human somatic cells. This dataset was previously used to showcase the performance of the UnionCom algorithm (Cao *et al.*, 2020). For comparison, we also run three state-of-the-art methods—SCOT, UnionCom and MMD-MA—on the same dataset. To judge performance, we employ three metrics: (i) the neighborhood overlap (Cao *et al.*, 2020), which is defined as the percentage of cells that can find their corresponding cells from the other dataset in a neighborhood of a given size around the cells that the algorithm matches them to, (ii) label transfer accuracy (Johansen and Quon, 2019), which measures how well cell type labels are transferred from one dataset to another and (iii) the FOSCTTM score, which quantifies the fraction of samples closer than the true match (Singh *et al.*, 2020) (see Section 2.4 for details). Hyperparameters for all methods were selected by minimizing the FOSCTTM score over a predefined grid (Section 2.5).

Our results show that Synmatch outperforms all three methods in finding the correct matching between cells across modalities. Synmatch achieves the best (i.e. lowest) FOSCTTM score of 0.19 compared with 0.20 for SCOT, 0.22 for MMD-MA and 0.23 for UnionCom. Synmatch also performs well according to the neighborhood overlap score, where it exhibits a higher score than the competing algorithms across neighborhoods of size < 50 (Fig. 2A). For larger neighborhoods, all four methods perform similarly. Finally, the label transfer accuracy for all four methods is similar, with a slight edge for Synmatch: Synmatch correctly transferred 60% of cell type labels from one modality to another compared with 58%, 59% and 59% for SCOT, UnionCom and MMD-MA, respectively.

Next, we ran Synmatch on subsets of cells from a SNARE-seq co-assay dataset (Chen *et al.*, 2019), which measures gene expression and chromatin accessibility. To assure robustness, we repeatedly subsampled 10 dataset of size 200 cells and we report the average performance for each algorithm. As before, we select hyperparameters by grid search, optimizing the FOSCTTM score. The neighborhood overlap for Synmatch is higher than those of the competing methods (Fig. 2B). Furthermore, Synmatch excels at correctly transferring cell type labels, improving over UnionCom on average by 0.33, over MMD-MA by 0.23 and over SCOT by 0.20 (Fig. 2C).

3.2 Synmatch successfully scales to thousands of cells

In practice, many multi-modal single-cell datasets contain thousands of cells and hence cannot be directly analyzed by Synmatch due to its memory requirements. Accordingly, we implemented and tested a recursive variant of Synmatch, which involves clustering the cells into a small number of meta-cells, matching those meta-cells with Synmatch and then matching the cells within each pair of matched meta-cells, again with Synmatch (Section 2.2). To validate the approach, we ran Synmatch on 10 random samples of 10 000 cells drawn from a SHARE-seq co-assay (Ma *et al.*, 2020), which profiles chromatin accessibility and gene expression. For the clustering step, we employed equal size K-means with $k = 250$ to group the cells into $c = 40$ meta-cells. As before, we compared Synmatch's performance with that of SCOT, MMD-MA and UnionCom and we used the same hyperparameter grid search procedure.

Synmatch performs well in this comparison. Synmatch's FOSCTTM score is the best 0.34 (0.36 for SCOT, 0.39 for MMD-MA and 0.42 for UnionCom). In terms of neighborhood overlap, Synmatch is often the best-performing method (8 of 10 neighborhood sizes that we considered) and when it is not the top-ranked method, it is always second-ranked (Fig. 3A). Synmatch also achieves the highest label transfer accuracy, exceeding the second-ranked method (SCOT) by 0.06 on average. Notably, Synmatch does a better job transferring cell type labels than MMD-MA and UnionCom in all 10 runs.

3.3 Investigating variants of the Synmatch algorithm

There are two critical components in the process of scaling Synmatch up to larger datasets: grouping the cells into meta-cells and matching the meta-cells between modalities. Accordingly, we explore these two steps in detail. We find that both of these steps

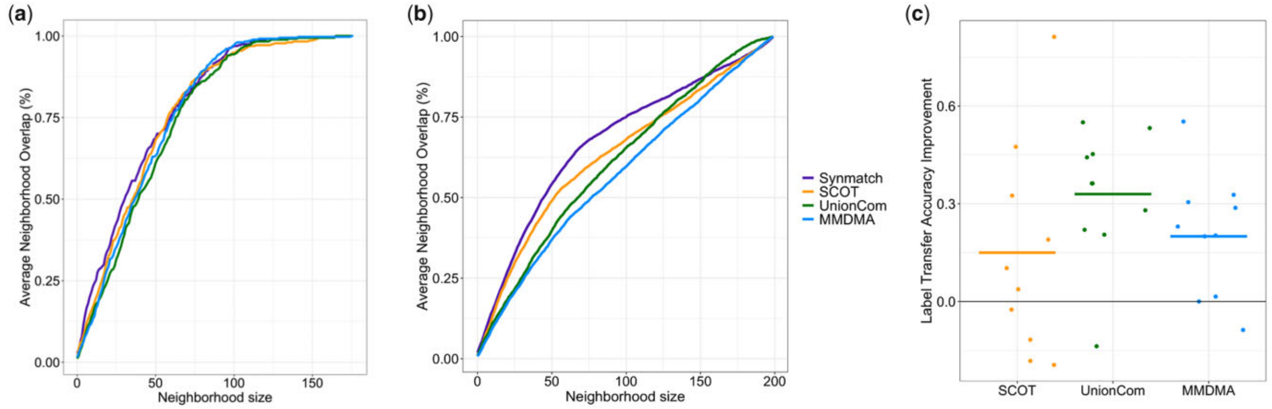

Fig. 2. Performance comparison on small datasets. (**A**) The figure plots the neighborhood overlap as a function of neighborhood size on the scGEM dataset. The four series correspond to Synmatch and three other state-of-the-art methods. (**B**) The figure plots the neighborhood overlap, averaged over 10 different datasets of size 200, drawn from the SNARE-seq dataset, as a function of neighborhood size. (**C**) The figure plots, for each of three competing methods, the difference in label transfer accuracy compared with Synmatch, with positive values representing an improvement by Synmatch. Each dot corresponds to a different randomly sampled subset of size 200 from the SNARE-seq assay

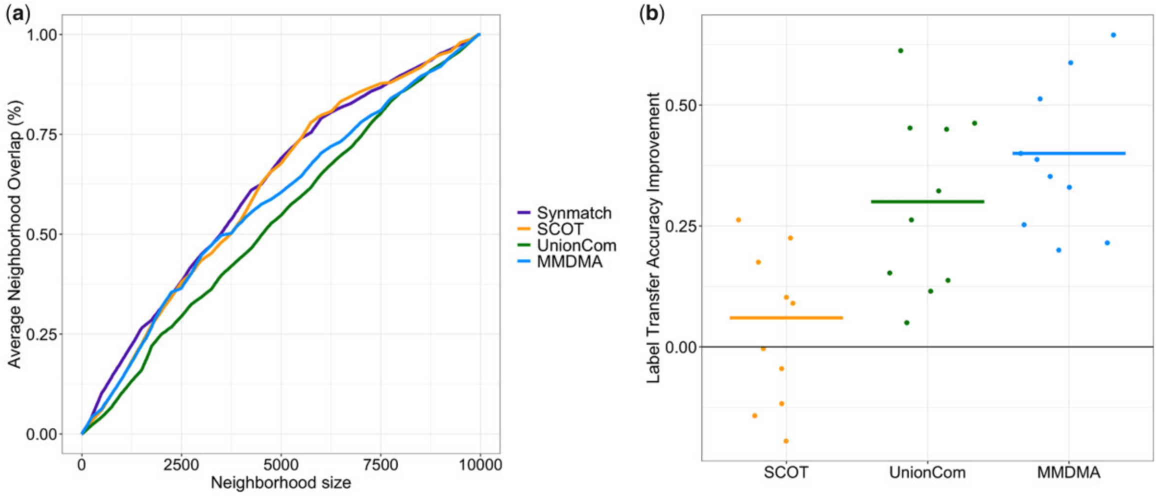

Fig. 3. Performance comparison on large dataset. (**A**) The figure plots the neighborhood overlap, averaged over 10 different datasets of size 10 000, drawn from the SHARE-seq dataset, as a function of neighborhood size. (**B**) The figure plots, for each of three competing methods, the difference in label transfer accuracy compared with Synmatch, with positive values represnting an improvement by Synmatch. Each dot corresponds to a different randomly sampled subset of size 10 000 from the SNARE-seq assay

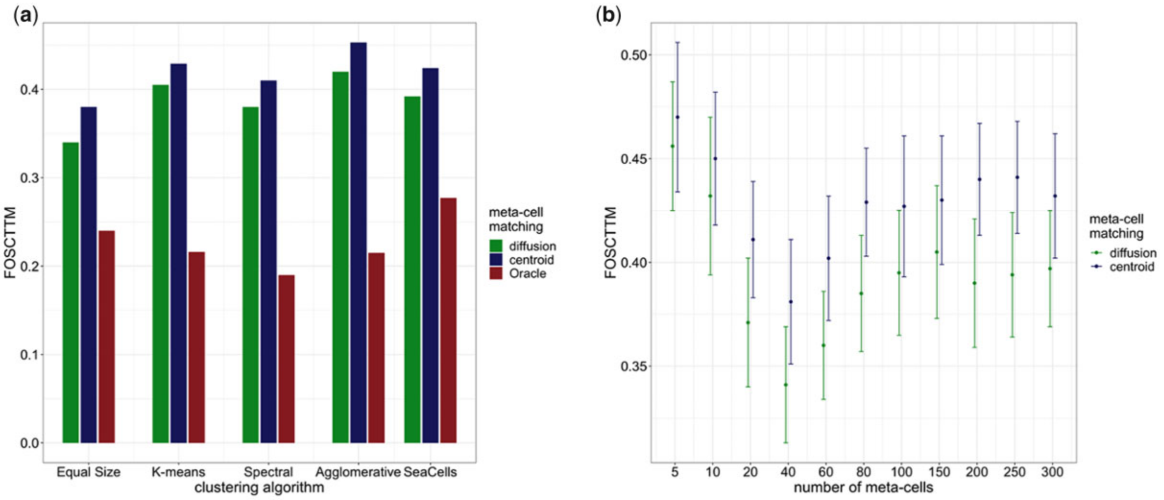

Fig. 4. Performance comparison between different clustering and meta-cell matching strategies. (**A**) The figure plots the FOSCTTM score, averaged over 10 different datasets of size 10 000 drawn from the SHARE-seq dataset, for several clustering approaches (groups of bars) with number of clusters $c = 40$ and three different meta-cell matching strategies (the color bars). (**B**) The figure plots the average FOSCTTM score over 10 different datasets of size 10 000 for various numbers of meta-cells (clusters) using equal size K-means with the two corresponding meta-cell matching strategies

have a significant impact on the performance of Synmatch and variations in either step can lead to very different results.

First, we test Synmatch using several other clustering strategies: regular K-means, agglomerative hierarchical clustering and spectral clustering. We also tested SeaCells (Persad *et al.*, 2022), a recently published method specifically designed for deriving meta-cells from single-cell data. We find that, on average, each of these clustering algorithms performs worse than equal-size k-means, when evaluated based on the FOSCTTM score (Fig. 4A). Further investigation shows that these clustering methods yield very imbalanced clusters, with some clusters containing only a handful of cells and others containing hundreds. Thus, if in the second step of our algorithm a meta-cell m_1 with less than 10 cells is matched to a meta-cell m_2 with more than 100 cells, then the subsequent matching of the individual cells from these two meta-cells will leave the majority of the cells unmatched. We attempted to resolve this problem by returning all the unmatched cells into a common pool and re-running Synmatch on them. This approach, however, leads to both a significantly slower performance and worse overall matching.

Second, we explore two different strategies to match the meta-cells to one another. The first strategy represents each meta-cell by its centroid and runs Synmatch to match the centroids. The second strategy computes an average diffusion-based similarity (S_{MU} and S_{MV}) between the meta-cells in each modality, which Synmatch then uses (Section 2.2). As an upper bound for comparison, we also include a third strategy: an oracle provides the true 1–1 cell correspondence to find the best possible match between the meta-cells. Briefly, for every pair of possible meta-cell matchings, the oracle computes the number of cells that can possibly find their correct match if two meta-cells are matched, uses it to assign weight on the edge between the two meta-cells and finally uses the Edmonds Karp algorithm to find the maximum bipartite matching between the meta-cells. The average diffusion-based strategy consistently outperforms the centroid-based one across different clustering strategies, including regular k-means (Fig. 4A, blue bars are always taller than green bars). We hypothesize that the reason is that the centroids represent a crude and imperfect center of each meta-cell in each measurement space, whereas the S_{MU} and S_{MV} matrices better capture the similarity relationships among the meta-cells. The large gap between the oracle-based and the Synmatch-based matching strategies indicates that our method could achieve significantly better performance if the meta-cells were linked more accurately.

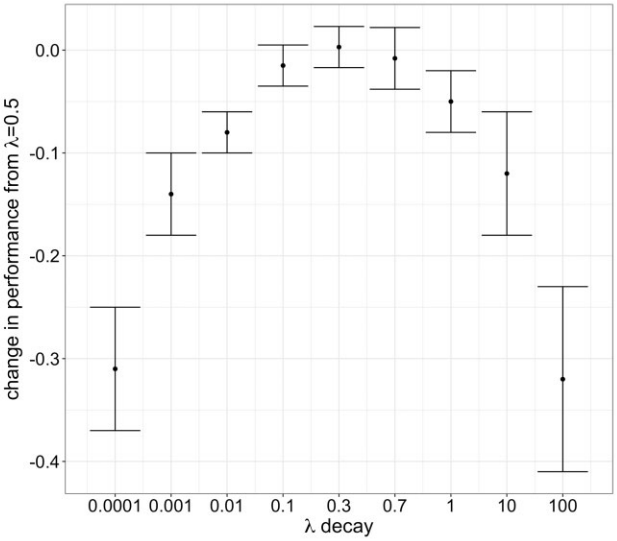

Fig. 5. Performance depends on adequately capturing neighborhood structure. The figure plots the change in performance, averaged over 10 different datasets of size 10 000, drawn from the SHARE-seq dataset, as the diffusion parameter λ is varied. The baseline is the default $\lambda = 0.5$

Third, we explore the number of meta-cells c we cluster the cells into. Because of memory constraints, we require $c \leq 300$ and test the performance of Synmatch for 10 different values of c (Fig. 4B). We observe that $c = 40$ yields the lowest average FOSCTTM score and that performance plateaus for $c > 100$. We also note that the diffusion-based linking strategy is consistently better than the centroid one for all c. Interestingly, using small number of meta-cells ($c = 5$) leads to the worst score. We suspect that the reasons for that are 2-fold. Given only a handful of cells, Synmatch cannot leverage neighborhood structure information since there are only a few neighbors possible. Further, these meta-cells are very large in size and if they are incorrectly matched this has a major negative downstream impact on the ability to correctly match the individual cells within them.

Finally, we investigate the impact of the diffusion decay parameter λ. By default we use $\lambda = 0.5$, which balances the importance of the local and global neighborhood structures. We observe that small changes in the value of this parameter ($\lambda \in (0.3, 0.8)$) do not affect significantly Synmatch's performance (Fig. 5). However, if λ is set to extreme values such as 10^{-4} or 10^{4}, the performance drops dramatically. This is expected since in both cases the diffusion does not capture the neighborhood structure—in the former it spreads almost uniformly across the cells and in the latter it is centered around a single cell.

4 Discussion

In this study we present Synmatch, an algorithm that directly maps cells across single-cell modalities that do not share any features or have any known cell-to-cell correspondence information. Synmatch exploits the neighborhood structure around the cells in each modality, seeking a matching that maps nearby cells in one modality to nearby cells in the other modality. The problem is framed as a discrete supermodular optimization and is solved efficiently. We demonstrate that Synmatch successfully matches cells in several small real single-cell multi-omics datasets and shows that it can scale to large dataset of thousands of cells. Synmatch compares favorably to state-of-the-art integration methods based on three commonly employed evaluation metrics.

From a theoretical perspective, our algorithm stands out from the majority of recently published work for two reasons: (i) it finds matching of the cells directly without the need to project the two modalities into a shared latent space and (ii) it uses a discrete optimization instead of the commonly employed optimal transport or deep learning auto-encoder-based architecture. As new tools for integration of single-cell omics data continue to emerge, those that aggregate cells into 'super-cells' or 'meta-cells' (Persad *et al.*, 2022) reflecting underlying biological properties could provide a better stepping stone for scaling up our approach.

Future work should focus on finding ways to improve the matching of the meta-cells, as our results indicate that this step has a significant impact on the overall performance. One element of our approach is that it does not immediately provide soft cell mappings, for example, when a cell in one modality is probabilistically matched to cells in the other modality. It can, however, be extended to the probabilistic case by using log-supermodular probability distributions or approaches where we exclude certain cells from a matching to arrive at score sensitivities that could be interpreted as probabilities. Our method can easily provide a many-to-one matching by relaxing the bipartite constraint to more general intersection of matroid constraints.

Funding

This paper was published as part of a special issue financially supported by ECCB2022. This work was supported by the National Institutes of Health award U01 HG009395.

Conflict of Interest: none declared.

References

Adossa,N. *et al.* (2021) Computational strategies for single-cell multi-omics integration. *Comput. Struct. Biotechnol. J.*, **19**, 2588–2596.

Amodio,M. and Krishnaswamy,S. (2018). MAGAN: aligning biological manifolds. In: Dy, J. and Krause, A. (eds), *Proceedings of the 35th International Conference on Machine Learning, Volume 80 of Proceedings of Machine Learning Research*. Stockholm, Sweden, PMLR, pp. 215–223.

Argelaguet,R. *et al.* (2020) MOFA+: a probabilistic framework for comprehensive integration of structured single-cell data. *Genome Biol.*, **21**, 111.

Bai,W. and Bilmes,J. (2018) Greed is still good: maximizing monotone submodular+supermodular (BP) functions. In: *International Conference on Machine Learning*, PMLR, pp. 304–313.

Bilmes,J. (2022) Submodularity in machine learning and artificial intelligence. arXiv. 2022.00132.

Cao,K. *et al.* (2020) Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, **36**, i48–i56.

Cao,K. *et al.* (2022a) Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics*, **38**, 211–219.

Cao,K. *et al.* (2022b) uniPort: a unified computational framework for single-cell data integration with optimal transport. bioRxiv. 2022.02.14.480323v1.

Cao,Z.-J. and Gao,G. (2021) Multi-omics integration and regulatory inference for unpaired single-cell data with a graph-linked unified embedding framework. bioRxiv. 2021.08.22.457275v2.

Chen,S. *et al.* (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.*, **37**, 1452–1457.

Cheow,L. *et al.* (2016) Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods*, **13**, 833–836.

Cui,Z. *et al.* (2014) Generalized unsupervised manifold alignment. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. and Weinberger, K. Q. (eds) *Advances in Neural Information Processing Systems*. Vol. **27**. Curran Associates, Inc., Montreal, Canada, pp. 2429–2437.

Demetci,P. *et al.* (2022) SCOT: single-cell multi-omics alignment with optimal transport. *J. Comput. Biol.*, **29**, 3–18.

González-Blas,C.B. *et al.* (2019) cisTopic: cis-regulatory topic modelling on single-cell ATAC-seq data. *Nat. Methods*, **16**, 397–400.

Hristov,B.H. *et al.* (2020) A guided network propagation approach to identify disease genes that combines prior and new information. In: *International Conference on Research in Computational Molecular Biology*. Springer, pp. 251–252.

Johansen,N. and Quon,G. (2019) scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.*, **20**, 1–21.

Kondor,R.I. and Lafferty,J. (2002). Diffusion kernels on graphs and other discrete input spaces. In: Sammut, C. and Hoffmann, A. (eds) *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann.

Korsunsky,I. *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, **16**, 1289–1296.

Liu,J. *et al.* (2019) Jointly embedding multiple single-cell omics measurements. In: Huber, K.T. and Gusfield, D. (eds) *19th International Workshop on Algorithms in Bioinformatics (WABI 2019), Volume 143 of Leibniz International Proceedings in Informatics (LIPIcs)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 10:1–10:13. PMC8496402.

Ma,S. *et al.* (2020) Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, **183**, 1103–1116.

Newman,M. (2018) *Networks*. Oxford University Press.

Persad,S. *et al.* (2022) Seacells: Inference of transcriptional and epigenomic cellular states from single-cell genomics data. bioRxiv.

Qi,Y. *et al.* (2008) Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.*, **18**, 1991–2004.

Singh,R. *et al.* (2020) Unsupervised manifold alignment for single-cell multi-omics data. *ACM BCB*, pp. 1–10. https://doi.org/10.1145/3388440.3412410.

Stanley,J.S. III *et al.* (2020) Harmonic alignment. In: *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, pp. 316–324.

Stanojevic,S. *et al.* (2022) Computational methods for single-cell multi-omics integration and alignment. arXiv. arXiv:2201.06725.

Stark,S.G. *et al.*; Tumor Profiler Consortium. (2020) SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics*, **36**, i919–i927.

Tsuda,K. and Noble,W.S. (2004) Learning kernels from biological networks by maximizing entropy. *Bioinformatics*, **20**, i326–i333.

Welch,J.D. *et al.* (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.

Zhang,Z. *et al.* (2021) Learning latent embedding of multi-modal single cell data and cross-modality relationship simultaneously. bioRxiv. 2021.04.16. 440230v2.

Zuo,C. and Chen,L. (2020) Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief. Bioinform.*, **22**, bbaa287.