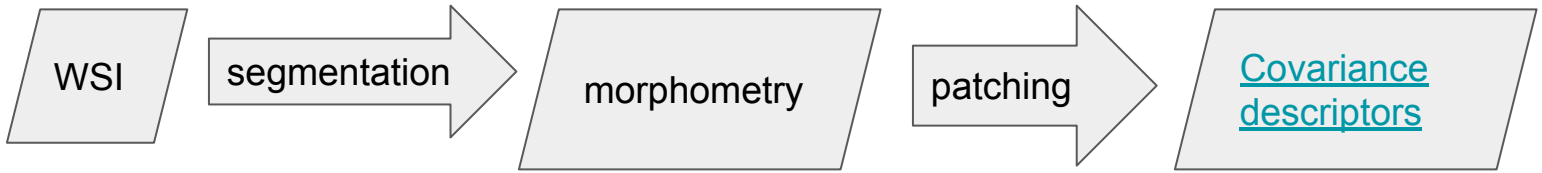
NucleAI

documentation

The idea



1. Segment a WSI to estimate nuclei locations and morphometrics (or work on the given polygonal dataset from a previous segmentation)
2. Group the nuclei into groups of proximal nuclei (this is parameter dependent)
3. For each group evaluate the covariance matrix of the morphological features of the nuclei contained in the patch

Step 1

1. One can segment the WSI or obtained pre-segmented data
2. In the case of pre-segmented data the downloaded dataset is formatted as a polygonal dataset (for each mask they provide the vertices of the polygon covering the mask)
   - Full path on garner1-P920: /media/garner1/hdd2/TCGA_polygons
3. Each slide is divided into rectangular patches, and each patch has its own polygonal dataset
   - /media/garner1/hdd2/TCGA_polygons/BRCA/TCGA-05-4245-01Z-00-DX1...svs.tar.gz/TCGA-05-4245-01Z-00-DX1...svs.tar.gz
4. The pipeline process the rectangular patches in parallel, to generate a list of morphological features for each of them
5. The output is stored in the data/features folder
   - path_to_repo/data/features/BRCA/TCGA-3C-AALI-01Z-00-DX1/...features.csv.pkl

Step 2

1. Since each WSI contains many nuclei, to speed up the computation, we subsample a fraction of them at random
2. For each chosen random nuclei, we consider its k nearest neighbours, and for this group of k+1 nuclei we generate the covariance matrix of the morphological features
3. The output is stored in data/covds
    a. path_to_repo/data/covds/BRCA/TCGA-3C-AALI-01Z-00-DX1/nuclei758121.numbCovd75812.freq10.covdNN50.pkl
    b. The file name contains information about the numb of nuclei in the WSI (758121), the numb of groups in which the nuclei have been divided (75812), the frequency at which we have subsampled the nuclei (1 every 10 nuclei), the numb of nearest neighbors around each sample nuclei that have been considered to generate the covariance matrix, ie the size of each group (50 nuclei)

Step 3

1. Each group of k+1 nuclei is an irregular (ie non rectangular) patch on the WSI
2. The covariance descriptors of all these irregular patches are stored in path_to_repo/data/covds

Data storage

- For convenience we have placed all features and descriptors in
  - /media/garner1/hdd2/features
  - /media/garner1/hdd2/covds
- Each directory contains subdirectories for each cancer type and sample ID
  - /media/garner1/hdd2/**features**/BRCA/TCGA-3C-AALI-01Z-00-DX1
  - /media/garner1/hdd2/**covds**/BRCA/TCGA-3C-AALI-01Z-00-DX1

Downstream analysis

1. For a given WSI, get list of covariance descriptors of irregular patches stored in data/covds
2. Evaluate the barycenters of the covariance descriptors
3. The barycenter will be a single point in a high-dimensional space that is associated to each WSI-sample
4. For a given cancer type, we can get the point cloud of all the samples belonging to that type
5. For different cancer types we can project into low dimensions the point-cloud of barycenters of all the samples

UMAP projection of descriptor vectors

- The covd descriptor vectors live in a high dimensional space
- UMAP can be used for dimensionality reduction both at the cancer-type level and at TCGA-level
  - Cancer type level: each sample is associated to a list of descriptors (one for each group of 50 proximal nuclei in the WSI); a single descriptor per sample is generated by evaluating the barycenter of the list of descriptors (remember to keep into consideration the geometric context as explained in the original paper and using a log-Euclidean framework); UMAP is applied to the barycenter descriptors of the cancer-type samples
  - TCGA-level: umap is applied to the barycenter descriptors of all the samples in the TCGA dataset

UMAP projection data

- Cancer-type level
  - For each cancer type the output of the UMAP projection is found here:

/media/garner1/hdd2/covds/{BLCA,BRCA,CESC,...}/descriptor_withI.umap.csv

- TCGA-level
  - At the moment of preparing this note there is a bug affecting UMAP which does not allow to process all TCGA dataset

Connecting morphology and molecular data

- Construct a representation of each sample morphology, as explained before
- Each sample comes with molecular information
- Try to predict the molecular data from morphology using some machine learning model (e.g. SVM or random forests)
    - Predictor: (x,y) coordinate pairs from UMAP representation of each sample
    - Target: quartile of CNA of each sample

Directory structure of the input/ouput data

- Output
  - /media/garner1/hdd2/covds ← contains the descriptors (size: 519G)
  - /media/garner1/hdd2/features ←contains the morphological features (size: 399G)
- Input
  - /media/garner1/hdd2/svs_{BLCA,BRCA,...} ←contains the svs images (empty; each dir contains the manifest file that can be used to download the images; after processing the images are deleted)
  - /media/garner1/hdd2/TCGA_polygons ← (size: 582G)