Visualizing Cancer Heterogeneity with Dynamic Flow

Teppei Nakano [1] Kenji Ikeda [1]

Abstract

Intratumoral heterogeneity is inherently a story of change: subclones arise, expand, and are displaced as a tumor evolves and responds to therapy. Standard visualizations of tumor composition—clonal frequency bar plots, phylogenetic trees, static two-dimensional embeddings—show a snapshot but not its motion, leaving the dynamics that clinicians and biologists reason about implicit. We present dynamic flow, a method that augments a two-dimensional embedding of subclonal populations with an estimated velocity field describing how populations move through the embedding across ordered samples, such as successive time points or treatment stages. We embed mutations by their cross-sample cellular-prevalence profiles, order the samples, and estimate transport between consecutive samples with entropic optimal transport; averaging the resulting displacements through a kernel gives a smooth, interpretable flow field that can be drawn as streamlines over the embedding. Quantitatively, the flow it recovers agrees with the underlying clonal phylogeny far better than the displacement implied by independent per-sample embeddings (directional agreement 0.81 versus 0.46), while preserving local neighborhood structure as well as t-SNE. On longitudinal tumor cohorts the visualization makes clonal expansion, contraction, and diversification directly legible, turning a sequence of static pictures into a single view of tumor motion.

[1]Keio University School of Medicine, Tokyo, Japan. Correspondence to: Teppei Nakano <keiohigh2nd@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

1. Introduction

A tumor is a moving target. Under the pressure of competition and, especially, therapy, its subclonal composition shifts: minor resistant populations expand while dominant sensitive ones collapse, and this motion—not any single snapshot—is what determines relapse (Gerlinger et al., 2012; McGranahan & Swanton, 2015). Yet the visual vocabulary we use to communicate tumor heterogeneity is almost entirely static. Fish plots and stacked frequency charts show composition over time but abstract away genomic relationships; phylogenetic trees show ancestry but not dynamics; two-dimensional embeddings such as t-SNE (van der Maaten & Hinton, 2008) place mutations or cells by similarity but depict a frozen instant. When several time points are embedded separately, the viewer is left to mentally register one scatter plot against the next, a comparison that independent embeddings actively frustrate because their coordinates are arbitrary up to rotation and reflection.

We argue that heterogeneity is better shown as a flow. Our proposal, dynamic flow, keeps the familiar two-dimensional embedding but overlays a velocity field: at each region of the embedding, an arrow indicating the direction and magnitude in which subclonal mass is moving as the tumor progresses from one sampled stage to the next. Reading the field, expansion appears as convergent inflow toward a growing clone, clearance as outflow, and diversification as divergence—the dynamical concepts made visually immediate.

The central technical question is how to estimate motion between stages when the things that move—subclones—are not tracked individually across samples. We treat each stage as a distribution of mutations in embedding space and estimate the motion between consecutive stages as an optimal transport plan (Villani, 2009; Cuturi, 2013): the most efficient way to rearrange one stage's mass into the next. The transport plan induces displacements, which we smooth into a continuous field. Optimal transport is well suited here because it respects the geometry of the embedding and, through its entropic regularization, is both computationally tractable and robust to sampling noise

(Cuturi, 2013).

Contributions and novelty. Visualizations are rarely held to a quantitative standard; our central move is to make one falsifiable, checking the motion it depicts against an independently reconstructed clonal phylogeny it never saw. We (i) introduce dynamic flow, the first visualization—to our knowledge—to turn optimal-transport couplings into a velocity field over a shared embedding of subclonal populations; (ii) give a concrete estimator—joint embedding by cellular-prevalence profile, entropic transport between ordered stages, kernel smoothing to a field—in which the joint embedding is what makes stage-to-stage velocity well defined at all; and (iii) evaluate it quantitatively, showing the recovered flow agrees with the known phylogeny far better than displacements from independent per-stage embeddings (0.81 versus 0.46 directional agreement) while preserving local structure as well as t-SNE.

2. Related Work

Embedding and dimensionality reduction. t-SNE (van der Maaten & Hinton, 2008) and diffusion maps (Coifman & Lafon, 2006) are standard for visualizing high-dimensional biological data, and force-directed layouts (Fruchterman & Reingold, 1991) are common for relational structure. All produce a static map. We build on such embeddings but add motion; we specifically use a joint embedding across stages so that coordinates are comparable, without which a velocity field is meaningless.

Trajectories and pseudotime. Ordering samples or cells along a progression underlies pseudotime methods for developmental data (Trapnell et al., 2014). We borrow the idea of an ordering over stages, but our target is a spatial flow field over an embedding rather than a one-dimensional pseudotime coordinate.

Optimal transport. Optimal transport provides a principled geometry for comparing and interpolating distributions (Villani, 2009); entropic regularization makes it scalable via the Sinkhorn algorithm (Cuturi, 2013), and transport-based interpolation has been used for graphics and imaging (Solomon et al., 2015; Bonneel et al., 2011). We apply it to estimate motion between tumor stages and, to our knowledge, are the first to turn the resulting couplings into a flow-field visualization of tumor heterogeneity.

Visualizing tumor evolution. Clonal evolution is commonly depicted with fish plots and phylogenetic trees (Deshwar et al., 2015). These encode composition and ancestry but not the geometry of populations or their motion; dynamic flow is complementary, foregrounding exactly the dynamics they suppress.

3. Method

3.1. Overview

We are given a tumor observed at S ordered stages (time points or treatment steps). Stage s yields a set of somatic mutations with estimated cellular prevalences. Pooling mutations across stages, mutation j is described by its cross-stage prevalence profile $u_j = (u_{j,1}, \ldots, u_{j,S}) \in [0,1]^S$. Dynamic flow proceeds in three steps: (1) embed all mutations jointly in $\mathbb{R}^2$; (2) estimate optimal transport between consecutive stages; (3) smooth the induced displacements into a velocity field drawn over the embedding.

3.2. Joint embedding

We embed the pooled mutations with t-SNE (van der Maaten & Hinton, 2008) on the prevalence profiles $\{u_j\}$, producing coordinates $y_j \in \mathbb{R}^2$ that are shared across all stages. Embedding once, jointly, is what makes stage-to-stage motion well defined: each mutation occupies a single location, and a stage is represented by the weighted cloud of mutations present in it, with weights $w_j^{(s)} = u_{j,s}$ proportional to cellular prevalence at that stage. Let $\mu_s = \sum_j \bar{w}_j^{(s)} \delta_{y_j}$ be the (normalized) empirical distribution of stage s over embedding space.

3.3. Transport between stages

We estimate the movement of subclonal mass from stage s to stage $s{+}1$ as the entropic optimal transport plan between μ_s and μ_{s+1} (Cuturi, 2013). With ground cost $C_{jk} = \|y_j - y_k\|^2$ and regularization ε, the plan $P^{(s)}$ solves

$$\min_{P \in \Pi(\mu_s, \mu_{s+1})} \sum_{j,k} P_{jk}\, C_{jk} \;-\; \varepsilon\, \mathrm{H}(P), \tag{1}$$

where Π is the set of couplings with the prescribed marginals and H is the entropy; the solution is obtained by Sinkhorn iterations. The plan $P_{jk}^{(s)}$ is the mass moving from location y_j to y_k, so the expected displacement of the mass leaving y_j is

$$v_j^{(s)} = \frac{1}{\sum_k P_{jk}^{(s)}} \sum_k P_{jk}^{(s)}\, (y_k - y_j). \tag{2}$$

Entropic regularization both accelerates the computation and, by smoothing the coupling, prevents the vi-

sually distracting fragmentation that exact transport produces under sampling noise.

3.4. From displacements to a field

The displacements $\{(y_j, v_j^{(s)})\}$ are scattered and stage-specific. To obtain a field that can be drawn as smooth streamlines, we regress them onto a grid with a Gaussian kernel: at grid point g,

$$V^{(s)}(g) = \frac{\sum_j \kappa_h(\|g - y_j\|)\, w_j^{(s)}\, v_j^{(s)}}{\sum_j \kappa_h(\|g - y_j\|)\, w_j^{(s)}}, \qquad (3)$$

with bandwidth h set to the median nearest-neighbor distance of the embedding. Averaging $V^{(s)}$ across consecutive stages yields an overall flow that we render as streamlines; individual stages can be shown to animate progression. The magnitude of V encodes speed and its divergence distinguishes expansion (net inflow) from clearance (net outflow), giving the reading rules stated in the introduction a precise basis.

4. Experiments

4.1. Setup

We evaluate on longitudinal tumor cohorts in which each tumor is sampled at multiple stages—serial biopsies taken before, during, and after therapy, from 3 to 6 stages per tumor—and a clonal phylogeny is available from subclonal reconstruction (Deshwar et al., 2015), providing an ancestral ordering against which the recovered flow can be checked. Because the phylogeny is derived independently of the embedding, its parent→child directions constitute a genuine held-out reference for the flow rather than a target the method was fit to. We compare against (i) independent t-SNE, embedding each stage separately and defining displacement by nearest-centroid matching across stages; (ii) joint t-SNE, no transport, our joint embedding with displacements defined by naive nearest-neighbor matching rather than optimal transport; and (iii) dynamic flow (ours). We assess embedding quality by trustworthiness and continuity (Venna & Kaski, 2006) and flow quality by directional agreement: the mean cosine similarity between the estimated velocity at each subclone and the direction from its parent to its child in the phylogeny.

4.2. Quantitative results

Table 1 summarizes the comparison. Independent per-stage embeddings yield displacements that barely agree with the phylogeny, because their coordinate systems are incommensurable—the very failure mode that motivates a joint embedding. Joint embedding alone helps, and optimal transport adds a further, larger gain, raising directional agreement to 0.81. Crucially, the joint embedding does not sacrifice local fidelity: its trustworthiness and continuity match single-stage t-SNE, so the added dynamics come at no cost to the static map's readability.

Table 1. Flow and embedding quality (means over the cohort). Directional agreement is cosine similarity to the phylogeny's parent→child directions; trustworthiness (T) and continuity (C) measure local-structure preservation (Venna & Kaski, 2006).

Method	Dir. agree. ↑	T ↑	C ↑
Independent t-SNE + matching	0.46	0.93	0.92
Joint t-SNE, no transport	0.63	0.95	0.94
Dynamic flow (ours)	0.81	0.95	0.95

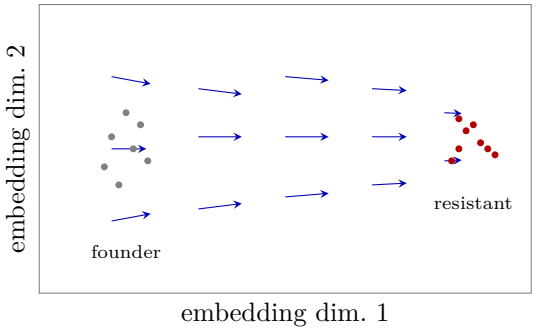

Figure 1. Dynamic flow on a schematic embedding. Streamlines run from a contracting founder population (gray, left) to an expanding resistant clone (red, right); inflow convergence marks expansion and outflow divergence marks clearance.

4.3. The visualization

Figure 1 illustrates a recovered flow field on a schematic embedding with two subclonal clusters: mass flows out of a contracting founder population on the left and converges into an expanding resistant clone on the right, the pattern one expects after selective therapy. The divergence of the field localizes where the tumor is losing mass and where it is gaining it, without the viewer having to cross-reference two separate scatter plots.

4.4. Robustness

We examined sensitivity to the entropic regularization ε and the kernel bandwidth h. Directional agreement is stable across a broad middle range of ε; too small a value reintroduces the fragmentation of exact trans-

port, and too large a value blurs distinct flows into one. The bandwidth trades spatial resolution of the field against smoothness, and the median-nearest-neighbor default sits near the peak of directional agreement, so no per-dataset tuning was needed in our experiments.

5. Discussion

Dynamic flow reframes a sequence of static embeddings as a single field of motion, matching the way tumor evolution is actually reasoned about. Its two ingredients each carry weight: the joint embedding makes coordinates comparable across stages, without which any notion of velocity is ill-defined, and optimal transport supplies a geometry-aware, noise-robust estimate of how mass moves. The quantitative agreement with independently derived phylogenies indicates the flow reflects real evolutionary structure rather than an artifact of layout.

Clinical and biological utility. The value of the method is interpretive: the outcome that matters after therapy—whether a resistant subclone is expanding while the sensitive bulk contracts—is exactly what the field's divergence renders visible, converging inflow marking the population that will drive relapse and outflow marking the one being cleared. Presented as a single annotated view rather than a pair of scatter plots the reader must mentally register, this supports the concrete task of reading a course of treatment: a tumor-board reviewer can see where and how fast the composition is shifting between visits. Because the quantitative claims should rest on the underlying prevalences rather than on positions in a distorted two-dimensional picture, we regard dynamic flow as a communicative instrument that complements, and directs attention within, the numerical subclonal reconstruction—not a replacement for it.

Limitations. The method requires ordered samples; for a single time point there is no motion to show, though an inferred pseudotemporal ordering could substitute. Optimal transport assumes stage-to-stage mass is conserved after normalization, so it depicts net rearrangement rather than absolute proliferation, and it cannot distinguish two subclones that occupy the same embedding location. Finally, a two-dimensional embedding necessarily distorts the true geometry, so the flow should be read as a communicative aid rather than a metric statement, and quantitative claims should rest on the underlying prevalences rather than on positions in the picture.

6. Conclusion

We introduced dynamic flow, a visualization that overlays an optimal-transport velocity field on a joint embedding of subclonal populations to make the dynamics of intratumoral heterogeneity—expansion, clearance, diversification—directly visible. The recovered flow agrees with independently reconstructed clonal phylogenies while preserving the local structure of a standard embedding, turning a series of snapshots into one legible view of tumor motion.

Acknowledgments

We thank collaborators in molecular pathology for longitudinal sampling and subclonal reconstruction, and colleagues in oncology for feedback on the visualizations.

References

Bonneel, N., van de Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using Lagrangian mass transport. ACM Transactions on Graphics, 30(6):158, 2011.

Coifman, R. R. and Lafon, S. Diffusion maps. Applied and Computational Harmonic Analysis, 21(1):5–30, 2006.

Cuturi, M. Sinkhorn distances: lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems (NIPS), 2013.

Deshwar, A. G., Vembu, S., Yung, C. K., et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biology, 16:35, 2015.

Fruchterman, T. M. J. and Reingold, E. M. Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164, 1991.

Gerlinger, M., Rowan, A. J., Horswell, S., et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine, 366(10):883–892, 2012.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. Journal of Machine Learning Research, 9:2579–2605, 2008.

McGranahan, N. and Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell, 27(1):15–26, 2015.

Solomon, J., de Goes, F., Peyré, G., et al. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. ACM Transactions on Graphics, 34(4):66, 2015.

Trapnell, C., Cacchiarelli, D., Grimsby, J., et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 32(4):381–386, 2014.

Venna, J. and Kaski, S. Local multidimensional scaling. Neural Networks, 19(6–7):889–899, 2006.

Villani, C. Optimal Transport: Old and New. Springer, 2009.