Contrastive trajectories inference

Given a multi-dimensional population dataset, the inference of contrasted pseudotemporal trajectories (and an individual pseudotime value) consists of three main steps:

(i) For high-dimensional datasets (e.g. ~40 000 transcripts), initial selection of features most likely to be involved in a trajectory across the entire population. We apply the unsupervised method proposed by Welch *et al.* (2016), which does not require prior knowledge of features involved in the process or differential expression analysis. Features are scored by comparing sample variance and 'neighbourhood variance'. Specifically, for a gene transcript g, its sample variance σ_g^2 across all samples

$$C_i = 100 \cdot \sum_{j=1}^{N_{cPC}} \left(\lambda_j^{norm} \cdot \frac{\omega_{i,j}^2}{\sum_{k=1}^{N_{genes}} \omega_{i,j}^2} \right) \quad (2)$$

where $\lambda_j^{norm} = (\lambda_j - \min\_\lambda) / \sum_{k=1}^{N_{total}} (\lambda_k - \min\_\lambda)$ is the normalized eigenvalue of the contrasted principal component j, min_λ is the minimum obtained eigenvalue, N_{total} is the original number of contrasted principal components, N_{cPC} is the number of contrasted principal components with λ_j^{norm} over a predefined cut-off value (i.e. 0.025), $\omega_{i,j}$ is the loading/weight of the gene transcript i on the component j, and $N_{features}$ is the total number of gene transcripts considered in the dimensionality reduction analysis. Similarly, the expected contribution value (cut-off) was calculated as in Abdi and Williams (2010):

$$C_{expected} = 100 \cdot \sum_{j=1}^{N_{cPC}} \left(\lambda_j^{norm} \cdot \frac{1}{N_{features}} \right) \quad (3)$$

The gene transcripts with total contribution C_i over the expected contribution value $C_{expected}$ were considered as highly influential to obtain the reduced representation space.

is calculated. Then, the 'neighbourhood variance' is computed as:

$$S_g^{2(N)} = \frac{1}{N_{transcripts} k_c - 1} \cdot \sum_{i=1}^{N_{genes}} \sum_{j=1}^{k_c} \left(e_{ig} - e_{N(ij)g}\right)^2 \quad (1)$$

where $N_{transcripts}$ is the total number of gene transcripts, e_{ij} is the expression level of the j^{th} transcript in the i^{th} sample, $N(i, j)$ is the j^{th} nearest neighbour of sample i, and k_c is the minimum number of neighbours needed to yield a connected graph. $S_g^{2(N)}$ is similar to the sample variance computed with respect to neighbouring points rather than the mean, measuring how much g varies across neighbouring samples. Intuitively, gene transcripts most likely to be involved in a trajectory should present a more gradual variation across neighbouring points than at global scale, which would correspond to a high ratio $\sigma_g^2/S_g^{2(N)}$. Thus, a threshold is applied to select those features with higher $\sigma_g^2/S_g^{2(N)}$ score, e.g. we kept the features with at least a 0.95 probability of being involved in a trajectory (i.e. ~3000 gene transcripts).

(ii) Data exploration and visualization via contrastive principal component analysis (cPCA) (Abid *et al.*, 2018). This novel technique identifies low-dimensional patterns that are enriched in a target dataset (e.g. a diseased population) relative to a comparison background dataset (e.g. demographically matched healthy subjects). By controlling the effects of characteristic patterns in the background (e.g. pathology-free and spurious associations, noise), cPCA (Abid *et al.*, 2018) allows visualizing specific data structures missed by standard data exploration and visualization methods (e.g. PCA, Kernel PCA). Specifically, if C_{target} and $C_{background}$ are the covariance matrices of the target and background data, the directions returned by cPCA are the singular vectors of the weighted difference of the co-variance matrices: $C_{target} - \alpha \cdot C_{background}$. The contrast parameter α represents the trade-off between having the high target variance and the low background variance. Multiple values of α are used (i.e. 100 logarithmically equally spaced points between 10^{-2} and 10^2). Instead of choosing a single α, the resulting subspaces for all the α-values are clustered (based in their proximity in terms of the principal angle and spectral clustering) (Ng *et al.*, 2002) in a few subspaces. The data are then projected onto each of these few subspaces, revealing different trends within the target data. While the original cPCA algorithm proposes to select the final subspace via visual examination, we chose automatically the subspace that maximizes the clustering tendency in the projected target data. For this, the 'gap' cluster evaluation criterion, implemented in the MATLAB function *evalclusters*, was used. When cPCA was applied to the selected gene expression transcripts [from step (i)], for each population, we obtained about six to eight contrasted principal components capturing the most enriched pathological properties relative to the background (i.e. subjects without cognitive deterioration and neuropathological signs). For ROSMAP, HBTRC and ADNI, sample sizes of the background populations were 177 (36%), 173 (23%) and 113 (15%), respectively. Selected α-values for these three studied datasets were: 11.76 (ROSMAP), 17.07 (HBTRC) and 11.76 (ADNI).

(iii) Subject ordering and gene expression-pseudotime calculation according to their proximity to the background population in the contrasted principal components space. For this, we first calculated the Euclidean distance matrix among all the subjects and the associated minimum spanning tree (MST). The MST was then used to calculate the shortest trajectory/path from any subject to the background subjects. Each specific trajectory consists of the concatenation of relatively similar subjects, with a given behaviour in the data's dimensionally reduced space. The position of each subject in his/her corresponding shortest trajectory reflects the individual proximity to the pathology-free state (the background) and, if analysed in the inverse direction, to advanced disease state. Thus, to quantify the distance to these two extremes (background or disease), an individual gene expression-pseudotime score is calculated as the shortest distance value to the background's centroid, relative to the maximum population value (i.e. values are standardized between 0 and 1). Finally, the subjects are ordered according to their gene expression-pseudotime values, from low (close to the background group) to high values (close to the most diseased subjects).

Additionally, to evaluate cPCA's performance versus other popular dimensionality reduction techniques, we repeated step (ii) using the traditional PCA (Abdi and Williams, 2010) and the recently proposed non-linear Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) approach (McInnes *et al.*, 2018). Subsequently, we reapplied step (iii), obtaining alternative subject orderings (and gene expression-pseudotimes) according to their proximity to the background population in the resulting PCA and UMAP components space.

Statistics

Data preprocessing

Before applying the contrastive trajectory inference (cTI) approach, each gene transcript's activity was adjusted for relevant covariates using robust additive linear models (Street *et al.*, 1988). Specifically, Dataset 1 gene expression was adjusted for post-mortem interval (PMI) in hours, age, gender and educational level. Dataset 2 gene expression was adjusted for PMI, sample pH, RIN, age and gender. Dataset 3 gene expression was controlled for RIN, plate number, age, gender and educational level. Also, each adjusted gene transcript activity was approximately transformed into a normal distribution via the Box-Cox transformation (Box and Cox, 1964), implemented in the MATLAB function *boxcox*.

Post hoc analyses

All predictive associations between grouping variables (e.g. Braak, CERAD and Vonsattel stages, clinical diagnosis) and the individual gene expression-pseudotimes were tested with ANOVA tests, familywise error (FWE)-controlled by permutations (Legendre and Legendre, 1998). For each dataset, the total contribution C_i of each gene transcript i to the obtained reduced representation space (and the genetic trajectories) was quantified as in Abdi and Williams (2010):