EARLY DISEASE DETECTION THROUGH COMPUTATIONAL PATHOLOGY

by

Virginia M. Burger

Mag. Mathematik, University of Vienna, 2008

Submitted to the Graduate Faculty of
the School of Medicine in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Virginia M. Burger

It was defended on

October 31, 2013

and approved by

Dr. Chakra Chennubhotla, Assistant Professor, Department of Computational and Systems Biology

Dr. Takis Benos, Professor, Department of Computational and Systems Biology

Dr. Yang Liu, Assistant Professor, Department of Medicine

Dr. Gary Miller, Professor, Department of Computer Science, Carnegie Mellon University

Dr. Lans Taylor, Professor, Department of Computational and Systems Biology

Dissertation Director: Dr. Chakra Chennubhotla, Assistant Professor, Department of Computational and Systems Biology

EARLY DISEASE DETECTION THROUGH COMPUTATIONAL PATHOLOGY

Virginia M. Burger, PhD

University of Pittsburgh, 2013

This thesis presents computational pathology algorithms for enabling early cancer detection in Barretts Esophagus (BE) and early subtype diagnosis in Interstitial Lung Diseases (ILD). BE is a condition affecting 10% of heartburn sufferers, for which 0.1% of patients develop esophageal adenocarcinoma each year. For most of the 130-200 diseases included in the class of ILDs, a full recovery is expected, but for a few of these diseases, the survival rate is less than three years. For both disease classes, treatment of the malignant forms would be harmful in patients with other forms, thus diagnosis is necessary prior to beginning treatment, and early treatment is most effective in eradicating disease. Early diagnosis of both of these disease classes is complicated by a high degree of sharing of subtle disease phenotypes, leading to high pathologist disagreement rates. Computational pathology methods can aid early diagnosis of these diseases through unbiased, data-driven algorithms.

To detect precancerous changes in patients with BE, we develop an automated algorithm which identifies epithelial nuclei in biopsy samples on which nano-scale optical biomarkers, related to cancer risk, can be quantified. The automated nuclei detector produces a higher quality selection of epithelial nuclei than manual detection, resulting in enhanced characterization of precancerous phenotype perturbations. To stratify ILD patients, we develop a novel quantitative representation of pathohistology samples that models lung architecture based on computed image features and insights from pathologists, and establish its utility as part of a diagnostic classifier. Algorithms such as these applied in a clinical setting can save pathologists time by filtering out obvious cases and providing unbiased reasoning to assist diagnoses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The field of computational pathology began in the mid-1980s with the goal of improving diagnosis and prognosis of tissues, [6, 7]. However, computers have only recently become powerful enough to accurately analyze tissue images at a practical scale [8]. In the past few years, fast slide scanners and increased computer storage have made systematic scanning of whole slide tissue images in medical laboratories a possibility, permitting large-scale computational analysis of pathological tissue. These studies have the potential to both assist pathologists in their traditional analyses through computer-aided diagnosis and prognosis and to discover novel features relevant for disease detection [9, 10, 11]. Similar to how automated screening of pap smears to filter out clearly heathy cases allows cytologists time to focus on ambiguous images, pathologists could gain time for analysis of diagnostically challenging images by prescreening their slides with computational pathology algorithms. As imaging is cheaper than genetic testing and images can be sent rapidly over the internet, computational algorithms implemented in telepathology platforms could bring expert medical insight to populations far from major hospitals.

As intra- and inter- pathologist variability is not uncommon [9, 11], computational aided diagnosis can provide an objective, quantitative assessment of ambiguous slides [10]. Computational Pathology is expected to be useful in resolving disagreement between pathologists and providing unbiased, explicitly reasoned, as opposed to intuitive, diagnoses. A recent example of the power of computational pathology to resolve pathologist disputes is seen in [12], where a computational measurement of lymphocytic infiltration is developed. While pathologic scores on the testing cohort lost prognostic strength due to pathologist disagreement, the computational score was able to differentiate between good and poor disease outcomes.

In addition to resolving discrepancies and saving pathologist time, computational algorithms

have discovered diagnostic and prognostic features in tissue data that have revealed unknown aspects of cancer progression. In 2011, Beck et al developed an algorithm termed C-path for computational prognosis of breast cancer [13]. Using machine learning on a large feature set including epithelial and stromal nuclei, they built a classifier which was able to predict 5-year patient survival with 89% accuracy. Interestingly, three of their most important features for prognosis were based on stromal nuclei, and these three features alone formed a better model of prognosis than a model built from the most predictive epithelial nuclei. This finding was striking, as pathologist grading criteria include only epithelial features. In a response to this study, the medical field has devoted more attention to intra-tumor stroma in the past few years, and recent findings have shown that stromal features can be used as prognostic parameters in colorectal cancer, esophageal and breast [14].

Early detection of disease is critical for treatment in many systems [15]. For many cancers, prompt removal of the tumorous region will reduce chance of metastasis. Recent studies have shown that removal of the whole cancer field may be necessary to prevent tumor regrowth [16]. Barrett's esophagus [BE], a common condition in the USA for which patients have an increased, but small, risk of developing cancer each year, provides an interesting platform for studying the cancer field, that is, the region around a tumor in which precancerous changes take place [17] because of the frequency of biopsies taken from BE patients to assess their risk of developing cancer. Not only is studying cancer development in BE important for developing methods for earlier diagnosis of cancer in these patients, but it also allows provides insight into the development of cancer in general, as patients who go on to develop cancer commonly have a biopsy record for how their tissue has changed between first being diagnosed with Barrett's Esophagus and eventually developing cancer.

Computational biology has focused in the past on segments of whole slide images and on tissue microarrays. This is partially due to the large size of whole slide images, but also do to he heterogeneity of the images [13]. Images with a single diagnostic label as cancerous may contain healthy tissue as well as tissue of many grades of cancer. This noisy ground-truth information challenges classification algorithms. Additionally, computational analysis of large whole-slide images requires accurate computer vision algorithms trained to identify objects in tissues images

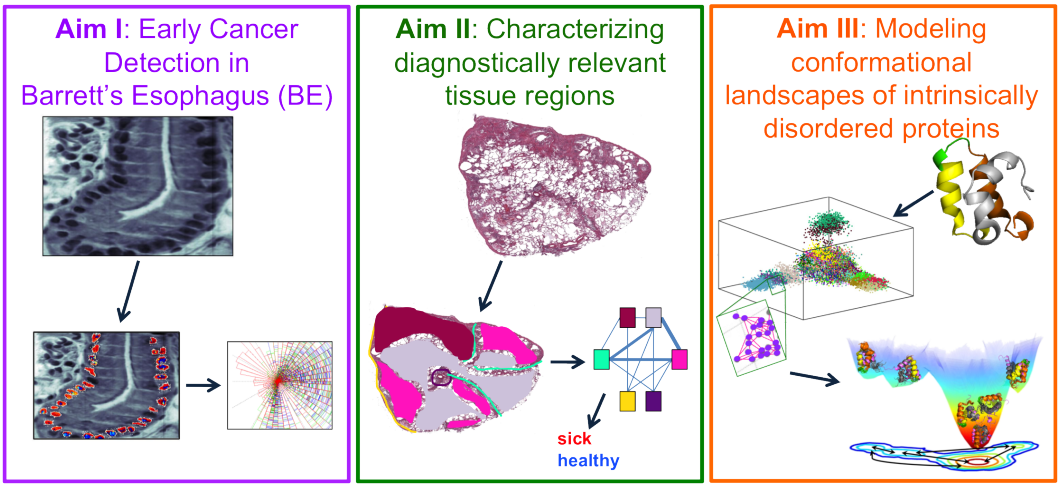

Figure 1: Overview of major aims

[18, 19].

Contributions

This thesis provides new methods for analyzing biological data on three scales (Fig. 1). On the micrometer scale, we present a method for classifying nuclei as epithelial or non-epithelial. This kind of classification is essential for analysis of cancerous tissue, as different types of features are indicative of cancer in epithelial, stromal, and lymphatic cells [13]. We demonstrate how optical phase computed on the epithelial nuclei can be used as an optical biomarker for detecting cancer in the field around a tumor, using spatial low-coherence quantitative phase microscopy [20]. On the centimeter scale, we present a method for representing heterogeneous whole slide images in terms of their spatial architecture with regards to homogeneous tissue components. We then demonstrate the use of this method for classification of interstitial lung diseases. On the nanometer scale, we present a method analogous to Aim II for determining the computational landscape of intrinsically disordered proteins, which are commonly implicated in cancers and other genetic diseases. This method can be used to elucidate bottleneck conformations in the protein's landscape, which could eventually be used as targets for therapeutic drugs.

2.0 BACKGROUND: NUCLEI SEGMENTATION

Nuclei are the smallest unit on which pathologists traditionally analyze features in histological images. Nuclei size, shape, and arrangement undergo documented changes as tissue becomes diseased. For example, in Barrett's esophagus, cells (and hence more visible nuclei) invade the lamina propria, a mucosa layer neighboring the epithelium, as cancer develops. In $20\times$ images, one pixel corresponds to 0.5 microns, so a nucleus with diameter of 10μm has on average 314 pixels. With this magnification, nuclei shape and average intensity can be used as features for analyzing images computationally. At higher magnifications, ($40\times$ indicates 0.25μm per pixel, $100\times$ indicates 0.1μm per pixel), nuclei contain many more pixels, and computational analysis of inter-nuclei features, such as symmetry and variations in intensity, are possible. However, the majority of images are scanned to only $10\times$ or $20\times$ magnification, as even $10\times$ scans can yield huge images. For example, a $1cm \times 1cm$ biopsy would yield a 10000×10000 scanned image at this magnification. Both storage and processing of such large images is challenging. Here, we discuss methods for processing these images, in terms of identifying nuclei and other cellular components.

Nuclei patterns differ naturally between cell types, organs, and diseases, and are captured differently depending on the slide preparation, staining, and imaging device, thus many system-specifc nuclei segmentation algorithms exist [13, 21, 22]. Several proprietary image segmentation methods have been released with scanners [23], but few open-source programs exist for segmentation of these large images. Hematoxylin and Eosin [H&E] images are currently more commonly analyzed computationally than fluorescent images, and several methods have been published for nuclei segmentation of these images based on active contours [24, 25, 26, 27], e.g. [28, 29, 30], as they perform well at detecting boundaries in the noisy images. For fluores-

cent images which are often challenged by low signal to noise rations, the watershed method is commonly used [31], as well as graph-based methods [32].

Any available methods, commercial included, must be fine-tuned to be as accurate on a different system as it is on the system it was designed for, and the challenge of adapting an old method to a new project often leads to researchers designing new methods for each system. In all systems, segmentation is challenged by cell density, nuclei density, overlapping nuclei, image contrast, background noise, and variations in nuclei morphology [33]. Specifically, [33] indicate that the main challenge stems from the face that tissue is a 2D section of a 3D sample. This results in nuclei being partially imaged, sectioned at odd angles, and damaged by sectioning. Additionally, the limited thickness of the section causes overlapping nuclei. In pathological samples, nuclei can have unnatural shapes and sizes, as well as variable chromatin texture. Importantly, tightly clustered nuclei and nuclei with unique morphologies are more difficult to segment than most nuclei, but also more likely to be indicative of disease [34]. The most basic methods are based on intensity thresholding, as nuclei are usually darker than their immediate surroundings [34]. However, variations in image intensity and both biological and experimental noise create many false positive nuclei. Diffusion and contrast normalization are often used to improve over intensity thresholding, followed by system specific methods to weed out false positive nuclei and break up clusters of nuclei

2.0.1 Common components of nuclei segmentation algorithms

While ideally one nuclei segmentation algorithm would be able to perfectly segment nuclei from any tissue image, variations in nuclei patterns in different tissues and using different stains cause it to be more practical to design specific algorithms for each image set. However, some standard tools are often applied as intermediate steps in nuclei segmentation algorithms like the ones mentioned above. Here, we briefly describe some of the most common image processing methods for identifying nuclei.

- Thresholding: An image is a matrix or stack of matrices filled with intensity values. Each matrix entry corresponds to an image pixel. Color images are typically stacks of three matrices, for example, one matrix for the each of the red, blue, and green channels. Gray-scale

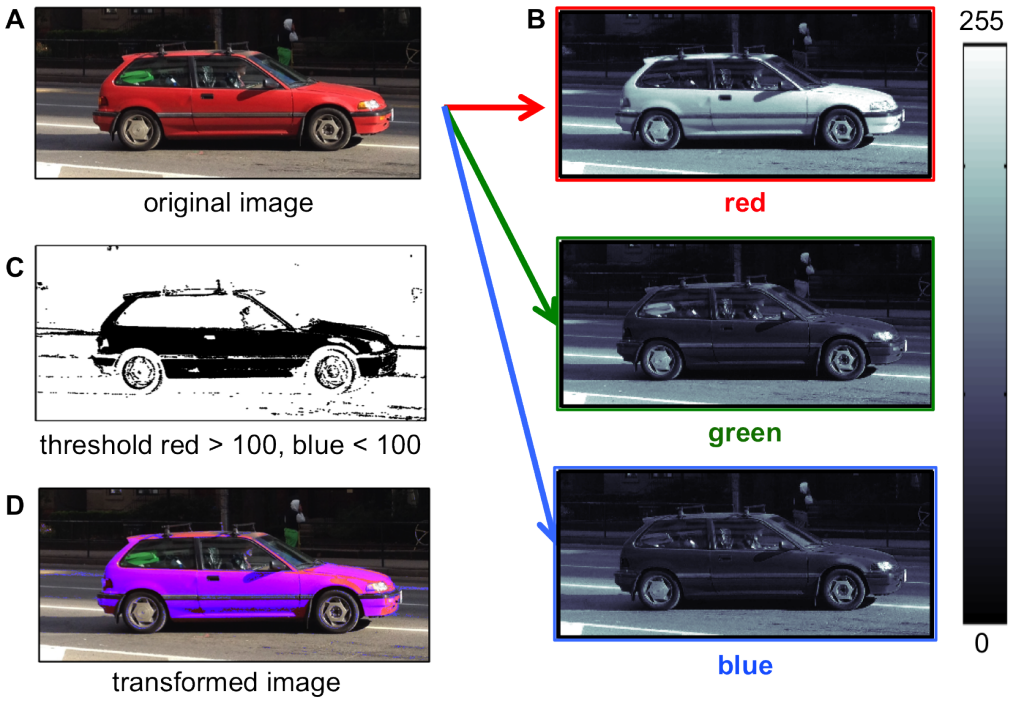

Figure 2: Intensity thresholding. A: A RGB image of a red car with a green box in the back. B: Red, green, and blue image channels. Whiter pixels correlate to more intensity in a channel, darker pixels correlate to less intensity in a channel. Note how the car is bright white in the red chanel, the green box is white in the green channel, and the street is relatively white in all channels. C: A binary mask is formed by thresholding for pixels with intensity greater than 100 in the red channel and less than 100 in the green channel. Pixels selected for by the threshold ("masked") are shown in black. D: The blue-channel intensity in the masked pixels is increased to turn the car purple.

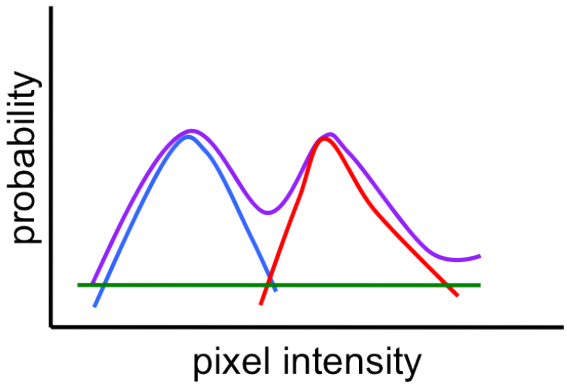

Figure 3: Gaussian Mixture Model: The two gaussian distributions (red and blue), together with the uniform distribution (green) are fit to explain the purple distribution.

images are simply images with only one matrix describing their pixel intensities. Thresholding in an image channel or combination of channels is a simple way of selecting pixels that have a specified intensity. For example, in Figure 2.0.1, pixels associated with the car have red channel intensity greater than 100 and green channel intensity less than 100 (Panel B). By thresholding for these pixels, a mask of the car can be computed (Panel C). However, thresholding does not typically produce a perfect segmentation due to color variations in images and existence of unrelated image objects with the same intensities. For example, here, the light shining on the car above its front tire changes the intensity in this area, and these pixels are not included in the mask. Additionally, many pixels belonging to the road markings do fall in the mask, although undesired. Preprocessing or post processing, for example with other methods described here, is often necessary to produce an accurate segmentation using thresholding. In Aims two and three, we use thresholding as part of the algorithms for nuclei segmentation. In nuclei segmentation, Otsu's method, which automatically selects an optimal threshold, is commonly used for creating a nuclei mask [35, 36, 37].

- Gaussian Mixture Models/: Mixture models separate a distribution into a set of sub distri-

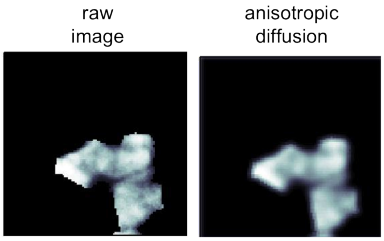

Figure 4: Anisotropic Diffusion of Hoechst image.

butions that together represent the full distribution [38]. For example, if a distribution has two peaks, it might best be described by two Gausssian distributions 2.0.1. In Aim 2, our gray-scale stained tissue images consist mainly of three components: nuclei, cytoplasm, and lumen. As the nuclei pixels mostly have very low intensities, the lumen pixels mostly have very high intensities, and the cytoplasm pixels fall in the middle, and there are thousands of pixels from each class in a given image, we assume that the intensities for each of the three regions are gaussian distributed and fit three gaussian distributions and a uniform background distribution to the intensities distribution of the entire image. This allows us to avoid choosing a specific threshold for nuclei intensities that must hold for every image, as regardless of absolute image intensities, the nuclei will always belong to the gaussians with the lowest intensities. We create a binary mask by thresholding all pixels with a minimal probability of belonging to the gaussian designated as belonging to nuclei.

- Anisotropic Diffusion: [39] Variations in pixel intensity due to experimental noise or signal noise are often smoothed using diffusion. Anisotropic diffusion smooths the image at each pixel according to the local gradient at that pixel, in that way respecting edges in the image. For example, Figure 2.0.1 shows a fluorescent image of a group of tightly clustered nuclei with large amounts of pixel intensity variation due to image noise. By smoothing with anisotropic diffusion, the intensity within each nucleus becomes uniform, while the barriers between nuclei remain intact. After diffusion, thresholding can be applied to identify the individual nuclei. The edge-preserving nature of anisotropic diffusion can lead to some ar-

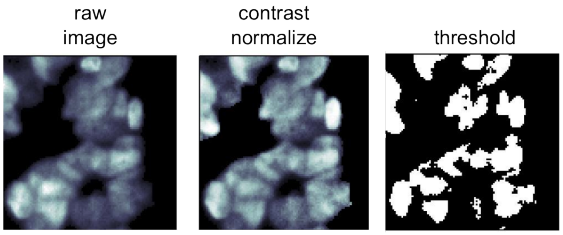

Figure 5: Contrast Normalization of Hoechst image prior to thresholding

tifacts, such as artificial edges forming from image noise, which can be mistaken for nuclei [40]. Thus, our nuclei segmentation algorithm in aim two includes a post-processing step after thresholding anisotropically diffused images to remove putative nuclei that do not fit the standard size or shape of nuclei.

- Total-variance denoising: An alternative method to anisotropic diffusion for reducing image noise while respecting edge is total-variance denoising [41, 42]. This method balances the intensity of each pixel with the intensity of its neighbors through a graph representation of the image, so that inconsistent pixel intensities due to noise are smoothed out. We apply total-variance denoising as a pre-processing step prior nuclei segmentation on stained gray-scale images in aim 2. The regularization parameter λ, which controls the amount of influence a pixel's neighbors have on its denoised intensity, can be varied to improve the segmentation in different ways. If λ is high, the neighboring pixels dominate over a pixel's own intensity, and the denoised nuclei are much smoother and less likely to be over-segmented (one nucleus is mistaken for several nuclei). However, high regularization also causes clustered nuclei to be merged, resulting in under-segmentation (several nuclei are mistaken for one large nucleus) of closely neighboring nuclei.

- Contrast Normalization: Contrast normalization adjusts pixel intensity with respect to the intensity of surrounding pixels [43]. This process highlights pixels that are much lighter or darker than their surroundings, and is very helpful in images with variations intensities. In tissue imaging, nuclei often appear brighter or darker due to the amount of stain they have

retained or their depth in the tissue. In 2.0.1 (left panel), we see an example of this artifact in a fluorescent image of a cluster of nuclei, in which a few nuclei are very bright. Thresholding would not be able to identify the nuclei in the initial image, because in order for the darkest nuclei to be selected by the threshold, cytoplasm pixels around the lightest nuclei would also be selected by the threshold. By first contrast normalizing the image (middle panel), all of the nuclei become are transformed to equally light intensities, as they are all lighter than their surroundings, and all of the cytoplasms are transformed to equally dark intensities, as they are all darker than their surroundings. At this point, thresholding is able to detect the nuclei (right panel).

- Watershed Segmentation: Conceptually, the watershed transformation views the gray-scale image is viewed as a topology map with low intensities corresponding to basins and high intensities corresponding to peaks. If one imagines rain pouring down on the map, and flowing to basin points, for each basin point, all pixels from which water would flow downhill to that point are assigned to the same cluster [44]. As nuclei typically have lower intensities than their surroundings, the watershed transform is a natural method for segmenting nuclei and has been applied to this task for decades [33, 45]. However, the segmentations often result in over segmentation, as multiple basins are commonly found within the same nuclei. Thus, this method is usually combined with other methods as part of a multi-step algorithm [46]. Here, we apply the watershed transformation as a post processing step on the segmented nuclei to adjust image boundaries.

- Canny edge detector: Edge detectors, which looks for lines or curves along which there is an intensity change in an image, is a common step in many image segmentation algorithms. In nuclei segmentation algorithms, edge detection can be used to identify nuclei boundaries in order to improve segmentation, as well as tissue and cell boundaries to delineate tissue architecture within images (Figure 2.0.1) [47, 48]. Canny edge detection is a commonly used multi-step edge detector which tries to reduce false edges created from noise [3]. For our applications, an important parameter of the canny edge detector is the size of the Gaussian smoothing filter. For epithelial cell classification in aim two, we use both large filters to find major edges in the stained H& images corresponding to cell and lumen boundaries and small

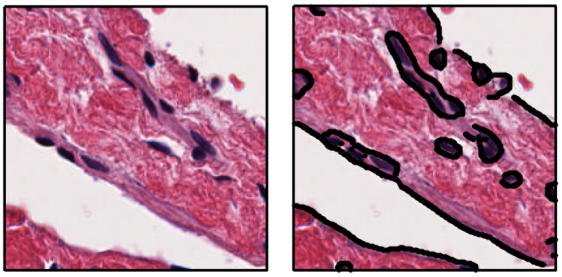

Figure 6: Canny edge detector finds tissue boundaries and nuclei in H$ image of lung tissue

filters to roughly identify the location of neighboring cells.

- Dilation / Erosion: Dilation and erosion can be used to expand (dilation) or shrink (erosion) binary mask. This is often used to correct binary masks when intensity variations have caused pixels to be missing from the mask. For example, in Figure 2.0.1, there is a gap between the two stick figures. By dilating and then eroding, the two figures are combined into one figure. In Aim two, we use dilation and erosion to fill gaps in the cell boundary mask.

Example of recent nuclei segmentation algorithm

Al-Kofahi, et al, focus on whole side images, for which computationally efficient algorithms are also necessary [33]. They perform automatic image binarization using a mixture of two Poisson distributions, which they find to be more appropriate than the traditional mixture of Gaussians. Furthermore, they minimize an energy function (with terms for labeling and continuity) to find an optimal labeling of the image as foreground and background. To identify individual nuclei,

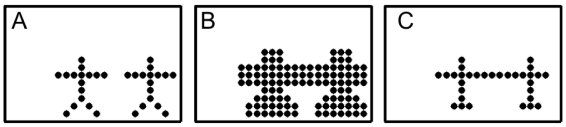

Figure 7: Dilation and erosion of a binary image. A: initial image. B: dilation. C: erosion of dilated image.

Figure 8: Object cropped from larger image.

they use size-constrained clustering, instead of watershed, which they find to be just as fast as watershed, while resulting in less over segmentation. They developed a parallel method based on graph coloring to adjust boundaries of non-neighboring nuclei. Users are able to merge/split nuclei for final corrections using a GUI. They suggest using a segmentation confidence score to screen for nuclei that are likely to be poorly segmented, and only presenting these nuclei to the user for correction, so that the user does not have to screen the entire image. They define "encroachment errors" as errors involving incorrect boundaries. Due to the pixelization of the actual image and the possibility of manual error, they only consider encroachment errors that correspond to at least 25% of the nucleus error. Overall, they have 94% accuracy, if only over- and under-segmentation errors are considered, and 86% accuracy if encroachment and binarization errors are also considered. In their images, under segmentation results from highly clustered with weak borders and over segmentation results from elongated or highly textured chromatin.

Holistic Scene Understanding

When pathologists look at tissue images, or when humans look at any scene, they are guided by global features. For example, when a person first looks at the scene, they can usually instantly recognize whether the scene is inside or outside, contains people, is in a city or nature, etc. After assessing the scene as a whole, they examine individual objects in the scene. Consider . It would take most people a bit of time to identify the object/s in the scene, unless they are very familiar with such objects. However, if shown the entire scene (Figure), it is very easy to identify the objects shown in the snippet in Figure . Upon looking at the entire scene, it is instantly recognizable as a picture of boats in a lake on the mountains. Upon a second glance, a person might observe that they are in Switzerland, because one of the boats has a swiss flag

and Switzerland is known for having snowy mountains. At that point, it becomes clear that the objects are tarps covering the boats for the winter.

Holistic scene understanding algorithms use global, local, and scene information together to guide computational image parsing (assignment of each image pixel to a semantic class). Recently, [49] introduced a method which jointly performs image parsing, object detection, and scene classification to reduce errors from individual tasks. To reduce computational complexity, they represent the image as a hierarchy of segments and super segments, instead of pixels, which are determined using contour detection [50]. Using a type of random field (2.0.1), specifically a holistic conditional random field, likelihood of specific object pairs occurring in the same scene is modeled, as well as likelihood of specific objects occuring in specific scene types. As holistic scene understanding is dependent on results from multiple tasks (object detection, scene classification, pixel grouping), Parikh, et al examined the amount of improvement possible in holistic algorithms through ideal results in individual tasks by replacing outputs from each machine task in [49] with human outputs [51]. One of their findings is that although humans perform slightly worse at isolated superpixel classification than machines, the overall algorithm performs better with human superpixel classification. This indicated that the mis-classifications by humans were less deleterious than machine mis-classifications. They analyzed the human and machine classification errors and found superpixel class features were important for subsequent algorithm steps, which they used to adjust the machine segmentation protocol to produce an overall more accurate algorithm.

Another recent example of holistic scene segmentation is from Lazebnik, et al [52]. They designed a two part classification algorithm with combines bounding-box detectors scanning for specific objects with region-based segmentation using a support vector machine [53]. After obtaining an initial set of labels, they smooth the labels so that neighboring segments agree using a Markov Random Field. To obtain their initial region labels, they define a probability score which computes the log-likelihood ratio between the probability that a pixel belongs to a certain class and the probability that the pixel does not belong to that class. While it takes several days to initially train the algorithm, the average running time for an individual image is approximately 3 minutes, and the MRF inference takes only 6.8 seconds per image. On average their accuracy in

Figure 9: Outdoor scene.

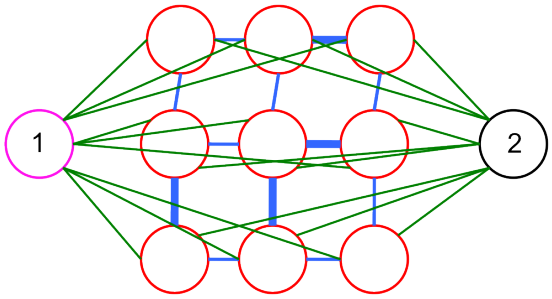

Figure 10: A simple Markov Random Field with nine nodes and two labels (1 and 2). Green lines represent edges between superpixel nodes and label nodes, and blue lines represent edges between superpixel nodes. Thicker lines indicate stronger edge weights.

assigning pixels to the correct class is 83.9%, where some (rarer objects, smaller objects) classes have lower accuracy, and others classes being easier (buildings, sky).

Markov Random Fields (MRF)

Holistic segmentation algorithms often use a Markov Random Field to connect image superpixels with neighboring superpixels and with labels [49] . In this model, each superpixel in a node in a graph (red circles in Figure 2.0.1) and there is an additional node in the graph for each label (black and pink circles in Figure 2.0.1). Edges connect neighboring superpixels as well as superpixels and labels. A cut is sought to find the optimal assignment of labels to superpixels. This is equivalent to solving the optimization problem

$$\min_{x} \sum_{i=1}^{n} \sum_{j \sim i} \psi(x_i, x_j) + \sum_{i=1}^{n} \phi(x_i, y_i),$$

which finds the optimal set of labels $\vec{x}$ for nodes with values $\vec{y}$, such that labels on neighboring pairs are probable (controlled by binary probabilities, ψ) and labels on nodes agree with the node value (controlled by unary probabilities, ϕ) [24, 26, 33, 54]. We describe MRFs in more depth in 3.5.2.

3.0 AIM I: EPITHELIAL CELL CLASSIFICATION IN BARRETT'S ESOPHAGUS FOR ANALYSIS OF PRE-CANCEROUS CHANGES IN NUCLEI

3.1 ABSTRACT

We present a methodology for enabling early cancer detection in Barretts Esophagus (BE). BE is a condition affecting 10% of heartburn sufferers, for which 0.1% of patients develop esophageal adenocarcinoma each year. BE patients undergo endoscopic surveillance for low grade dysplasia (LGD), a pre-malignant lesion. Both diagnosis of LGD and establishment of treatment course suffer from high pathologist disagreement rates, due to shared disease phenotypes between LGD and non-malignant conditions, as well as the propensity of LGD to regress without intervention. As treatment is not completely harmless, extent of dysplasia and degree of cancer risk must be established before treatment can begin. Computational pathology can aid early detection of high-risk LGD through unbiased, data-driven algorithms.

We develop an automated algorithm which identifies epithelial nuclei in biopsy samples on which nano-scale optical biomarkers, related to cancer risk, can be quantified. Specifically, by modeling each tissue image as a Markov Random Field on putative nuclei within the image, we incorporate context-based features describing epithelial nuclei to find an optimal labeling of all image pixels as belonging to epithelial nuclei, other nuclei, or background. Our method identifies 97% of nuclei within our data set, and correctly labels over 90% of those nuclei as epithelial or non-epithelial. We show that a nano-scale biomarker measured on epithelial nuclei, computed through spatial-domain low-coherence quantitative phase microscopy, varies significantly between patients with BE and no dysplasia, BE and high grade dysplasia, and BE with esophageal adenocarcinoma, establishing its utility as a clinical measure for dysplasia. The au-

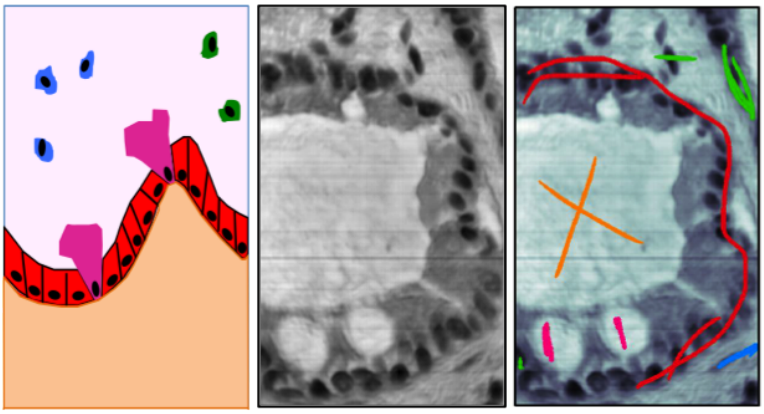

Figure 11: Benign BE tissue. Left-most panel shows cartoon image of BE. Black circles indicate nuclei. Following image pair shows labeled biopsy slice, labeled by DH. Color legend for left- and right-most images: red= epithelial cell, orange= lumen, blue= stromal cells, green= lymphocytes, pink= goblet cells. [1].

tomated epithelial nuclei detector produces a higher quality selection of epithelial nuclei than manual detection, resulting in enhanced characterization of pre-cancerous phenotypes. Algorithms such as these applied in a clinical setting provide unbiased reasoning to assist diagnosis of ambiguous cases, save time by filtering out obvious cases, and can help establish degree of cancer risk for individual patients.

3.2 INTRODUCTION

Barrett's esophagus (BE) is a pre-malignant condition occurring in 10% of gastro-esophageal reflux patients [55, 56, 57], in which esophageal epithelium undergoes benign metaplasia [55]. Specifically, columnar epithelium containing glandular cells replaces the normal squamous epithelial lining of the lower esophagus. The prevalence of BE is estimated to be around $1-2\%$ in Europe and predicted to be up to 6% in the USA, [58], with 23 million cases in 2001 [59]. These

BE patients are $30 - 125$ times more likely to develop esophageal adenocarcinoma [EAC], cancer of glandular epithelial cells, then the average population [57]. As EAC has one of the fastest growing incidence rates of all cancers and a five-year survival rate of less than 5%, decreasing mortality requires early identification of the BE patients who are at risk for developing EAC [59].

Traditionally, BE is diagnosed by both endoscopy and biopsy, where the biopsy must show intestinal metaplasia, evidenced by glandular goblet cells 11 [55, 56, 60]. After an initial biopsy confirms BE, patients undergo routine biopsy surveillance, with the frequency of biopsies increasing if the patient develops dysplasia, abnormal changes in cell nuclei [57, 60]. Barrett's esophagus is understood to progress along a metaplastic - dysplastic - carcinomic pathway from non-dysplastic metaplaspia (ND), through low-degree dysplasia (LGD) and high degree dysplasia (HGD), to esophageal adenocarcinoma (EAC) [57, 61], although some patients never progress past an early stage. Early recognition of epithelial tissue likely to progress to HGD would enable targeted anti-cancer treatment to begin before onset of dysplasia.

Carcinomas, which include around 80% of human cancers, originate in epithelial nuclei [62], thus pathologists have traditionally examined epithelial cells when diagnosing cancer (red cells in fig. 11). Optical technologies, such as spatial-domain low-coherence quantitative phase microscopy (SL-QPM, 3.3), that seek to identify early characteristics of cancer at the nano-scale, that is before cancer is evident through tissue architecture, also focus on epithelial cell nuclei, as these will show precancerous changes earlier than other stromal, and other, cell nuclei (blue and green cells in Fig. 11).

Pathologists identify epithelial cells in BE tissue using a mix of holistic insight and local information. Specifically, they look for chains of columnar cells that surround a lumen area, with apical sides facing the lumen (Fig. 11). While in cartoon examples the epithelial cells are usually easily identified (Fig. 11,left), in reality discerning epithelial cells from other cells can be difficult, leading to non-trivial rates of inter- and intra-pathologist disagreement [19]. and distinguishing goblet cells within chains of epithelial cells can often be challenging. Consider, for example, the biopsy tissue sample shown in Fig. 11 (middle, right). Here, a pathologist (DH) has labeled nuclei and regions as either lumen, goblet cells, epithelial nuclei, stromal nuclei, or lymphocytes. The two goblet cells most likely have nuclei within this image, but the pathologist

has given no label to the four-five nuclei residing below the goblet cells, as it is not clear which of these belong to the epithelium and which are goblet cells. The region in the upper-right corner marked as lymphocytes could easily be mistaken for a chain of epithelial cells. Here, the lack of neighboring lumen region and shape of nuclei must have lead to the classification of these nuclei as lymphocytes. It is not possible to classify an individual nucleus as epithelial or non-epithelial without knowledge of the nucleus's local and global surroundings. However, only local information is needed to classify an object in the image as a nucleus. **We present an algorithm that mimics the strategy used by pathologists to identify epithelial nuclei**: we first identify all regions in the image that could be nuclei, and then use holistic image segmentation, encoded in a Markov random field, to classify nuclei as epithelial or non-epithelial. Our automated epithelial classification system significantly reduces the manual labor required by researchers to label epithelial nuclei within cell images, while eliminating bias in their selection.

Contributions

We first present a versatile automatic nuclei segmentation algorithm together with a GUI that can be used to manually improve nuclei boundaries and select nuclei for further analysis. We then present an automated epithelial classification algorithm that incorporates contextual clues learned from pathologists to label nuclei as epithelial or non-epithelial. Finally, we show that an optical biomarker computed with SL-QPM can be measured on epithelial nuclei to stratify healthy tissue from Barrett's esophagus patients according to their likelihood of neighboring cancerous tissue and that automated segmentation provides a higher quality quantification of this biomarker than manual segmentation.

3.3 BACKGROUND

Early Detection of Disease/Cancer

Due to the high likelihood of patients with BE developing EAC, BE patients undergo routine esophagus biopsies. However, there is a high degree of inter- and intra- pathologist disagreement at the critical LGD/HGH stage, where a small number of patient's eventually develop cancer,

but most do not [63]. Additionally, biopsies are limited in size and number, and diagnostically critical dysplastic or carcinomic regions may be overlooked. The ability to detect early signs of cancer outside of dysplastic regions, and prior to onset of tissue architectural changes, would allow better identification of those patients who should begin anti-cancer treatment.

Spatial-domain Low-coherence Quantitative Phase Microscopy (SL-QPM)

"Normal" tissue, predisposed to carcinogenesis, displays molecular changes on the nanoscale level indicative of carcinogensis [2]. Conventional microscopy visualizes tissues at the micron scale, at which these chromosomal level changes are not apparent. Spatial-domain low-coherence quantitative phase microscopy is a novel optical method which detects structural changes at the sub-nanometer level [20]. The presence of molecular alterations in tissue predisposed for cancer has been evidenced in several cancers, including breast and esophagus[64].

A technical explanation of SL-QPM can be found in [65], and an overview is given here: As cancer originates in DNA, it is to be expected that pre-carcinomic alterations would be apparent in DNA packing and arrangement, before the alterations cause changes in cell and nucleus structure. Traditional microscopy methods can not visualize changes within the cell nuclei. However, DNA packing and organization influences local density within nuclei. Light passes through media with different densities at different speeds, and thus light will, on average, pass through pre-cancerous nuclei at different rates than through healthy nuclei. By measuring how light of a large array of wavelengths passes through each pixel of each the nuclei in an image, the phase of light passing through each pixel can be computed. The phase is computed at several depths of interest to identify a measurement that best resolves diagnostic differences between tissue. The method has been shown to be robust against small variations in experimental factors, such as staining and tissue thickness (on average 4μm thick).

Image Analysis for Nuclei Detection

Nuclei segmentation is a key step in computational pathology algorithms, as many biomarkers are measured on cell nuclei [10, 34]. Nuclei patterns differ naturally between cell types, organs, and diseases, and are captured differently depending on the slide preparation, staining, and imaging device, thus many system-specifc nuclei segmentation algorithms exist [13, 21, 22]. In all systems, segmentation is challenged by cell density, nuclei density, overlapping nuclei, im-

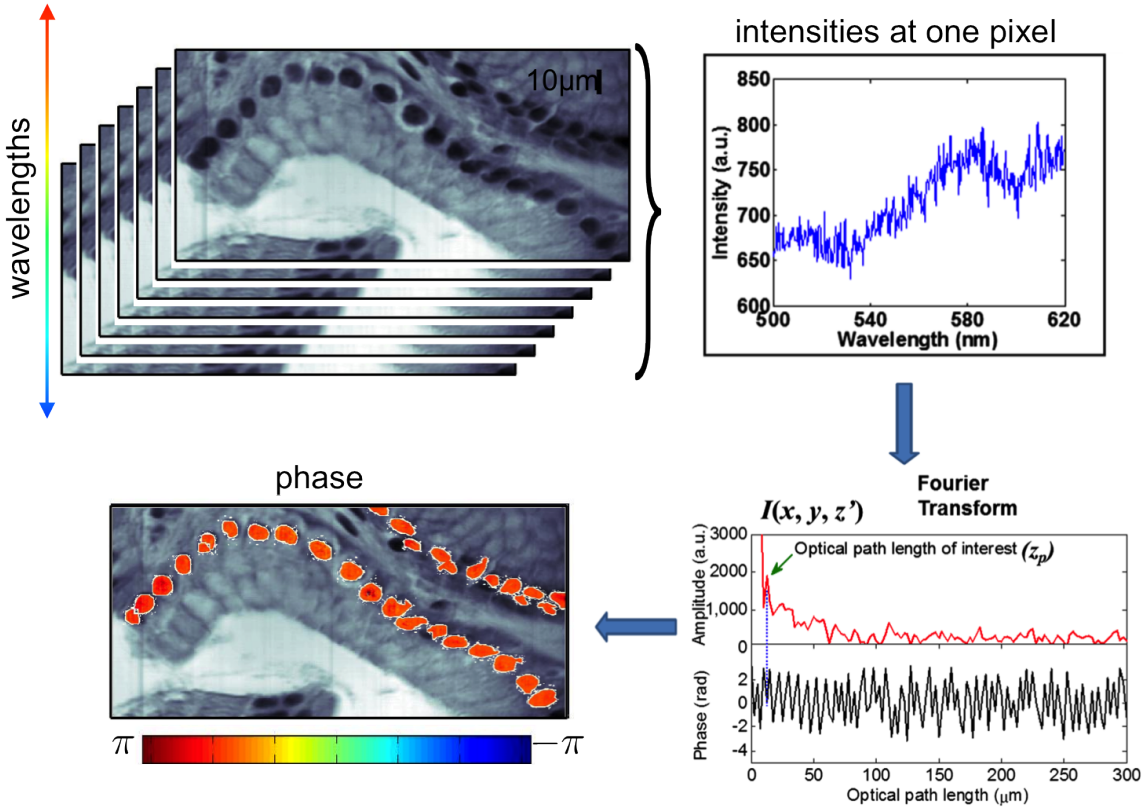

Figure 12: Adapted from [2]: Overview of SL-QPM system. Left-top: Tissue is imaged at 1004 wavelengths. Right-top, bottom: For each pixel, the graph of all intensities is Fourier-transformed. Left-bottom: Optical path length for all epithelial nuclei pixels, derived from Fourier Transform, is converted to phase, and mapped on to initial tissue image.

age contrast, background noise, and variations in nuclei morphology [33]. Importantly, tightly clustered nuclei and nuclei with unique morphologies are more difficult to segment than most nuclei, but also more likely to be indicative of disease [34]. The most basic methods are based on intensity thresholding, as nuclei are usually darker than their immediate surroundings [34]. However, variations in image intensity and both biological and experimental noise create many false positive nuclei. Diffusion and contrast normalization are often used to improve over intensity thresholding, followed by system specific methods to weed out false positive nuclei and break up clusters of nuclei. The tissue studied here is imaged at a $40\times$ magnification, higher than the majority of scanned images. As such, more detail within nuclei is likely to lead to over-segmentation of nuclei using standard algorithms that are used to lower resolution data. Additionally, the higher resolution allows algorithms for segmenting these nuclei to aim for better performance on highly clustered nuclei, as they are better resolved. Thus, designing a nuclei segmentation algorithm specific to this optical system should provide a more accurate segmentation than a packaged algorithm repurposed for this data would.

Machine Learning for Epithelial Classification

Recent papers have shown the utility of measuring biomarkers individually on epithelial and stromal nuclei for both cancer detection and uncovering of novel cancer biomarkers. Linder, et al, trained a support vector machine on texture features describing small blocks of tissue ($42\mu m \times 42\mu m$, containing around 5 nuclei per region) in $10\times$ magnification colorectal tissue slides [66], with 97% agreement between human and classifier. However, the resolution of their classifier is not at the nucleus level, but at the block level, so individual stromal cells interspersed within a block containing a chain of epithelial cells would be labeled as epithelium. Beck, 2011, labels superpixels in $20\times$ magnification images as epithelial or stromal using L1-regularized logistic regression learned on a training set of images, with 89% accuracy [13]. While the superpixels provide a more nuclear-specific labeling than the blocks used in [66], the classification can still mislabel individual nuclei within a superpixel.

Holistic Segmentation by Encoding Context

When analyzing histological images, pathologists rely heavily on contextual information to understand the cellular environment. For example, the arrangement of nuclei in glands or chains

helps identify particular nuclei as being epithelial or stromal (See Fig. 11). The utility of including context information has been demonstrated in computational image parsing, where overall segmentation and object classification has been improved by incorporating contextual features into image understanding [67]. For nuclei classification, we aim to improve our classification model by both the inclusion of learned contextual clues from ground-truth images (e.g. average distance between epithelial nuclei), as well as the inclusion of knowledge-based contextual clues from discussion with expert pathologists (e.g. orientation of epithelial chain to lumen).

3.4 DATA

As part of the SL-QPM protocol, each tissue sample is imaged at 1004 wavelengths, yielding 1004 separate images for a single sample. We use the average of these 1004 images for nuclei segmentation and classification, and refer to the average image simply as the image.

Our learning data consisted of 414 stained histology images at $40\times$ magnification ($0.25\mu m$ per pixel) from healthy (BE-normal) tissue taken from 89 patients, with each patient yielding four to five images. The average image size in the data base is (531×363) pixels, or $133\mu m \times 91\mu m$, with image size ranging between 32770 and 359840 pixels. From this data set, 47 patients (215 images) were diagnosed with Barrett's Esophagus, no dysplasia [BE-normal], 28 patients (131 images) were diagnosed with Barrett's Esophagus and High Grade Dysplasia [BE-HGD], and 14 patients (68 images) were diagnosed with Barrett's Esophagus and Esophageal Adenocarcinoma [BE-EAC].

In addition to the 414 images from our learning set used to train and test the nuclei segmentation and epithelial classification algorithms, we obtained a set of 424 stained histology images of the same magnification and in the same size range, for which phase information was calculated using SL-QPM to evaluate the usage of SL-QPM for early cancer detection in Barrett's Esophagus [64]. Images in this "experimental set" came from the same set of patients as the training set, with diagnoses of BE-normal, BE-HGD, and BE-EAC, and again contain only healthy (BE-normal) tissue.

Importantly, while the images come from patients of three diagnostic classes, the tissue selected for imaging is in all cases healthy BE, with no dysplasia. It should **not** be apparent, even to an expert, that any of the tissue samples actually come from patients with an increased risk of cancer over BE-normal patients, as here we are studying cell changes in the field adjacent to carcinoma.

Ground Truth

Nuclei Segmentation: Ground truth labeling of nuclei boundaries was performed by VB using a matlab GUI designed for the task to label nuclei boundaries, and verified/edited by pathologist DH on a random sample of 10 BE-normal images, 10 BE-HGD images, and 10 BE-EAC images. *Epithelial Classification:* On a subset of 38 images from the same set of 89 patients, but unique from the learning and experimental sets, image regions were marked by DH as belonging to epithelial cells, stromal cells, inflammatory cells, goblet cells, lymphocytes, other non-epithelial cells, or lumen. The 414 image set was then labeled accordingly by VB and verified/edited by DH. For the 424 image set, nuclei boundaries were automatically predicted using Phase I of our algorithm, and then putative nuclei were labeled as epithelial or non-epithelial by KS.

Evaluation

We evaluate our nuclei segmentation and epithelial classification methods according to true positive rate (TPR), false positive rate (FPR), and accuracy. For epithelial classification, the TPR is defined as the percent of nuclei with ground-truth label epithelial, that are also predicted to be epithelial. The FPR is the percent of nuclei with ground-truth label non-epithelial, that are predicted to be epithelial. The accuracy is defined as the total number of correctly classified putative nuclei, divided by the total number of putative nuclei. For nuclei segmentation, we define a true positive as any predicted nucleus that overlaps with a ground-truth nucleus, a false positive as any predicted nucleus that does not overlap with any ground-truth nuclei, and a false negative as any ground-truth nucleus that does not overlap with any predicted nuclei. The total number of true nuclei is the number of ground-truth nuclei, and the total number of false nuclei is the number of false positives. As our definition of true positive is very weak, in that we only require one pixel overlap for a putative nucleus to be considered correct, we used two additional measures to establish the quality of the predictions while tuning our nuclei segmentation algo-

rithm. The %- covered measures the number of pixels shared by the putative nucleus and its corresponding ground-truth nucleus, divided by the total number of pixels in the ground-truth nucleus. The %-wasted measures the number of pixels from the putative nucleus that are not also in its corresponding ground-truth nucleus, divided by the total number of pixels in the putative nucleus.

3.5 METHODS

Epithelial segmentation proceeds in two phases. In Phase I, putative nuclei are identified in the image. We outline the nuclei method used for this data set (3.4) below, which we designed to obtain accurate nuclei with respect to ground-truth nuclei boundaries. This method does not seek to minimize the number of false positives (tissue regions mistaken for nuclei), but instead tries to maximize the number of true positives, as the epithelial classification algorithm in Phase II is able to identify most false positives, but suffers when epithelial nuclei are missing from epithelial chains, making global information incorrect. In Phase II, nuclei are labeled as belonging to epithelial or non-epithelial cells using a conditional Markov random field (MRF).

3.5.1 Phase I: Nuclei segmentation

We have developed a nuclei segmentation method that identifies putative nuclei in stained tissue images. While different image sets/techniques (staining, magnification, cell-type, etc) will require different parameters or perhaps additional steps, we have found that this method accurately identifies nuclei in several tissue image data sets. Consider segmenting the nuclei shown in Fig. 11 (middle). While many nuclei can be easily identified as black circles, the nuclei on the top left are tightly clustered and hard to resolve, and some nuclei near the bottom right have weaker intensities than the majority of the nuclei. Additionally, there are several dark regions in the image that could be mistaken for nuclei, while they are actually simply variations in cytoplasm/lumen intensity. Some nuclei can also have intensity variations, causing over segmentation of the nuclei into several smaller nuclei. These intensity variations can have biological explanations, such as

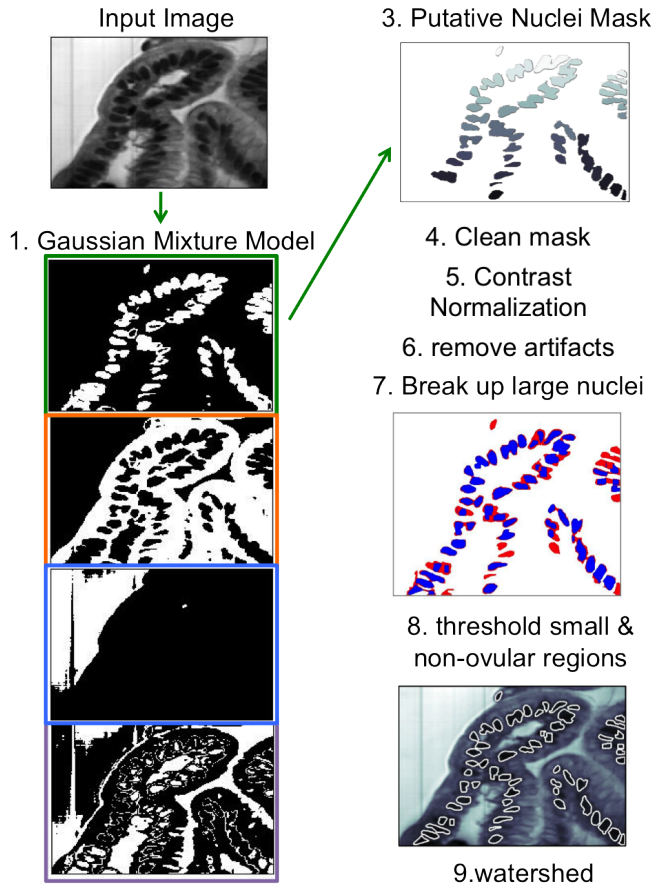

Figure 13: Nuclei segmentation overview. In step 1, a mixture of three Gaussians and a background distribution are fit to the image. Each of the four boxed images corresponds to one of the distributions (maroon corresponds to the background distribution), where white pixels indicate pixels that are most accounted for by that distribution. The green-boxed distribution is automatically identified to correspond to the nuclei, based on the size and shape of its connected components. A mask of putative nuclei is formed from pixels accounted for by this distribution in step 3. Here, each color indicates a putative nucleus. Note that some of the putative nuclei actually correspond to several closely neighboring nuclei, and must be further processed in steps 4-7. After step 7, the initial putative nucleus mask is shown in red, and the processed large nuclei are replaced by blue nuclei. Steps 8 filters out particles that are lacking typical characteristics of nuclei, and step 9 smooths nuclei boundaries using watershed.

chromosome location within nuclei, but can sometimes also be due to equipment/experimental error.

First, to reduce intensity variations in cytoplasm regions that can be mistaken for nuclei and variations in nuclei regions leading to oversegmentation, the image I is denoised using total-variance denoising with a range of smoothing factors (λ) to form the denoised image I^{λ} (Fig. 3.5.1) [42, 68]. Total variation denoising minimizes the total variation with respect to the true signal x_i and the observed signal y_i at pixel i, $|x_i - y_i|$, such that the true values of neighboring pixels are close, where the distance between true values of neighboring pixels i and j is given by $(x_i - x_j)^2$. The smoothing factor λ controls how much weight is given to the total variation term, that is how much more or less important is the variation of the true signal from the observed signal than the closeness of the true values of neighboring pixels. The denoised solution is found by optimizing $\min_x \sum_{i\tilde{j}} (x_i - x_j)^2 + \lambda \sum_i |x_i - y_i|$ [42]. Using Chin's implementation, built on the fast Laplacian solver [69], this step is completed in nearly linear time [68].

Nuclei segmentation

Nuclei segmentations are performed on a set of denoised image transformations I^{λ} of the initial image I and then merged (Fig. 3.5.1). Here, we show that merging multiple segmentations on multiple denoised images provides better nuclei coverage than simply segmenting any one image. While using a low smoothing threshold can produce putative nuclei that are over-segmented and miss nuclei that have strong variations in pixel intensities, low thresholds have the advantage of being able to distinguish closely packed nuclei. In contrast, using a high smoothing threshold can cause incorrectly grouping of tightly packed nuclei into a single putative nucleus, but high thresholds are less likely to over-segment nuclei and are able to sidentify nuclei with significant pixel intensity variation. By segmenting at multiple thresholds and then merging the results, more nuclei are identified and the nuclei boundaries agree better with the ground truth. A disadvantage of this method is that more false positive putative nuclei are found, that is, more regions that are not part of nuclei are falsely labeled as nuclei, but these false positives should be largely removed by the epithelial classification schema. Additionally, running multiple rounds of TV-denoising is time-consuming, thus employing a local scaling factor would improve efficiency in future efforts.

In Table 3.5.1, we show the average false positive and true positive rate for nuclei segmenta-

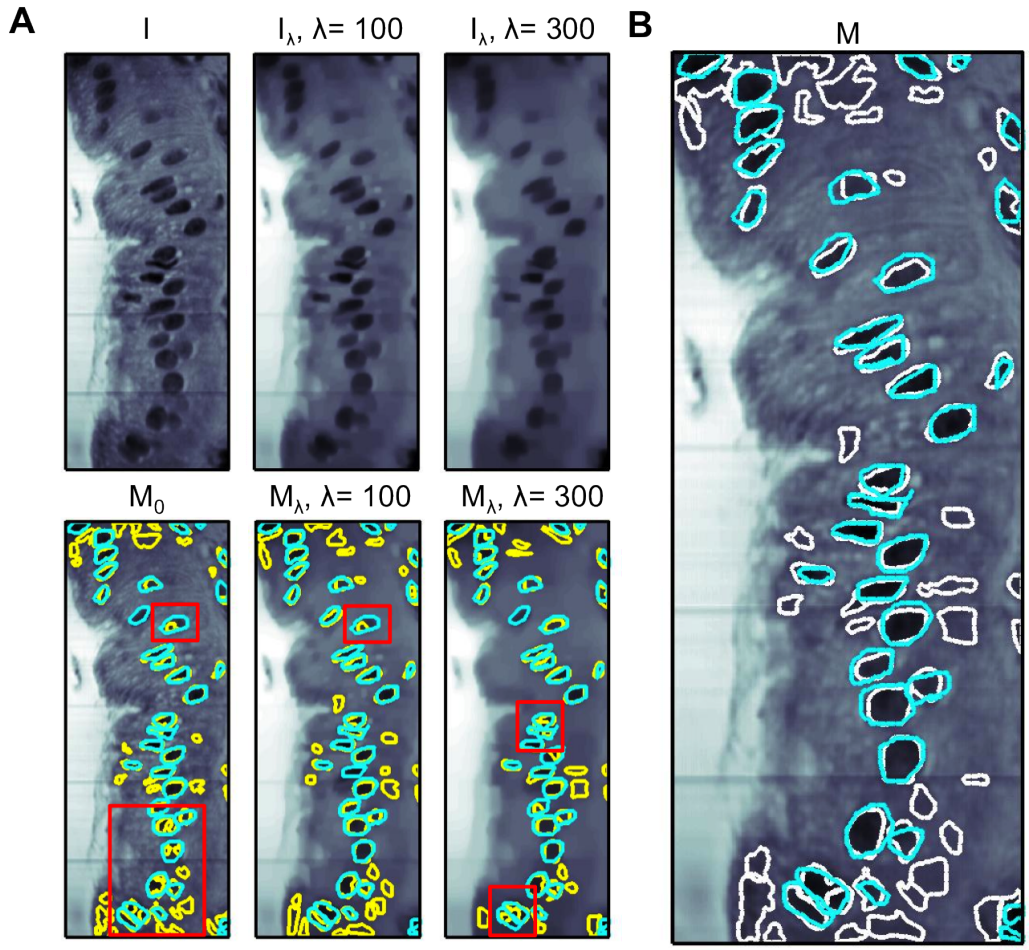

Figure 14: Image denoising and nuclei segmentation. A: Top row, left: raw image (I_0), middle,right: image denoised with $\lambda = 300$ (I_{300}), $\lambda = 100$ (I_{100}). Bottom row: Nuclei segmentations (yellow) according to Phase I on I_0, I_{100}, and I_{300} to form masks M_0, M_{100}, and M_{300}. Cyan: ground-truth segmentation. Red boxes indicate nuclei that were incorrectly segmented with each λ. B: White: Merged nuclei segmentation M. Cyan: ground-truth segmentation.

tion across the 30 hand-segmented ground-truth images. For FPR and TPR, we consider putative nucleus a true positive if any pixel in the putative nucleus belongs to a ground-truth nucleus. Similarly, we consider a putative nucleus a false positive if no pixel in the putative nucleus belongs to a ground-truth nucleus. Additionally, we show the precision/sensitivity of these regions in terms of the percent of pixels in a ground-truth nucleus that are covered by its corresponding putative nucleus % covered), and the percent of pixels in a putative nucleus that are also covered by its corresponding ground-truth nucleus (% not wasted).

λ	0	5	25	50	100	150	200	300	merged
FPR	34	32	37	36	35	33	32	27	43
TPR	93	92	90	92	90	91	89	74	98
% covered	66	65	63	65	65	65	64	51	80
% not wasted	70	68	65	65	64	63	61	48	64
$\frac{\%\text{not wasted}+\%\text{covered}}{2}$	68	67	64	65	64	64	63	50	72

Merging the nuclei segmentations provides identification of 98% of nuclei, 5% better than the identification achieved by any single segmentation. Additionally, 80% of nuclei pixels are covered using the merged segmentation, 14% better than the coverage achieved by any single segmentation. The merged segmentation 'wastes" approximately the same number of pixels as any of the single image segmentations. While the false positive nuclei identification rate is significantly higher using the merged method, most of the false positives do not strongly resemble epithelial nuclei and will be removed by the epithelial classification algorithm. Thus, the advantages of identifying more nuclei with closer agreement to ground-truth than any of the single segmentation methods makes the merged segmentation the best nuclei mask to feed into the epithelial classification schema.

Second, an intensity range corresponding to nuclei is identified for the image. As intensities can vary between tissue images due to staining methods and biological factors, we do not specify a specific intensity range for nuclei for a given system. Instead, for each image, we fit three Gaussian components to the distribution of intensities within the image. These correspond to nuclei, cytoplasm, and stroma/lumen. This removes the need for normalizing all images in the data set to the same background intensity, thus avoiding normalization artifacts. Additionally,

this allows this algorithm to be ported between many systems and tissue types without having to reparametrize intensity thresholds. On each smoothed image I^λ, the following steps are performed:

1. Fit a gaussian mixture model to the image's intensity distribution with three gaussians (typically corresponding to nuclei, cytoplasm, lumen/stroma) and a background distribution.
2. Using intensities and region sizes of the pixels described by each Gaussian component, identify the gaussian G_g component that most likely corresponds to nuclei.
3. Define the nuclei mask M^λ as a $(n_x \times n_y)$ binary matrix, where $M^\lambda(x, y) = 1$ if pixel (x, y) is accounted for by at least $r\%$ by G_g, and 0 otherwise. The cutoff r is empirically set to be $0.45\cdot$maximal percent that a pixel is accounted for by G_g.

Third, at this point, M^λ is equal one for any pixel that may be part of a nucleus. Each connected component in M^λ is considered a putative nucleus. However, M^λ may contain many large connected regions that are actually made up of several closely neighboring nuclei, and it may be missing pixels belong to nuclei that were not captured by G_g, e.g. lighter intensity pixels inside nuclei due to intensity variations. The next few steps work to break up large regions into individual nuclei and smooth out nuclei boundaries.

1. Clean up mask M^λ by removing holes and isolated/bridge pixels.
2. Contrast normalize mask. This is helpful in finding individual nuclei in large regions.
3. Remove thin lines of pixels included in nuclei mask, which are often caused by "wrinkles" in cytoplasm.
4. Further process large regions to break into individual nuclei:
    a. First find average size of putative nuclei at this point by determining the median nucleus radius r_{med} and setting $A_{\text{med}} = \pi r_{\text{med}}^2$. Set an upper bound for large regions as any putative nucleus with area greater than $1.75A_{\text{med}}$. The factor 1.75 was determined empirically. This bound will cause many nuclei of reasonable size to be included in the group of large regions, but if they are sufficiently uniform in intensity, they will be returned unchanged to the set of putative nuclei after the following steps. Additionally, compute some statistics on shape (such as eccentricity and convexity) to determine reasonable bounds on nucleus shape.

b. Remove any large region with very high intensity (light in color), by requiring that the darkest pixel in large regions must be at least as dark as the median intensity pixel in small regions.

c. On each large region, iteratively perform anisotropic diffusion followed by contrast normalization and thresholding, until the region has been broken into multiple regions. The new regions will be added to the set of large regions, if they are also larger than $1.75A_{\text{med}}$, or added to the set of putative nuclei. If a large region does not break into multiple regions, but is of reasonable shape and size, it is also added to the set of putative nuclei, and otherwise discarded.

Finally, at this point, the large regions will all have been broken into smaller regions or deemed to be of reasonable shape and size. We update the parameters for size (A_{med}) and shape using the revised set of putative nuclei.

5 Remove very small regions, defined by any putative nucleus with size less than $\frac{A_{\text{med}}}{3}$.

6 Expand each putative nucleus using watershed to smooth out nuclei boundaries.

This method yields a putative nucleus mask, M^{λ} for each smoothed image I^{λ} (Fig. 3.5.1A, bottom row). We combine these masks so that each pixel is assigned to the largest putative nucleus across all λ at that pixel, to yield a final putative nucleus mask M (Fig. 3.5.1, panel B). The putative nuclei at this point may contain regions that are not actually nuclei, but the second phase of the algorithm should be able to identify these regions as non-epithelial components. Thus, we strive here to have a high True Positive rate, with less concern about achieving a low False Positive Rate.

Results: Nuclei Segmentation

Nuclei segmentation methodology and parameters were optimized on an independent data set of 38 images, taken from a subset of the same 89 patients, but not included in the 414 image set. To establish the accuracy of the method, nuclei were hand-segmented on a validation set of 30 images from the 414 image data set, ten from each of the three diagnostic classes. The hand-segmentation was performed initially by Virginia Burger, and corrected/verified by pathologist Dr. Doug Hartman.

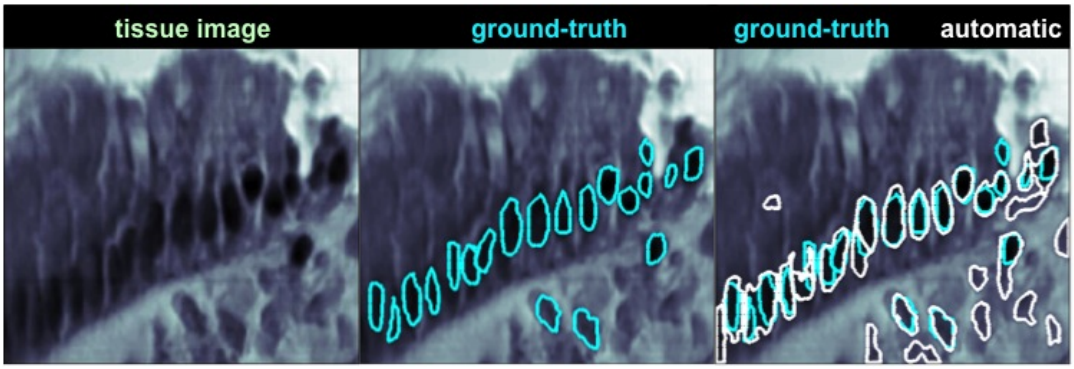

Figure 15: Manual and automated segmentation are shown for a sample image. This example has a 94% TPR and 33% FPR.

	number images	TPR ($\%$	FPR ($\%$)
overall	30	94	33
BE-normal	10	96	34
BE-HGD	10	94	29
BE-EAC	10	94	36

We show the overall segmentation accuracy on the 30 images in Table 3.5.1, as well as the performance on each diagnostic class. There is little variation in segmentation accuracy between classes, which is to be expected since the images all show healthy tissue. A representative ground-truth hand-segmentation and computational nucleus segmentation is shown in Figure 15.

Nuclei Segmentation GUI for manual epithelial classification

To compare automated epithelial classification with manual epithelial selection, we built a MatLab GUI which allows a user to hand-pick putative nuclei as epithelial nuclei for phase analysis. As visualization of the putative nuclei boundaries are distracting and can bias the user, the GUI displays only the raw tissue image. The user clicks on a point in the image within a nucleus to select that nucleus. If the nucleus is part of the putative nucleus set, the GUI displays the boundaries of the putative nucleus at that point. If the nucleus is not part of the set, the GUI uses

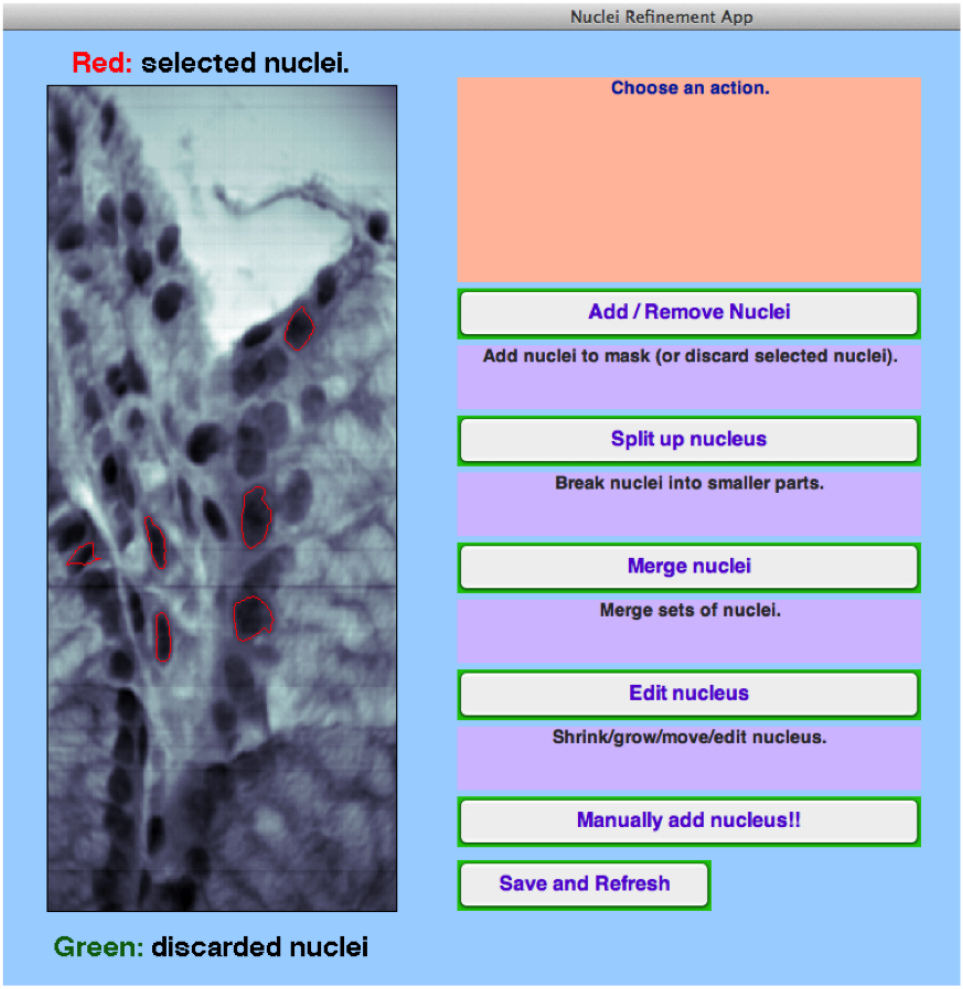

Figure 16: Screenshot of epithelial classification app.

watershed to compute a nucleus at that point and displays its boundaries to the user. The GUI then allows the user to grow or shrink the nucleus, merge two nuclei, or split a predicted nucleus into two nuclei. These actions are all performed using steps from the nuclei segmentation algorithm described above. Additionally, the user can hand-trace a nucleus boundary if unsatisfied with the predicted nucleus at that point.

3.5.2 Phase II: Epithelial Classification

Pathologists use context information, in addition to nuclei descriptors, to identify epithelial nuclei in images. For example, while epithelial nuclei in a particular type of tissue are known to have a certain radius, e.g. $\approx 10\mu m$ in esophagus epithelium, many other nuclei can have this same size. The location of a nucleus with respect to other nuclei and tissue structures complements this information, allowing pathologists to determine specifically which nuclei make up the epithelium. To analogously combine intrinsic and context information while identifying nuclei, we employ a Markov Random Field (MRF) [citation] encoding unary and binary classifiers.

Unary Classifier

Unary classifiers give the probability that a nucleus is epithelial, independent of the labels of its neighboring nuclei. We measured a total of 94 features 2 on each putative nucleus, and built a classifier using AdaBoost to label each putative nucleus with a probability of being epithelial [70]. The feature sets includes descriptors measured on isolated nuclei, such as size, intensity, and convexity, as well as features dependent on the environment, such as distance to cell boundary or next closest nucleus. Used independently, each classifier was only weakly predictive 2. We used AdaBoost with MatLab's default parameters (binary classifier, learning rate of 1, 100 learners) to combine the set of 94 weak classifiers into a stronger classifier, $\psi : \vec{x} \in \mathbb{R}^{94} \rightarrow [0, 1] \subset \mathbb{R}$, where $\vec{x}$ is the feature vector for nucleus x.

Pairwise Classifier

Pairwise classifiers give the probability that a nucleus is epithelial, conditioned on the label (epithelial or non-epithelial) of each of its neighbors. Pathologists use many contextual clues to classify nuclei, e.g. epithelial nuclei tend to form a chain along a lumen region, neighboring epithelial nuclei have similar orientations to the lumen, and size/shape of neighboring epithelial

nuclei are similar.

Epithelial Classification

Initially, a set of n_p features encoding such contextual clues were measured on all pairs of nearby nuclei, where the threshold for "nearby" was set to be a function of the median distance between nuclei within an image.However, due to the randomness of individual nuclei, these pair-wise features alone could not distinguish pairs of same-class nuclei from pairs of mixed-class nuclei (epithelial & epithelial, non-epithelial & non-epithelial, or epithelial & non-epithelial). Thus, to encode more global image information, the tissue architecture within the image was captured in terms of a) location of epithelial cell boundaries and b) arrangement of nuclei in a "tree", with the longest chain of nuclei making up the trunk (Fig. 3.5.2). These tissue architecture features encode the contextual clues used by pathologists: chains of nuclei (described by the "tree trunk") along the lumen border (described by the epithelial cell boundaries).

The nuclei pairs were then divided into eight architecture-categories according to their location with respect to the epithelial cell boundaries and their position on the nucleus tree. To find the tree, we use a greedy algorithm with initiates a trunk at the nucleus with highest unary probability of being epithelial, and adds nuclei to the trunk in either direction, ensuring that added nuclei are close together, form a relatively straight line, and have similar unary probabilities, orientation, and size, where parameters for close, straight, and similar were determined empirically. Once no more nuclei can be added while remaining within the restraints specified by the parameters, all remaining nuclei are added iteratively onto branches, where each nucleus is simply attached with a branch to its closest neighbor already on the tree. This trunk/branch model tends to place epithelial cells on the initial trunk, and any other chains of epithelial form branches of the tree. Thus, most nuclei pairs within the same architectural-category are of the same type: nuclei pairs on trunks tend to be epithelial, nuclei pairs at junctures between branches or the trunk and a branch tend to contain mixed nuclei, and nuclei pairs on branches are often either both non-epithelial or both epithelial. This architectural layout largely removes the randomness of individual nuclei pairs that handicapped the classification of pairwise features, when applied to arbitrary nuclei pairs. To determine cell boundaries, we used a Canny Edge Detector [3], with a Gaussian smoothing filter selected to have width 20μm, representing twice the length

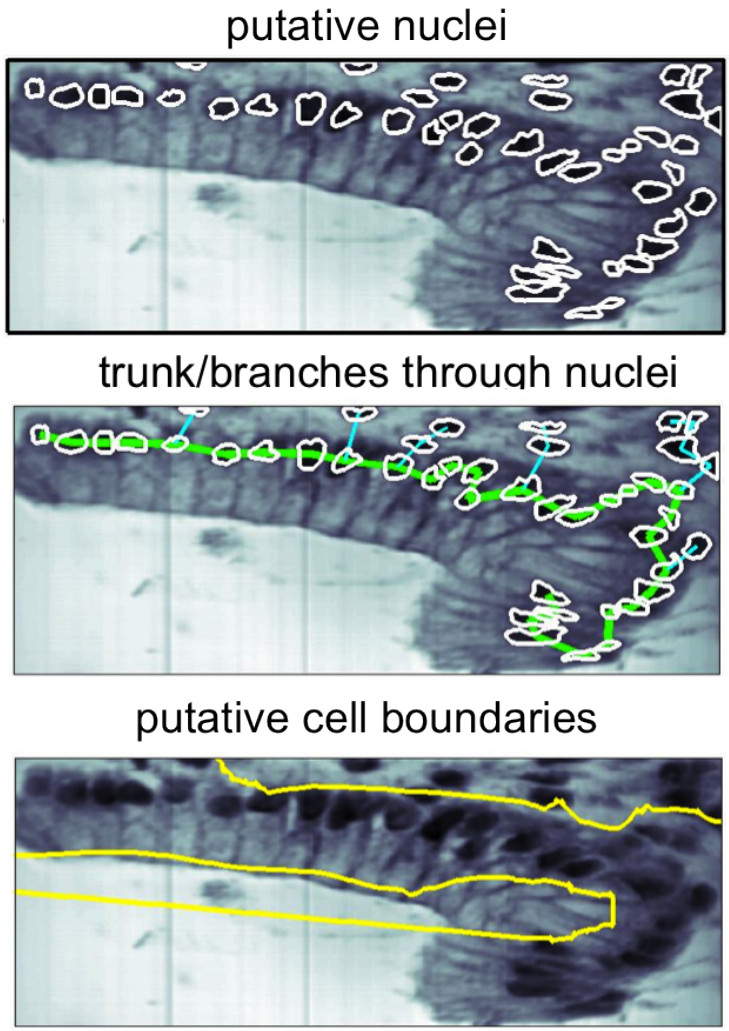

Figure 17: Encoding Context: In Phase I, putative nuclei are predicted. The second row shows a trunk (green) with branches (cyan) built to model the nucleus architecture. The bottom row shows results from a Canny Edge detector meant to epithelial capture cell boundaries [3]. Note that neither result is a perfect model, use an approximation.

of a nucleus. Long, smooth edges representing cell boundaries were formed by first connecting nearby edges with similar slopes at their termini, and then removing short edges. Parameters for short edges, nearby edges, and similar slopes were determined empirically on the training set. Both of these algorithms are described in more detail below 3.5.3.

For each of these eight classes, a classifier function was trained using AdaBoost on the initial set of n_p pairwise features. Specifically, for each architecture-category $c \in C$, the conditional probabilities were learned that a nucleus is epithelial, given that its neighbor is epithelial ($\Psi_e^c : (\vec{x}, \vec{y}) \in \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \rightarrow [0,1] \subset \mathbb{R}$) and the probability that a nucleus is epithelial, given that its neighbor is non-epithelial ($\Psi_n^c : (\vec{x}, \vec{y}) \in \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \rightarrow [0,1] \subset \mathbb{R}$). Here, e denotes epithelial, n denotes non-epithelial, and the probabilities are symmetric ($\Psi_e^c(\vec{x}, \vec{y}) = \Psi_e^c(y, x)$ for the pair of nuclei (x, y) with feature vectors $\vec{x}, \vec{y} \in \mathbb{R}^{n_p}$, analogous for Ψ_n^c).

Conditional Markov Random Field [22, 71]

Maximization on a conditional random field yields an optimal class labeling (as epithelial or non-epithelial) for the putative nuclei in an image according to that field. Note that we still carry the term "putative nuclei" because some regions assigned to the non-epithelial class may not be nuclei at all; we only seek to classify these regions as not being epithelial, regardless of whether or not they are nuclei. We build an undirected graph in which each putative nucleus is a node, and place edges between nearby nuclei, as defined in the previous section. Let N denote the number of nodes (nuclei) in the graph, E denote the set of epithelial nuclei, and $\bar{E}$ denote the set of non-epithelial nuclei. The edge between nodes x and y, belonging to architectural category c, with feature vectors $\vec{x}$ and $\vec{y}$, is weighted with the pairwise conditional probability matrix

$$\begin{pmatrix} P(x \in E \,|\, y \in E) & P(x \in E \,|\, y \in \bar{E}) \\ P(x \in \bar{E} \,|\, y \in E) & P(x \in \bar{E} \,|\, y \in \bar{E}) \end{pmatrix} =$$

$$\begin{pmatrix} \Psi_e^c(\vec{x}, \vec{y}) & \Psi_n^c(\vec{x}, \vec{y}) \\ 1 - \Psi_e^c(\vec{x}, \vec{y}) & 1 - \Psi_n^c(\vec{x}, \vec{y}) \end{pmatrix},$$

for architectural class $c = c(x, y) \in C$. Each node x is also attached a pair of unary probabilities $\big(P(x \in E), P(x \in \bar{E})\big)^{\mathrm{T}} = \big(\psi(\vec{x_i}), 1 - \psi(\vec{x_i})\big)^{\mathrm{T}}$. The pairwise probability matrices are assembled for all nuclei pairs into the $(2N \times 2N)$ **b**inary probability matrix B, and the $(2N \times 1)$ **u**nary

probability vector $\vec{u}$. Let ω be a scalar factor determining the weight of the pairwise term in the optimization problem. Then, we solve:

$$\max_{v} \vec{u}^T v + \omega \, \vec{v}^T B \vec{v},$$

where $\vec{v}$ is a vector of N concatenated (2×1) vectors $\vec{v}_i$, such that $||v_i||_1 = 1, \, \forall i$. We adapt the two-phase algorithm from [71], which finds the optimal solution to this problem by first finding a global solution to a related problem in which the constraint $||\vec{v}_i||_1 = 1, \, \forall i$ is relaxed, then projecting the solution into the space of binary, unit-norm v_i's, and finally finding a local solution in the space of binary, unit-norm v_i's. As the labeling that maximizes the unary probabilities, v^U, already tends to be close to the ground-truth solution (see Table 3.5.4), we condense this process by performing local optimization directly, using v^U as a starting point. The algorithm is:

0. Initialize $t = 0$, $\vec{v}_t = \vec{v}^U$, score$_t = \vec{u}^T v_t + \omega v_t^T B v_t$, score$_{t+1}$ = score$_t + 2\epsilon$.
1. While $|\text{score}_t - \text{score}_{t+1}| > \epsilon$
   i. $t = t + 1$
   ii. $\vec{v}_t = \omega \, B \vec{v}_{t-1} + \vec{u}$
   iii. Normalize $\vec{v}$ on each node i such that $||\vec{v}_i||_1 = 1$.
   iv. score$_t = \vec{u}^T \vec{v}_t + \omega \, \vec{v}_t^T B \vec{v}_t$.

This method is a variant of the power iteration for finding the first eigenpair of a matrix and will converge [71, 72]. As our starting point is usually very close to the optimal solution, the convergence is usually rapid.

Example

Consider an image with only three nearby nuclei (Fig. 18). According to the ground-truth, nuclei 1 and 2 are epithelial and nucleus 3 is non-epithelial. Here, we demonstrate how the cMRF described above can predict these labels. We begin by assigning unary probabilities to each nucleus using the unary classifier $\psi(\vec{x}_i)$, where $\vec{x}_i$ is a set of 94 features computed on nucleus i, $i = \{1, 2, 3\}$, and $N = 3$. Let

$$\psi(\vec{x_1}) = 0.9, \; \psi(\vec{x_2}) = 0.9, \text{ and } \psi(\vec{x_3}) = 0.52.$$

In this case, the unary probabilities are strong indicators that nodes 1 and 2 are epithelial, but the unary probability of node 3 only slightly favors the epithelial label, which is actually false. By adding contextual information through pairwise probabilities, the labeling should be corrected. Assume nodes 1 and 2 are related with architectural class 1, and node 3 is related to each of these nodes with architectural class 3. Then we define:

$$\big(P(x_1 \in E \,|\, x_2 \in E\,), P(x_1 \in E \,|\, x_2 \in \bar{E}\,)\big) = \big(\Psi_e^1(\vec{x_1}, \vec{x_2}), \Psi_n^1(\vec{x_1}, \vec{x_2})\big) = (.95, .25),$$

$$\big(P(x_1 \in E \,|\, x_3 \in E\,), P(x_1 \in E \,|\, x_3 \in \bar{E}\,)\big) = \big(\Psi_e^3(\vec{x_1}, \vec{x3}), \Psi_n^3(\vec{x_1}, \vec{x_3})\big) = (.3, .8), \text{ and}$$

$$\big(P(x_3 \in E \,|\, x_2 \in E\,), P(x_3 \in E \,|\, x_2 \in \bar{E}\,)\big) = \big(\Psi_e^3(\vec{x_3}, \vec{x_2}), \Psi_n^3(\vec{x_3}, \vec{x_2})\big) = (.3, .8).$$

The probability that nodes 1 and 2 are the same class is high, because they have similar sizes, shapes, orientations, and other pairwise features, and are included in architectural category 1. The probability that node 3 is in a different class than nodes 1 and 2 is high because nodes 3 has a very different size, shape, orientation than both nodes 1 and 2, and the pairs (x_1, x_3) and (x_2, x_3) are in category 3, which encourages nodes to have different labels.

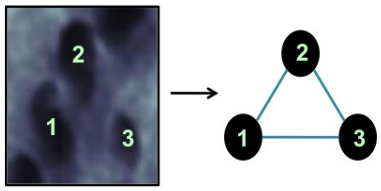

Figure 18: Example: Three nearby nuclei in image represented as three interconnected nodes in graph.

Using the unary probabilities, we define the unary probability vector U as: $U = (0.9.0.1, 0.9.0.1, 0.52, 0.48)^T$. The binary probability matrices define the matrix

$$B = \begin{pmatrix} 0 & 0 & 0.95 & 0.25 & 0.3 & 0.8 \\ 0 & 0 & 0.05 & 0.75 & 0.7 & 0.2 \\ 0.95 & 0.75 & 0 & 0 & 0.3 & 0.8 \\ 0.05 & 0.25 & 0 & 0 & 0.7 & 0.2 \\ 0.3 & 0.8 & 0.3 & 0.8 & 0 & 0 \\ 0.7 & 0.2 & 0.7 & 0.2 & 0 & 0 \end{pmatrix}$$

We initialize the solution vector $\vec{v}$ according to the maximal unary probability, therefore $\vec{v_0} = (1, 0, 1, 0, 1, 0)^T$. To find the optimal value of $\vec{v}$, we iteratively compute:

$$\vec{v_t} = \lambda B \vec{v_{t-1}} + U$$

and normalize on each node until convergence. We set $\lambda = 0.1$, as determined on the training set. One iteration gives us:

$$\vec{v_1} = B\vec{v_0} + U = (1.03, 0.18, 1.03, 0.18, .58, 0.62)^T.$$

The maximum argument of $\vec{v_1}(1:2)$ gives us the label assignment of node 1 at iteration 1, the maximum argument of $\vec{v_1}(3:4)$ gives us the label assignment of node 2 at iteration 1, and the maximum argument of $\vec{v_1}(5:6)$ gives us the label assignment of node 3 at iteration 1. Thus, the solution vectors are $v_1 = [1, 0]$, $v_2 = [1, 0]$, and $v_3 = [0, 1]$. Therefore, nodes 1 and 2 are labeled as epithelial, while node 3 is labeled as non-epithelial. The iteration continues until the sum $U^T v + \lambda v^T B v$ converges, at which point the final solution vectors v_i are computed and the labels are assigned. Here, the solution vector converges to $\vec{v_5} = [0.870.130.870.130.490.51]$, therefore nodes 1 and 2 are epithelial, and node 3 is non-epithelial. Thus, the context information encoded in the pairwise term are able to correct the initial unary probabilities to find the most logical class labeling of the entire image.

Correcting for isolated epithelial nuclei with a local smoothing factor

If one nuclei in a pair of non-epithelial nuclei has very different features than its neighbor, then the probability that this nuclei is epithelial, conditioned on its neighbor being non-epithelial, can be higher than the probability that both nuclei are non-epithelial, since the probabilities are trained to assign very different neighboring nuclei to different classes. In most cases, the unary probability that this nucleus is non-epithelial is strong enough to overpower the pairwise probability that it is epithelial, and the nucleus is correctly labeled as non-epithelial. However, if the nucleus is in a group of non-epithelial nuclei, and is very different than its neighbors, then the combined pairwise probabilities from all the neighbors that the nucleus is epithelial may outweigh the unary probability that that nucleus is non-epithelial, and the nucleus will be labeled as epithelial.To adjust for this, instead of a single smoothing factor λ, we scale λ for each node according its the number of neighbors, specifically: $\lambda_j := d(j)$, where $d(j)$ is the degree of node j. We show in Table 3.5.2 how using a local smoothing factor improves the MRF. Additionally, we have further improved the classification accuracy by following the MRF with a correction

	Training			Testing		
	FPR	TPR	accuracy	FPR	TPR	accuracy
MRF, sc.λ	6.7	89.2	91.6	10.4	82.5	86.7
MRF, loc. λ	8.3	88.6	90.4	11.6	84.0	86.6
MRF, sc. λ + corr.	6.6	89.2	91.7	10.2	82.5	86.8
MRF, loc. λ + corr.	5.4	88.2	92.0	8.9	80.9	87.0

step, in which isolated nuclei labeled as epithelial nuclei are re-assigned a label according to their maximal unary probability (See Table 3.5.2).

Choice of pairwise classifiers and parameters

To determine the most appropriate method for epithelial nuclei detection, we sampled a range of pairwise classifiers with a range of smoothing parameters. That is, for the problem $\max_{\vec{v}} \vec{u}^T \vec{v} + \omega \vec{v}^T B \vec{v}$, we varied B and ω, as well as the degree of connectivity. We sampled all combinations of the following cases:

- As an alternative to pairwise classification functions $\Psi^c(\vec{x}, \vec{y})$ dependent on both feature vectors and the archictecture-category c of each nuclei-pair, we employed fixed pairwise classification probabilities dependent on solely the architecture-category of the pair.
- We employed both local and scalar smoothing factors ω.
- We computed results with and without the correction step for isolated epithelial nuclei.
- We considered edges between only nuclei connected along the computed nuclei tree, versus edges between all spatially nearby nuclei.
- We considered two methods for predicting the nucleus tree.

Additionally, as the architecture-category of each nucleus pair is predicted using a greedy algorithm designed to model nuclei as a trunk with branches and a Canny edge detector to estimate cell boundaries, and may be imperfect, we computed the ground-truth architecture-category of each nucleus pair in terms of the nucleus trunk, the cell boundaries, or both. For these "ideal" cases, we also sampled each of the above classification functions over a range of smoothing parameters, to determine how well the algorithm would perform if these intermediate values were perfect.

The accuracy of each method was computed on the testing data set for a large range of

smoothing parameters ω. For each method that improved the accuracy by at least 1% over the accuracy with only the unary probabilities for some ω, we selected a subset of ω's close to that method's optimal ω, and evaluated the performance of the method on a random validation set (83 images randomly selected from the combined training/testing sets) 3.5.2. We chose the method with highest accuracy on the validation set as our classifier, together with the optimal ω for that method on the validation set. (Note that the results shown on testing set need not be greater than 1% over the unary classifier, as ω is first optimized on the validation set, and so a different ω may be used for the overall results than was initially used to select methods to test on the validation set.) This classification method was then used to predict epithelial nuclei on the experimental set (See Results).

3.5.3 Tissue architecture features

We initially defined a set of pairwise features between nuclei with the goal of discriminating pairs of epithelial nuclei from pairs of non-epithelial nuclei or mixed pairs of epithelial and non-epithelial nuclei. However, due to the large degree of randomness in the nuclei, these features did not sufficiently discriminate the nuclei. Pathologists use large-scale architectural features to identify regions containing epithelial nuclei and discriminate epithelial from non-epithelial nuclei within those regions. For example, pathologists identify chains of nuclei perpendicular to a lumen region which largely consist of epithelial nuclei, and identify goblet cell nuclei within these chains due to their different shape and orientation to the lumen compared to those of epithelial nuclei. By identifying chains of nuclei and cellular boundaries, we mimic pathologists' methods for using overall tissue architecture to identify epithelial nuclei 3.5.2. Once nuclei pairs have been assigned architectural classes, we use the initial set of pairwise features to discriminate between epithelial nuclei pairs and nuclei pairs containing non-epithelial nuclei.

Nucleus "tree"

To identify chains of epithelial nuclei, we build a spanning tree on the subset of nuclei that has a high unary probability of being epithelial, where the parameter ν is used as a threshold for high unary probability and is determined empirically on the training set, to optimize the number of epithelial nuclei found on a trunk. The greedy algorithm presented below seeks a spanning tree

method					Validation set results				Overall Results
tree	correction?	Ψ_c	λ	edges	FPR	TPR	acc.	Training acc.	Testing acc.
greedy	y	f	$\vec{\lambda}$	spatial	5.2	87.6	91.9	92.0	87.0
greedy	y	f	λ	spatial	6.5	88.0	91.3	91.7	86.8
greedy		f	λ	spatial	6.6	88.1	91.2	91.6	86.7
MST	y	f	$\vec{\lambda}$	spatial	8.1	88.8	90.7	90.8	87.7
MST		f	$\vec{\lambda}$	spatial	8.4	89.2	90.6	90.6	87.4
MST		f	λ	spatial	8.2	88.7	90.5	90.6	86.9
MST	y	f	λ	spatial	8.1	88.4	88.2		86.6
MST	y	f	$\vec{\lambda}$	tree	8.8	89.5	88.4		87.1
MST	y	f	λ	tree	9.0	89.6	88.4		87.1
MST		f	λ	tree	9.2	89.7	88.3		86.8
MST		f	$\vec{\lambda}$	tree	10.6	89.3	89.4		86.8
MST	y	l	λ	tree	9.3	87.3	86.9		86.8
MST		l	λ	tree	9.4	87.3	86.8		86.8
unary classifier, greedy					10.9	88.4	88.8	88.3	85.7
unary classifier, MST					11.2	88.6	88.7	88.3	85.8

Table 1: Results on validation, training, and testing set for best performing classifiers (any binary classifier that improved accuracy on testing set by more than 1% over unary classifier. Columns 1-5 describes the parameters used for each binary classifier.) The final two rows show results with only the unary classifier. The first column (tree) indicates whether the greedy tree algorithm described here was used to find the tree, targeting a straight trunk, or if a standard Minimum Spanning Tree (MST) algorithm was used to find the tree. The next column (correction?) indicates whether nuclei labeled as epithelial, but not neighboring any other epithelial nuclei, were assigned a corrected label according to their unary probability (y) or maintained their original label ( ). The third column indicates whether the pairwise terms were functions (f) or fixed values (l). The forth column, λ, indicates whether a local smoothing factor was used ($\vec{\lambda}$), or not (λ). The fifth column indicates whether edges in the MRF were placed only between nuclei with edges on the tree (tree), or between all nearby nuclei (spatial).

with a straight trunk and branches extending to nuclei that do not fit in a straight line along the trunk.

Setup: We define a graph H in which every nucleus with unary probability of being epithelial is larger than ν is a vertex. Edges connect nuclei pairs whose minimal distance is smaller than θ. Edge weights between nuclei pairs are given by $dtree$, a function of the distance between the nuclei, the ratio of nuclei sizes and orientations, and the largest angle formed by placing one of the nuclei on the vertex, one of the nuclei on a leg, and each possible neighboring vertex on the other edge.

Algorithm: Greedy Trunk

Initiate: We begin the greedy search by identifying the vertex $v \in G$ with highest probability of being epithelial for which there is a neighboring nucleus with edge weight smaller than δ. This vertex forms the initial trunk node.

Iterate: Identify the vertex with minimal distance to one of the two (or one in first iteration) trunk termini. If this distance is smaller than δ, add vertex to trunk.

Terminate: If minimal distance to trunk termini is larger than δ, terminate.

Algorithm: Add branches

Initiate: Let $T =$ trunk. Add all putative nuclei from image to H, including nuclei with unary probability $< \nu$. Let S $= \{$vertices in H which are not in $T.\}$

Iterate: Find vertex pair $(v \in S, u \in T)$ such that $v = \min_{s \in S, t \in T} dtree(s,t)$. If $dtree(v,u) < \delta_b$, add branch edge between vertices v and u, and move v from S to T.

Terminate If $dtree(v,u) > \delta_b$, terminate and leave remaining nuclei off tree.

After running *Greedy Trunk* and *Add branches* once, check if any branches are overall straighter or longer than the tree extended from the branch terminus. If so, replace the tree region with the branch.

Nuclei pairs are then labeled as being neighbors on the trunk, on a branch, at trunk-branch or branch-branch juncture, or in space. These tree labels, together with orientation of nuclei pairs with respect to cell boundaries, as defined below, determine the architectural class of each nuclei pair.

Cell boundaries

As epithelial cell cytoplasms tend to have a slightly different intensity than their surroundings, we use a Canny Edge detector to identify the boundary between epithelial cells and neighboring regions [3]. The width of the Gaussian smoothing factor selected for the Canny Edge detector is chosen to be approximately twice the diameter of the average cell nucleus. The initial edges predicted tend to have gaps, where intensity differences between epithelial cells and surroundings were not significant. Additionally, many false edges are detected due to color variation in the tissue, nuclei boundaries, and experimental artifacts such as wrinkles in the tissue. We first scan the image for neighing putative edge pairs with similar slopes (both near their termini and end-to-end slope) , and connect these edges to remove gaps in cell boundaries. Next, we remove all edges that are not at least as long as $\frac{1}{3}$ of the longest edge found in the image. Parameters for the canny edge detector and the edge threshold were determined empirically on the training set, in order to optimize the number of epithelial nuclei that fell on the same side of an edge. Nuclei pairs are labeled as crossing an edge or being on the same side of an edge. These cell-boundary classes, combined with the four classes predicted by the nuclei tree, form the eight architectural classes used in the Markov Random Field.

Ideal tissue architecture features

While learning parameters for the pairwise features, it was observed that some of the intermediate steps did not yield perfect results. Specifically, the nucleus chain, which ideally would place all epithelial nuclei along the main chain and non-epithelial nuclei on branches, does not always capture the longest chain of epithelial nuclei, and the epithelial boundaries detector sometimes incorrectly labels boundaries. In order to determine how much better the algorithm would perform if either of these methods were perfect, we created ideal versions of these features for each images, and ran the MRF using the ideal versions. We found that if both the nuclei chain and epithelial edge labelings were ideal, and a look-up table of probabilities learned from the datareplaced the continuous pairwise classification functions, we could achieve a true positive rate of 94% and a false positive rate of 4% on the testing data.The incorrect labelings in this case were largely due to ambiguous nuclei.The look-up table is expected to give better results than continuous functions in the case that the tree and cell boundaries are ideal, as in an ideal tree all trunk nuclei are epithelial and all branch nuclei are non-epithelial, and nuclei on oppo-

site sides of epithelial region boundaries will always have opposite labels, and thus a continuous function would permit unnatural options that are prevented by the look-up table. Using function classifiers, the true positive rate decreased to 91% and the false positive rate increased to 9% We analyzed classification results on the training and testing set using pairwise probabilities computed from varied combinations of ideal features, computed features, look-up tables, and functions. Overall, look-up tables are not significantly better classifiers than continuous functions, unless the underlying data is ideal. We found that using both putative cell boundaries and the predicted nuclei tree provided the most accurate predictions of predicted epithelial nuclei, compared to using only one of the two architectural descriptors or only unary probabilities (data not shown).

3.5.4 Results: Epithelial Classification

The images were randomly split into a training set of 331 images (80%) and a testing set of 83 images (20%). To validate the epithelial classification, all putative nuclei were labeled as epithelial or non-epithelial on all 414 images by Virginia Burger, and corrected/verified by pathologist Dr. Doug Hartman. Parameters for both unary and pairwise classifiers were learned on the training set. In Table 3.5.4, we show the epithelial classification results using (a) only the unary classifier, and (b) both unary and pairwise classifiers, on both the training and testing sets. Improvement in both increased true positive rate (TPR) and decreased false positive rate (FPR) are observed with the addition of the pairwise classifier. Figure 3.5.4 shows an example of improvement in accuracy through addition of context information encoded in the MRF.

	Training			Testing		
	FPR (%)	TPR (%)	acc.	FPR (%)	TPR (%)	acc.
unary	10.9	87.3	88.3	13.3	84.0	85.7
cMRF	8.4	89.5	92.0	10.9	85.6	87.0

On a 2012 MacBook Pro (2.9GHz Intel Core i7, 8GB memory), initial nuclei segmentation takes approximately 120 seconds for an average sized pixel image. The epithelial classification takes around 60 seconds, thus the algorithm spends on average of 180 seconds per image. By running the algorithm overnight, significant time is saved over the several minutes required for a researcher to manually outline each epithelial nucleus in an image.

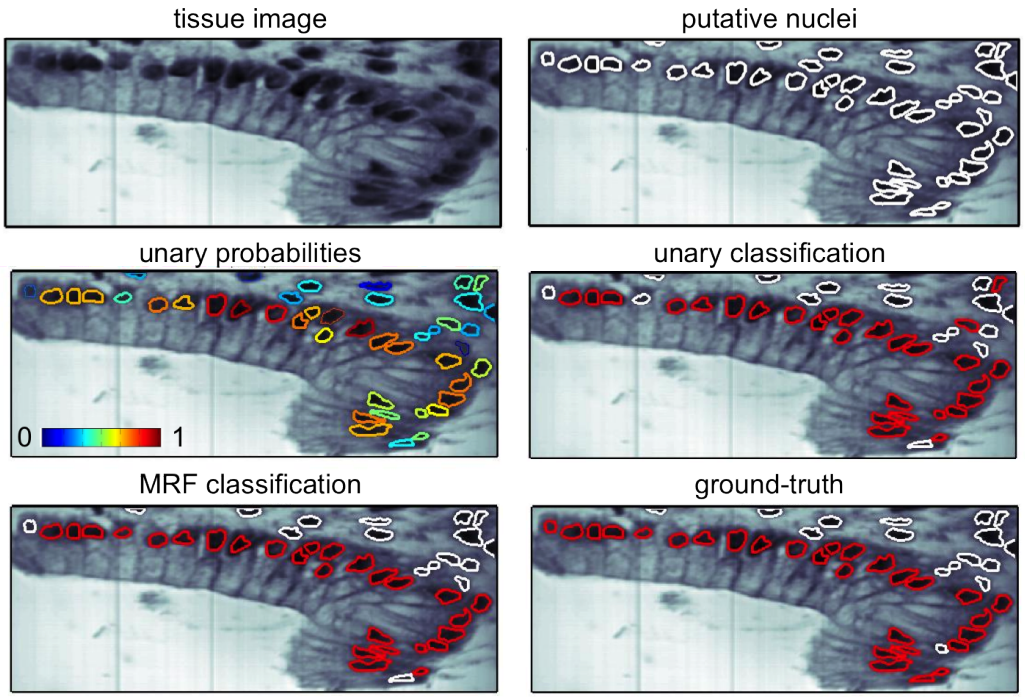

Figure 19: Epithelial classification: For the initial image (top-left) with ground-truth nuclei labeling as in bottom-right (red = epithelial cell nuclei, white = other nuclei), putative nuclei are predicted in Phase I (top-left). The unary probability of these nuclei being epithelial is shown in the middle-left, and all nuclei with unary probability greater than 0.5 could be classified as epithelial, as in middle-right. By using contextual information encoded in a MRF, the classification improves (bottom-left).

Automated versus manual selection

On the experimental data set of 424 images, an independent researcher, KS, generated putative nuclei using the above nuclei segmentation algorithm and manually selected around 10 epithelial nuclei from the set of putative nuclei from each image for phase analysis. Overall, 4095 nuclei were manually selected, while 7045 nuclei were automatically selected. For 3224 of the manually selected nuclei (78.7%), an automatically selected nucleus shared at least half of its pixels. For 80.1% of the manually selected nuclei, an automatically selected nucleus overlapped by at least one pixel. Note that we expect many more nuclei to be selected automatically than manually, as the algorithm seeks every epithelial nuclei, while KS only sought around 10 representative nuclei per image. The automated algorithm tended to miss epithelial nuclei that were isolated, as the pairwise nature of the algorithm encourages epithelial nuclei to appear in chains.

Feature	FPR	TPR	FPR	TPR
median pixel intensity	28.4	90.1	32.1	88.7
Area	26.8	77.1	26.9	75.3
mean-nuc-back	31.9	76.3	34.3	76.7
med-nuc-back	40.3	91	43.5	91
χ_2-distance between nucleus and surrounding intensities	36.3	71.1	34.5	69.1
average length of closest canny edge, $\sigma = 20$	33.6	59.9	36.8	67.8
distance to closest cell boundary	36.9	66.1	37.8	65.1
shared-edges-double-20	34.3	60.5	39.3	66.8
average length of closest canny edge, $\sigma = 25$	42.8	69.4	44.7	73.5
shared-edges-single-20	36.8	60.9	38.3	64.6
length of second closest canny edge, $\sigma = 20$	33.2	54.2	34.7	60.7
distance to second closest canny edge, $\sigma = 30$	38	61.3	38.7	64.4
length of first closest canny edge, $\sigma = 25$	38.5	58.9	38.7	63.2
shared-edges-single-25	47.1	70.8	46.1	72.8
average length of closest canny edge, $\sigma = 15$	36.4	59.9	38.4	62.5
-avg-dist-to-two-edges-sig-30	43.4	67.7	44.6	69.8

average distance to closest two cell boundaries	27.8	51.7	30.9	55.8
mean-neigh-orien-diff	42.9	65.1	39.6	63.3
average distance to two closest canny edges, $\sigma = 25$	36.8	59.3	38.1	61.2
distance to second closest canny edge, $\sigma = 45$	38.6	59.7	38.8	61.5
length of second closest canny edge, $\sigma = 15$	38.4	58.2	39.3	61.8
avg-length-edge-sig-30	33.5	57.8	35	57.8
average distance to two closest canny edges, $\sigma = 45$	41.8	62.5	40.4	62.7
angle to closest cell boundary	39.3	58.7	40.8	63.1
distance to second closest canny edge, $\sigma = 25$	34.7	57	36.5	58.6
shared-edges-double-10	30.7	54.6	33.3	55.9
shared-edges-double-15	37.7	62.9	43.3	65.4
dist-to-second-edge-sig-40	37.6	60.7	39.7	61.2
dist-to-second-edge-sig-35	46.4	71.1	48.1	71.1
length-first-edge-sig-30	48.3	67.5	48.5	71.4
length-first-edge-sig-20	43.2	64.6	45.9	67.2
perc-nuc-overlap-with-edge-sig-15	34.6	53.1	34.1	55.1
-avg-dist-to-two-edges-sig-35	47.8	72	48.6	71.3
-avg-dist-to-two-edges-sig-10	44.3	71.6	47.3	69
shared-edges-double-30	36.7	58.9	39.7	59.6
shared-edges-single-30	41	65.2	39.6	59.4
-avg-dist-to-two-edges-sig-40	44.3	66.9	43.3	63.2
dist-to-first-edge-sig-45	38.8	59.6	33.6	54.1
shared-edges-single-15	35.8	54.6	37.2	56.9
dist-to-second-edge-sig-10	45.3	72.4	48.4	69.9
shared-edges-single-35	43.7	64.1	41.5	60.8
dist-to-first-edge-sig-35	47.3	69.1	45.3	65.2
dist-to-first-edge-sig-25	42.2	60.4	41.4	60.6
length-first-edge-sig-35	41.3	59.2	44.3	63.9
-avg-dist-to-two-edges-sig-20	38.8	56.9	38.5	57.7

dist-to-first-edge-sig-40	42.1	62.9	39	58.1
length-second-edge-sig-25	29	49.5	32	52.3
shared-edges-double-25	47	70.6	50.5	72.6
dist-to-first-edge-sig-30	49.9	71.2	49.7	70.9
shared-edges-single-45	49.1	65.8	45.3	63.6
dist-to-second-edge-sig-20	45.6	64	44.8	62.9
length-second-edge-sig-30	45.9	68.6	47.9	67
length-first-edge-sig-15	45.4	65.4	48.3	67
shared-edges-single-40	45.6	63.2	40.1	57.4
perc-nuc-overlap-with-edge-sig-10	45.8	69.4	49.2	67.3
avg-length-edge-sig-40	45.5	66	51	70.1
shared-edges-double-35	42.4	60.3	42.5	58.9
length-first-edge-sig-40	44.2	61.1	48	64.2
area-convexarea	47.4	66.7	48	64
length-second-edge-sig-10	36	51.5	36	51.8
avg-length-edge-sig-35	27.4	46.2	31.7	48.5
median-int-back	30.7	56.3	37.6	52.6
only-one-close-edge-25	59.5	79.1	57.9	82.7
length-first-edge-sig-10	54.4	72.7	53.6	71.8
shared-edges-double-40	42.5	58.7	39	53.5
only-one-close-edge-30	60.2	80	58.3	82.5
length-second-edge-sig-40	41.2	59	45.5	58.4
avg-length-edge-sig-45	35.5	52.8	33	47.9
length-second-edge-sig-35	39	58.4	44.9	57.5
only-one-close-edge-20	58.9	76.3	58.7	79.2
only-one-close-edge-45	61.5	77.6	57.3	75.7
dist-to-first-edge-sig-20	41.2	51.2	40.5	52
avg-length-edge-sig-10	28.3	43.9	28.7	44.1
length-second-edge-sig-45	46	61.2	51.3	61.3

length-first-edge-sig-45	49.4	65.7	54.1	65.3
only-one-close-edge-15	56.5	67.5	56.3	68.6
only-one-close-edge-40	62.3	79.6	60.4	77
-avg-dist-to-two-edges-sig-15	53.1	67.8	54.6	65.4
only-one-close-edge-35	62.5	82	62.8	84.2
third-neigh-dist	24.6	43.5	31	43.1
perc-nuc-overlap-with-edge-sig-30	64	82.4	62.4	81.4
std-nuc+neigh	59.4	74.1	59	69.5
perc-nuc-overlap-with-edge-sig-25	68.2	85.4	65.1	84.6
avg-distance-to-closest-3-edges	19.1	32.4	24	37.5
shared-edges-double-45	59.9	70	56.5	62.1
only-one-close-edge-10	54.2	60.3	53.5	57.1
dist-to-second-edge-sig-15	62.7	77.9	64.8	74.8
perc-nuc-overlap-with-edge-sig-35	69.8	87.2	69.5	86.6
dist-to-first-edge-sig-15	68	80.5	68.2	80.2
dist-to-first-edge-sig-10	67.1	88.3	70	88
perc-nuc-overlap-with-edge-sig-20	72.6	86.3	70.7	88.5
shared-edges-single-10	69.8	77	70.6	75.8
perc-nuc-overlap-with-edge-sig-40	73.1	89.2	73.9	87.3
perc-nuc-overlap-with-edge-sig-45	75.5	89.6	76	89.2
combined features (AdaBoost)	12.9	88.6	13.8	88.2

Table 2: True and False positive rates shown on training (columns 2-3) and testing (columns 4-5) sets for the unary features. The bottom row shows the FPR and TPR for the combined classifier generated with AdaBoost. Each training set consisted of around 15387 nuclei from 332 images, and each testing set consisted of around 3935 nuclei from 882 images. In total, there were 11459 ground-truth non-epithelial putative nuclei and 7863 ground-truth epithelial nuclei. For canny edge features, σ indicates the size of the Gaussian filter used for smoothing. Note that these results are for the combined nuclei set taken from all images - a single image can yield both training and testing nuclei. As epithelial classification depends on neighboring nuclei, entire images are labeled as either testing or training for validating epithelial segmentation. Thus, the overall training and testing accuracy here will be slightly different than the accuracy shown for the unary classifier in Table 3.5.4.

3.6 OPTICAL BIOMARKER FOR CANCER RISK IN BE

In this section, we show that (a) automatically selecting nuclei produces an equivalent or larger set of epithelial nuclei as manually selecting nuclei, and (b) distributions of features computed on the phase of epithelial cells can be used as an optical biomarker for cancer risk in BE. When computing phase on the predicted epithelial nuclei, we ignore nuclei on image boundaries, as the pixel intensities near the boundaries are generally much darker than in the image interiors do to

intensity fall-off.

An average nucleus has approximately 800 pixels, and phase is computed on every pixel in every epithelial nucleus. To summarize the distribution of phases on a nucleus, we compute the entropy as $H^b = -\sum_b p_b \log(p_b)$, where b indicates a binning index. We use 51 bins of length $\frac{\pi}{25}$ to discretize the phase at each pixel. Additionally, we analyzed the distributions of (1) mean phase on each nucleus, (2), mean standard deviation on each nucleus, (3) mean nucleus phase on each image, (4) mean amplitude on each nucleus, and (4) nuclei pixel phases, across each diagnostic set, and found each measure to have statistical significance for differentiating the diagnostic classes.In Figure 3.6, we see that the phase entropy within nuclei increases as the diagnostic class worsens from BE-normal to BE-HGD to BE-EAC, for depths 1-2. For depths 3-4, the entropy decreases along this same pathway (not shown).

We show the p-values describing the probability that the entropy distributions from any pair of diagnostic classes were generated from the same distribution in Table 3.6 for both automatically and manually selected nuclei. Given a cutoff for significance of p-value < 0.05, both the manual and automatic nuclei have significantly different distributions for each diagnostic class in at least one, and almost all, phase depths. The HGD and EAC classes are hardest to separate, while the BE-EAC classes are easiest to separate.

depth	BE-HGD	HGD-EAC	BE-HGD
	Manual		
1	0.0000	0.4719	0.0000
2	0.0000	0.8206	0.0000
3	0.0006	0.0009	0.0000
4	0.0219	0.0001	0.0000
	Automatic		
1	0.0017	0.0009	0.0000
2	0.0000	0.0954	0.0000
3	0.0000	0.0003	0.0000
4	0.0001	0.0005	0.0000

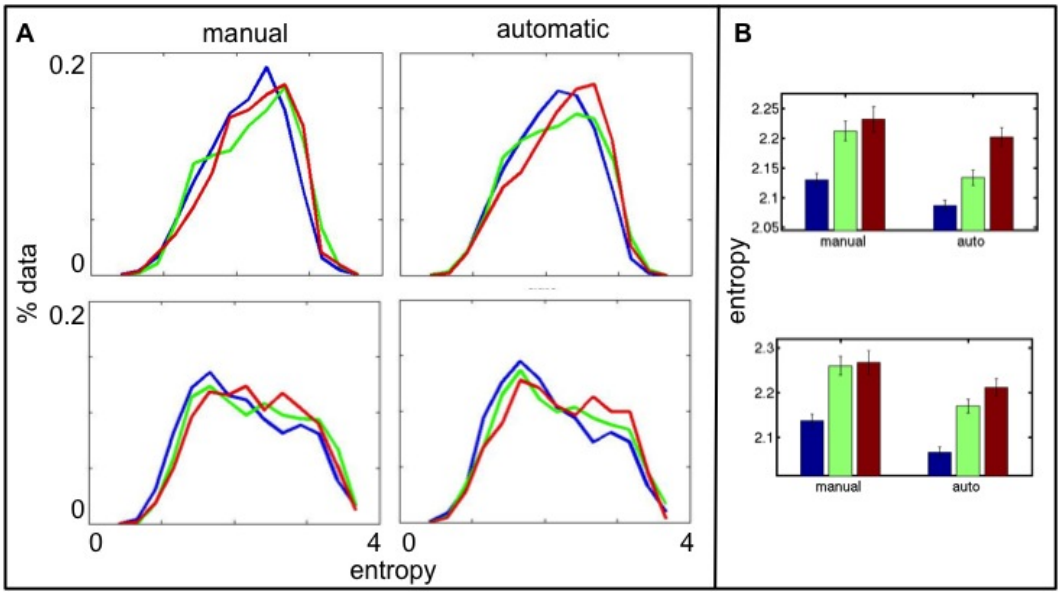

Figure 20: Entropy distribution on nuclei at depths 1 (top row) and 2 (bottom row), using manual or automatic selection. The right panel shows the mean, where error bars indicate standard error, for each diagnostic class, using manual or automatic selection. Blue indicates healthy tissue, green indicates HGD-adjacent tissue, and red indicates EAC-adjacent tissue.

3.7 DISCUSSION

Epithelial Classification The overlap between automatically selected nuclei and manually selected nuclei from the experimental set is around 80%. The epithelial nuclei detector specifically looks for chains of epithelial nuclei learned from pathologist annotations, and achieves an accuracy of around 90%. The independent researcher selected around 10 nuclei from each image from"columnar-shaped epithelial cells having similar morphological features such as intact nuclear boundary and no overlap of nucleus [2]", and his selections have not been validated by a pathologist. Thus, while the overlap between manual and automated selection is not perfect, there is no guarantee that the manual selection is perfect, and thus we have focused our analysis on agreement between automated and pathologist labelings. Moreover, an advantage of automated selection is the absence of user bias. Another significant benefit of automated nuclei selection is the time saved in nuclei selection; the automatic detector was able to identify almost twice as many epithelial nuclei as the manual detector with almost no time effort by the researcher.

As the automatic nuclei detector finds almost twice as many nuclei as manual nuclei selection, it would not be more surprising that the distributions between diagnostic classes are more often significantly distinguishable. However, if we remove a random set of the automatically selected nuclei, so that the distributions are of identical size, the automatic nuclei still yield significantly different distributions for each class (data not shown). Thus, the false positives in the automatic nuclei selection do not decrease the statistical significance of the results. The automatically selected nuclei may more often yield significantly differentiable diagnostic classes than the manually selected nuclei due to lack of bias in nuclei selection.

In Figure 3.6A (row 2), for depth 2, we see a second peak around phase $= 3$ in the entropy distributions for BE-EAC nuclei from the automatic nuclei, but not the manually selected nuclei. In both sets, there is a dip in the entropy distribution of BE-normal nuclei at the same entropy. Existence of nuclei with entropy within the range of this peak could be pursued as a possible discriminative feature for detection of early signs of cancer. Notably, the BE-HGD distribution falls between the BE-normal and BE-EAC distributions at this entropy. Figure 3.7 shows the phase distribution for each class at depth 1. In the zoomed figure, especially for the automatically

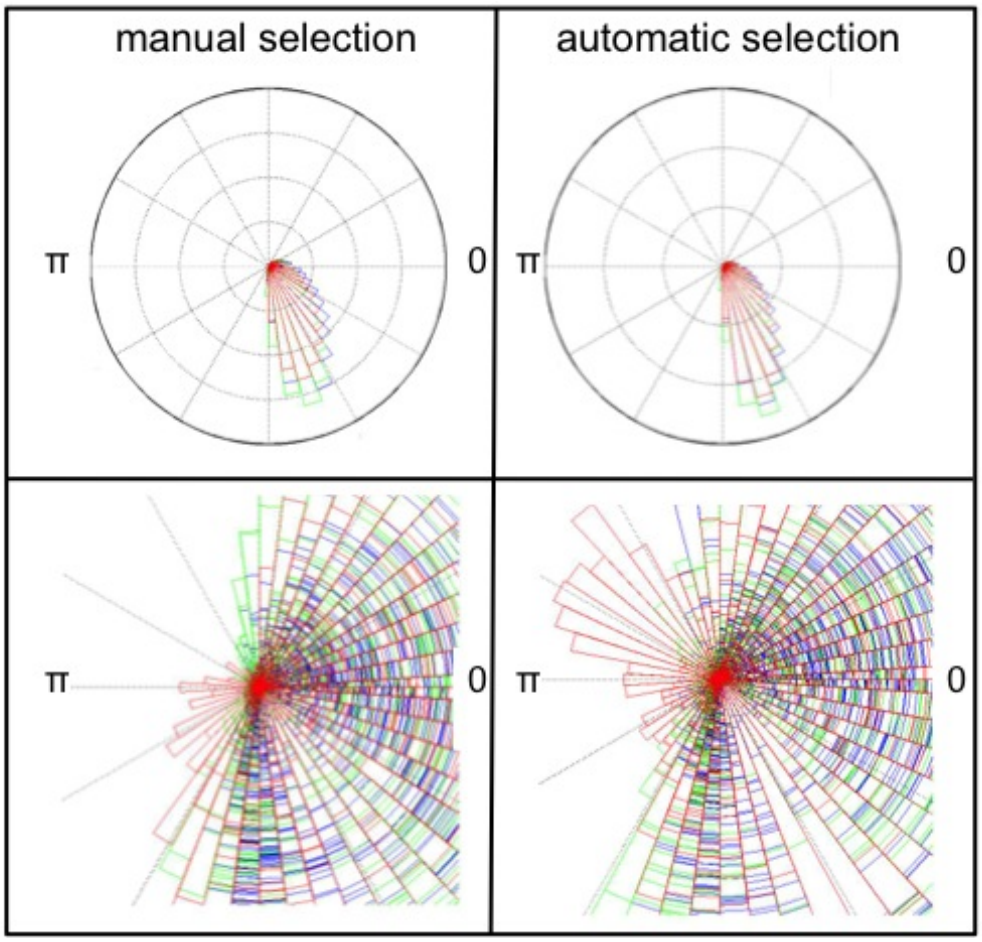

Figure 21: Average phase distributions on nuclei, averaged across each diagnostic class, at depth 1 are shown. Nuclei in the left panel were manually selected and nuclei in the right panel were automatically selected. Blue indicates BE-normal tissue, green indicates BE-HGD-adjacent tissue, and red indicates BE-EAC-adjacent tissue. The top row shows the full histograms, and the bottom row zooms in for visualization of low probability phases.

selected nuclei, we see that BE-EAC nuclei have phases in the range of π to $\frac{3\pi}{2}$, where almost no BE-EAC or BE-normal nuclei have phase density. This region could potentially be used as a classifier for detection of pre-cancerous changes in healthy tissue.

Recent discoveries have shown that micro-scale stromal nuclei patterns can also be indicative of cancer. As the features of stromal nuclei used to predict cancer are different than the features of epithelial nuclei, these two classes of nuclei must be examined independently. Thus, automized methods for epithelial classification are useful beyond the field of early cancer detection on the nano-scale.

4.0 BACKGROUND: HIERARCHICAL SPECTRAL CLUSTERING

4.1 BACKGROUND

Clustering is a natural approach to simplifying large sets of data by grouping similar data points into clusters. Consider the group of apples shown in 4.1. Each apple is a data point in this model data set. There are many ways of grouping these apples. For example, apples could be grouped according to their color: one cluster would contain only green apples, and a second cluster would contain only red apples. Additionally, apples could be clustered according to size, direction of step, or location in group. Furthermore, the apples could be grouped hierarchically, e.g. the apples could be first grouped according to color, and then each of these clustered could be further grouped according to size, and even further sub-grouped according to stem direction. Multiple potential clusterings exist for most data sets, and the objective of clustering must be considered when defining similarity between data points and specifying the number of desired clusters or the desired cluster size.

While the basic idea of finding similar data points and assigning them to a cluster is common throughout all algorithms, many different algorithms for clustering data exist. These algorithms vary slightly in their input, e.g. does the number or size of clusters need to be specified, and significantly in their methodologies. Linkage clustering is a set of local clustering methods that utilize similarity between neighboring data points, and group neighboring points that are similar (bottom-up) or split neighboring points that are dissimilar (top-down). K-means is commonly used local clustering method which iteratively adjusts cluster centers until all clusters contain points which are more similar to their cluster center than to any other cluster center. For k-means, the number of clusters, k, must be pre-determined.

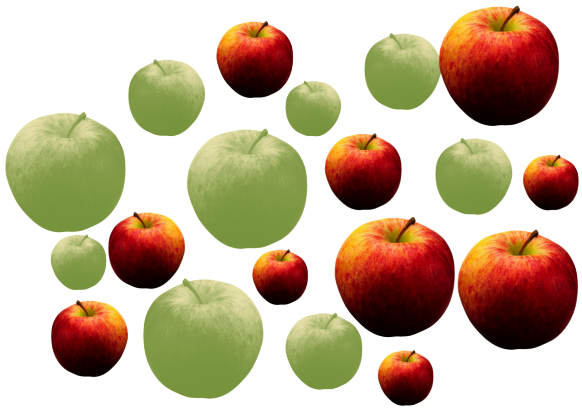

Figure 22: Group of apples.

Sometimes the number of clusters is not known. For example, in the group of apples, depending on how similarity is defined (color, color and size, stem-direction, etc.), different numbers of clusters would be desirable. For color or step-direction, there are clearly two clusters, while for size, there are three clusters. However, if we imagine that the data set is much larger than the group shown here (that is, that this is group is only a sub-sample of the entire group), and we consider that size and color are actually continuous variables, then there may be colors and sizes that are not included in this set. In this way, there may be more clusters than we see here. For example, there may be a small population of blue apples that were not included in the sample. If we cluster solely by color into two groups, then these blue apples would be included in the green cluster. However, if we let the number of clusters be unspecified, and cluster according to similarity in color, then we would obtain three clusters. For large data sets, we often can only preview a sub-sample of the data, and thus the number of clusters can be hard to accurately predict.

In Aims two and three, we seek to cluster large data sets into systems of states. In Aim 2, we look for similar image patches across a large set of whole slide lung tissue images on the size order of $20000 \times 20000 \times 3$ pixels per image. In Aim 3, we process long time-scale (on the order of 100000 frames) molecular dynamics simulation trajectories of a protein to find a set of conformational states visited by the protein. For both of these problems, we do not know how many states we expect to find, and the number of states would be expected to vary if we

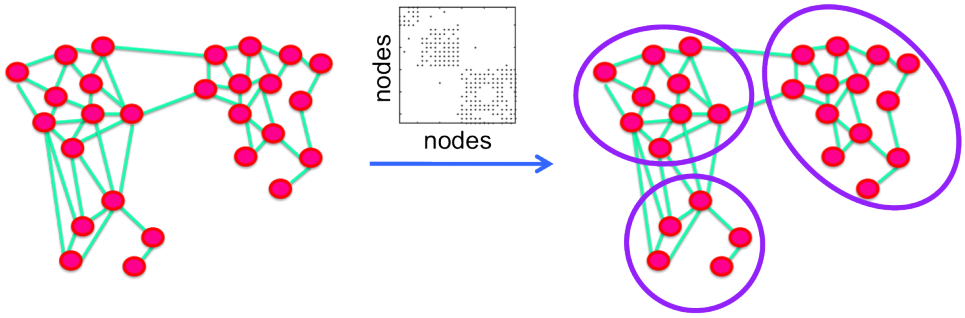

Figure 23: Random walk on a graph: A graph is shown in the left panel. This graph can be represented a connectivity matrix, shown above the arrow. By performing a random walk on the graph, clusters appear naturally between sets of highly connected nodes (right panel).

looked at images from a different organ or simulations from a different protein. For example, lung tissue and skin tissue can have different numbers of healthy and pathologic cell structure patterns. Similarly, different proteins have different numbers of states, depending on their degree of flexibility, their number of binding partners, and other factors. Additionally, for both problems we wish to produce a multi-scale representation of the data; we want to obtain sets of very homogeneous clusters with nearly identical data points, and also coarse clusters that group objects that have some shared features. In the case of tissue images, clusterings at different coarseness levels allow allow us to capture lung patterns of varied levels of homogeneity, corresponding to pathologic features at different size-scales. For protein simulations, clusterings of varied levels of coarseness allow us to describe protein transitions at varied time-scales. We thus choose to cluster both data sets using a hierarchical approach [73, 74, 75, 76, 77, 78, 79]. Specifically, we model the data as a graph, and perform a random walk on the graph to identify clusters. Each data point is a node in the graph and edges connect nearby nodes in the space. By performing a random walk on the graph, that is, starting at a node and moving successively from one node to another according to the strength of the edge weights, clusters in the graph occur naturally 4.1. The definition of "nearby nodes" is a crucial parameter for this (or any) clustering method. For example, in the group of apples shown in 4.1, apple color would probably be a more useful

feature for clustering the group of apples than the location of each apple's centroid. However, if the user desires apples grouped by size, not color, than size would be a more appropriate feature and color need not be included in the feature set. Similarity between a pair of apples would be defined as a function of the distances between each feature value for each apple. In chapters 5 and 6, we discuss how nearby nodes are identifies for Aims 2 and 3. In both cases, biological features drive the definitions of nearby, so that connected (neighboring) nodes according to our definition would also be deemed similar by experts.

4.2 ALGORITHM

Initiation: Let n_0 be the number of nodes (data-points) in the data set. Build an $(n_0 \times n_0)$ affinity matrix A describing the similarity between each pair of nodes. That is, $A(i,j) = 0$ if nodes i and j are not connected, and $A(i,j) =$ similarity between nodes i and j if the nodes are connected. Here, similarity is defined as a function of the distances between each feature for each node.In Chapters 5 and 6, similarity functions are defined for the set of lung images (Aim 2) and the set of protein conformations (Aim 3).

Ensure that the affinity matrix is connected and symmetric.If not connected (the matrix can be reduced into blocks), then perform the clustering on each component separately.

Set $A_0 := A$.

Iteration: For $t = 1$ until done:

1. Find Markov Transition matrix and stationary distribution of current graph: Compute the diagonal degree matrix D_{t-1}, with entries

$$D_{t-1}(i,i) = \sum_{j=1}^{n_t} A_{t-1}(j,i),$$

and

$$D_{t-1}(i,j) = 0 \quad \forall \quad i \neq j.$$

The degree matrix reflects the connectivity of the graph in that it contains the total number of connections to that node. Nodes with higher degrees can be seen as hubs, and nodes with very low degrees can be seen as isolates.

Then compute the Markov transition matrix

$$M_{t-1} = A_{t-1} D_{t-1}^{-1}.$$

The Markov transition matrix gives the probability of the random walker transitioning from one node to any of its neighboring nodes. The probability is 0 if the nodes are not neighbors. Each column sums to 1. Note that the Markov transition matrix is not usually symmetric. Next compute the normalized degree matrix

$$\pi_{t-1}(i) = \frac{D_{t-1}(i,i)}{\sum_j D_{t-1}(j,j)}.$$

The probability of a Markov Chain residing in a particular node after infinite iterations is given by its stationary distribution. For connected Markov transition matrices, the stationary distribution is trivially equal to the normalized degree vector, since $M_t \pi_t = \vec{1}\pi_t$.

2. Random Walk: Diffuse the Markov transition matrix by a multiplication

$$\hat{M}_{t-1} = M_{t-1} \times M_{t-1}.$$

This diffusion reveals distant connectivity and promotes cluster behavior by making probabilities within clusters more uniform.

3. Identify clusters: Prepare a kernel matrix K_t to carry network information from level $(t-1)$ of the hierarchy to level (t): First, find the nodes corresponding to local peaks of the stationary distribution ($\vec{\pi}_{t-1}$). Then, use the corresponding columns (kernels) of the diffused Markov transition matrix ($\hat{M}_{t-1}$) to form the $(n_{t-1} \times n_t)$ kernel matrix K_t, where n_t is the number of kernels found with $n_t \ll n_{t-1}$.

4. Build reduced graph: Solve

$$\vec{\pi}_{t-1} = K_t \vec{\pi}_t$$

for $\vec{\pi}_t$ with an expectation-maximization algorithm to find a low-dimensional representation $\vec{\pi}_t$ of the stationary distribution $\vec{\pi}_{t-1}$ [3].

5. Compute A_t and M_t, each of size $(n_t \times n_t)$, using $\vec{\pi}_t$ [3]:

$$M_t = \text{ diag}(\vec{\pi}_t)K_t^T \text{ diag}(K_t\vec{\pi}_t)^{-1}K_t$$

and

$$A_t = \text{ diag}(\vec{\pi}_t)K_t^T \text{ diag}(K_t\vec{\pi}_t)^{-1}K_t \text{ diag}(\vec{\pi}_t),$$

where K_t^T is the transpose of K_t and $\text{diag}(\vec{\pi}_t)$ indicates a diagonal matrix formed from the vector $\vec{\pi}_t$.

6. $t \rightarrow t + 1$

Termination: End if $n_t \leq 2$. Let $T = t$. At this point, the component has been divided into one or two segments.

4.2.1 "Goodness" of Clusterings

In Aims 2 and 3, we show goodness of clustering by comparing the overall similarity between nodes assigned to the same cluster to the overall similarity of nodes assigned to different clusters. In both aims, we have data points from multiple experiments (different images in Aim I, different trajectories in Aim II). We expect to see that some clusters contain data from multiple experiments, whereas other clusters may contain only data from one experiment. For example, most images are expected to contain some healthy tissue, so a cluster containing healthy tissue should contain tissue from multiple images. If this is not the case, there could be imaging artifacts (staining, shadows), that cause healthy tissue to be assigned to vary between images and thus be assigned to different clusters. In contrast, carcinomic tissue is expected to occur only in those patients with cancer. Thus, a cluster containing carcinomic tissue would be expected to only contain tissue from those patients. Similarly with protein simulations, highly stable protein conformations should be sampled by most trajectories and thus clusters containing these conformations should contain data from many trajectories. In contrast, rare states may only be accessed by a few trajectories, and thus clusters corresponding to these rare states would only contain data from the corresponding trajectories. Therefore, we examine the degree of mixing of trajectories

within clusters as a means of assessing the quality of the clustering approach and the definition of similarity for a particular data set.

In both aims, detailed ground-truth is lacking. For tissue images, we have diagnostic information at the image (patient) level, but not at the pixel level. Thus, to establish the goodness of the clustering, we show that the clustering agrees with the diagnostic labels (e.g. clusters exist which are specific to patients with carcinoma), and we obtain expert validation that each cluster contains diagnostically similar tissue. For protein simulations, we show that conformations assigned to the same cluster share biophysically relevant features, such as radius of gyration and internal energy.

4.2.2 Hierarchy Level

Similar to the apples 4.1, the image data in Aim 2 and protein simulations in Aim 3 can be clustered to varied degrees of homogeneity, each with its own use. In the following chapters, we discuss the information derivable from the clusterings at each hierarchy level and how a final hierarchy level could be chosen to best represent the data with respect to a particular question.

5.0 AIM II: COMPUTATIONAL STRATIFICATION OF DISEASE PROGRESS

We present a computational pathology schema for enabling early subtype diagnosis in Interstitial Lung Diseases (ILD). For most of the 130-200 diseases included in the class of ILDs, a full recovery is expected, but for a few of these diseases, the survival rate is less than three years. Treatment of the malignant forms of ILD would be harmful in patients with other forms, thus diagnosis is necessary prior to beginning treatment, and early treatment is most effective in eradicating disease. Early diagnosis is complicated by a high degree of sharing of subtle disease phenotypes, leading to high pathologist disagreement rates. To stratify ILD patients, we develop a novel quantitative representation of pathohistology samples that models lung architecture based on computed image features and insights from pathologists, and establish its utility as part of a diagnostic classifier. Unbiased, data-driven algorithms such as these applied in a clinical setting can save pathologists time by filtering out obvious cases and providing unbiased reasoning to assist diagnoses.

5.1 INTRODUCTION

Idiopathic interstitial pneumonias (IIPs) are a set of around $130 - 200$ chronic lung disorders, usually involving fibrosis of the lungs [80]. The diagnosis of these diseases has long been difficult because the diseases share many overlapping clinical, histologic, and radiologic features. Additionally, many IIPs are very rare, so many clinicians have limited experience with each subtype to rely on when making a diagnosis [81]. Since 2001, the new ATS-ERS classification, established by the American Thoratic Society and the European Respiratory Society, has been

followed for classification of these diseases. Interstitial pulmonary fibrosis [IPF] is the most common IIP and has the worst prognosis; patients with IPF have a median life expectancy of around three years, while most patients with the other IIPs have a high likelihood of recovery, especially if any environmental factors causing the disorder are removed. As different treatments are applied for each condition, early diagnosis is essential to begin appropriate treatment.

There is significant overlap in the diagnostic features for each IIP subtype and related disease. Interstitial pulmonary fibrosis is often referred to as usual interstitial pneumonia (UIP), which is the term for the morphologic pattern present in IPF. Non-specific insterstial pneumonia (NSIP) is the second most common IID, and its fibrotic subtype is commonly confused with UIP. Homogeneity of the lung tissue is a cardinal sign of NSIP, whereas IPF is hallmarked by heterogeneous tissue. The morphologic pattern of NSIP is also seen in hypersensitivity pneumonitis (HP), connective tissue diseases, and drug disorders, but NSIP itself is idiopathic. The smoking related IIDs, respiratory bronchiolitis-insterstitial lung disease (RB-ILD) and desquamative interstitial pneumonia (DIP) are believed to fall along a pathomorphologic continuum, in which DIP is the extreme form of RB-ILD. However, while a diagnostic criterium for RB-ILD is smoking, DIP may also occur in non-smokers and can be insidious. RB-ILD and DIP are not differentiable by standard histopathology methods.

While the histopathologic entitities (fibroblastic foci, lymphoid aggregates,) are common between the diseases, their locations with respect to each other and architectural components of lungs, such as the pleura or the interstitium (See Figure 5.1), are distinguishing factors between the diseases. Thus, **pathologists must make use of context information while analyzing the image.** For example, the spatial and temporal homogeneity in NSIP is a key feature in differentiating it from UIP, which has patchy lung involvement. The difficult in assessing the degree of homogeneity in histopathology slides has led to a high degree of inter-observer variation in distinguishing NSIP from UIP [82]. Here, we present a simplified representation of lung histology samples through their "architectural signature" (Figure 33). The architectural signature of a tissue is a 2D matrix describing the pairwise spatial arrangements of a set of histopathologic entities. **We hypothesize that these matrices can be used to describe the architectural layout of a tissue in terms of its pathohistology, and that architectural signatures can be used as**

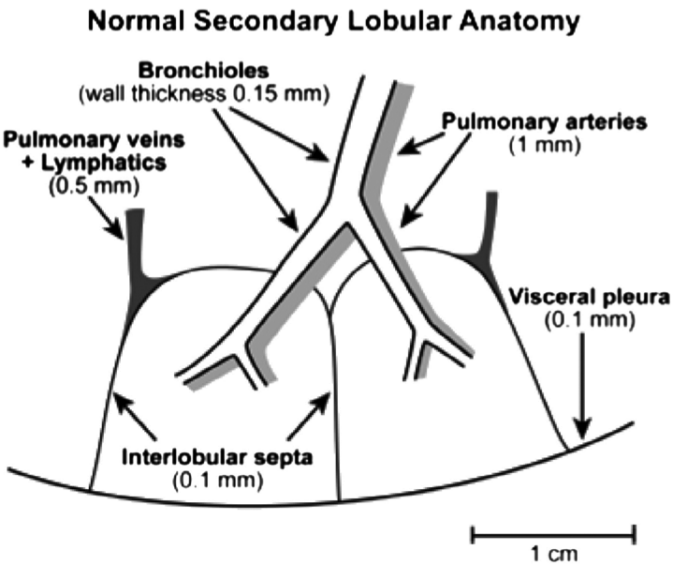

Figure 24: Secondary Pulmonary Lobule. Taken from Devakonda, 2010 [4].

part of a computational diagnosis strategy.

5.2 BACKGROUND

Histologic criteria is the basis for IIP classification. However, to capture histopathologic information, biopsies must be performed. Thus, non-invasive computed tomography (CT) is used in advance of biopsy to determine necessity of biopsy based on diagnostic information from the CT scan and to select a location for eventual biopsy. Overall, patterns detected in CT scans correlate well with histologic patterns and most computational image analysis for lung disease has focused on radiology images from CT. However, in ambiguous difficult cases, pathologic images must be used. Due to the high degree of inter-observer variability in diagnoses, a computational classification schema would be useful to provide a fast, unbiased diagnosis. Unlike pathologists, computers do not use intuition or risk intra-observer variation, and thus the computer can also provide specific reasons for its choice of diagnosis, as well as a confidence interval. While most

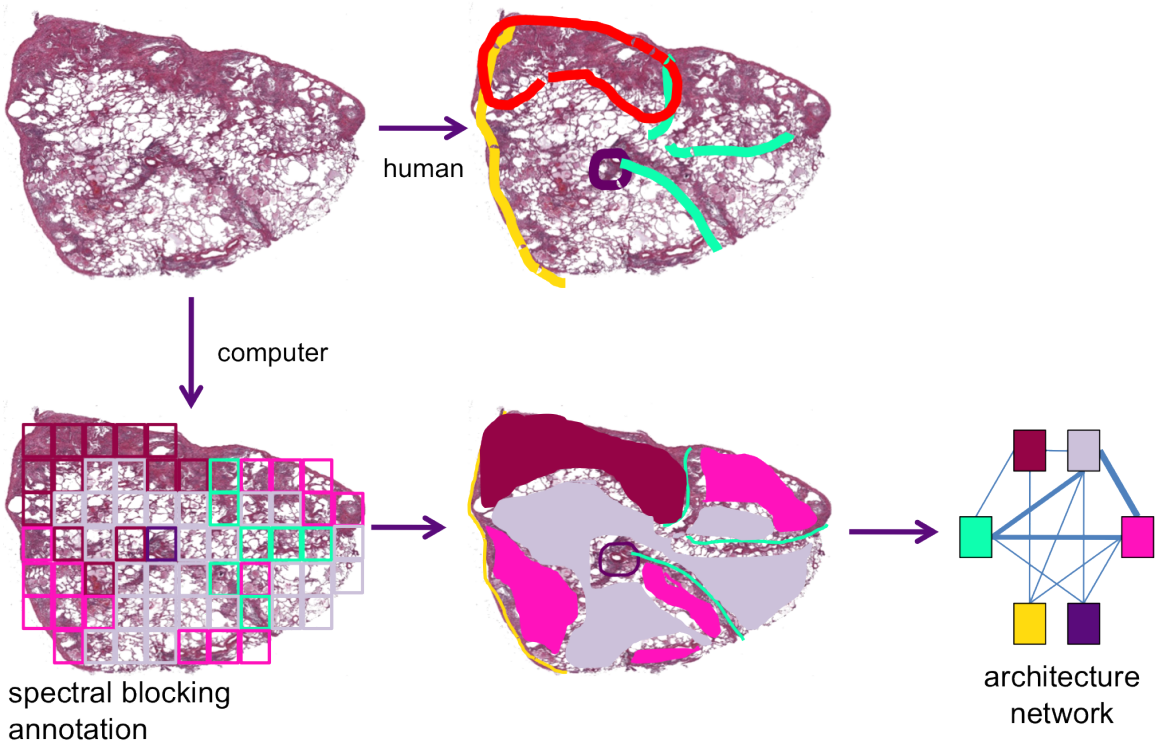

Figure 25: Quantifying context: While a pathologist observes higher-order architectural structures in lung tissue along with low-level diagnotic features, a computer sees only pixels. We train an algorithm which identifies homogeneous tissue regions, groups these regions to form diagnostically relevant tissue components, and build a spatial architectural matrix encoding context, which can be used as input to a diagnostic classifier.

computational work on lung disease has been focused on CT scans, computational algorithms have been developed for diagnosis and prognosis using whole slide histopathology images in many systems [83], such as neuroblastomas [84], prostate [85], and breast [86]. Most algorithms perform hierarchical analysis of the tissue in order to reduce computational complexity and incorporate global image aspects into local analyses [10]. [87]

Computational algorithms for histopathology commonly define a set of tissue classes and describe images as a weighted sum of these classes [85]. For lung tissue, for which diagnosis is very dependent on spatial locations of pathologic tissue, we take this representation a step forward by creating a simplified spatial model of the lung which describes the location of its tissue classes with respect to each other. We then demonstrate the potential of this model in a diagnostic classifier. The majority of computational models for lung disease focus on automated analysis of CT scans. However, in difficult cases, pathologists must look at lung biopsies to determine a diagnosis, and we present here one of the first computational analyses of whole-slide lung tissue images.

5.3 DATA

Our data, provided by the Lung Tissue Research Consortium (LTRC: http://www.ltrcpublic.com), consists of 63 whole slide H&E-stained images from 63 patients. The images ranged in size from 124×10^6 to 788×10^6 pixels, with average image size 21000×21000 pixels. For each image, clinical information was provided from throughout the patient's history. However, as the tissues provided were not necessarily diagnostic themselves, pathologist Frank Schneider (FS) labeled each image as diagnostic of one of seven categories: UIP, NSIP, fibrotic, other, control, emphysema, honeycomb, or non-diagnostic. Additionally, he selected a subset of 14 images which were clearly diagnostic of either UIP or NSIP to establish the potential of our method for differentiating diseased tissue 4. The majority of images that are control for IID in this data set originated in patients with a diagnosis of carcinoma.

The mean background image intensity was adjusted in each image so that all images have a

diagnosis	UIP	NSIP	fibrotic	control	other	honeycomb	emphysema	non-diagnostic
count	13(8)	6(6)	7	14	8	1	6	8

Table 4: Number of images that are diagnostic of each disorder / pattern. For UIP and NSIP, the number of clearly diagnostic images is provided in parenthesis.

mean background intensity in each channel of around 255 (white). To accomplish this, we define background as the set of tissue-free pixels within 1000 pixels of the image boundaries. The mean intensity was computed in the red, green, and blue channels on this set and the difference between 255 and this mean in each channel was added to all image pixels (capping at 255). As the initial difference in mean background intensities was less than 8 gray-levels (on a scale of 0 - 255), and all image features are computed over binned intensities, further normalization did not seem necessary and would have moved the analysis further from the raw data.

5.4 METHODS

As pathologists use both local and global information while determining a diagnosis, we seek to build an analogous multi-scale representation of the tissue image. In order to capture image information at multiple scales, we determine a hierarchy of increasingly coarse tissue histology [TH] states, and assign each image a state composition vector at each coarseness level. At any given coarseness level, we capture the spatial layout of the tissue through an architectural matrix, which describes the location of each TH-state with respect to every other state in a given image. Below, we first describe our approach to finding homogeneous tissue components. Next, we describe our hierarchical clustering method for grouping these homogeneous regions into increasingly coarse tissue component states. At this point, we introduce state composition vectors, which represent the percentage of each tissue made up of each TH. Then, we explain how we build spatial architecture matrices using the spatial arrangements of the TH-states in an image,

and, finally, we discuss how these matrices may be used to classify lung tissue data according to diagnosis.

5.4.1 Clustering

We represent each whole slide image as a vector of TH-state memberships at a set of coarseness levels. To determine the set of TH-states, we cluster across the entire image set to find unique **tissue histologies**, described through nuclei architecture and degrees of H&E staining. While image patches need not be neighboring to be assigned to the same TH-state, they must have similar tissue histologies. At coarseness level h, we define a set of N_s^h states, and describe the j^{th} image I^j as a vector $\vec{c^h} \in R^{N_S^h}$, where c_i^h = the percent of image j found in state i, for $i = \{1, \ldots, N_s^h\}$. At the finest hierarchy level, N_s^h is very large and the states are very homogeneous. At the coarsest hierarchy level, N_s^h is very small, and the states contain heterogeneous data. Our hierarchical clustering algorithm proceeds as follows: (1) Partition each image individually into a set of many homogeneous "microstates". These microstates are defined independently of other images. (2) Combine representative image patches from each microstate across all images into a large set $\mathcal{C}$. Cluster $\mathcal{C}$ iteratively into a hierarchy of TH-state sets with increasingly coarser clusters. (3) At each coarseness level, assign each microstate to a TH-state, and define the composition vectors $\vec{c}$ for each image. We describe each step in detail below.

Preproccesing on individual images: finding homogeneous microstates

Each image I_j is partitioned uniformly into a set of N_b square "blocks", which are large enough to capture local nuclei arrangements. Similar blocks are recursively grouped through spectral clustering to find a sets of microstates describing common histological patterns found in the image. A block size of 200 pixels, containing around 50 nuclei per block, is used. This block size was chosen empirically to consistently capture sufficient nuclei for distinguishing local architecture, while not being so large that blocks would commonly contain heterogeneous nuclei patterns.

Encoding pathologist knowledge into image features

Block similarity is defined through both stain similarity, defined by image intensities in the red, green, and blue [RGB] channels and through a set of features chosen to capture diagnostically

relevant patterns in the images [*5.4.1*]. Such features include size, shape, and arrangement of nuclei, guided by discourse with pathologists. We do not consider nucleus intensity, as variations in intensity due to staining across and between images would bias the small amount of expected intensity variation within nuclei. However, we do consider the intensity of the pixels immediately neighboring the nuclei, as these intensities can vary widely depending on the type of cell (white around lymphocytes, pink around epithelium) and pathologists consider cell types while inferring diagnoses, e.g. more lymphocytes could be indicative of inflammation. Additionally, we include Haralick features [88], which have been used to capture texture features for ILD classification in high-resolution computed topography images [89]. All features are normalized to have 0 mean and unit standard deviation.

In Figure 5.4.1, we show how the nuclei features are able to differentiate between blocks with nearly identical RGB distributions, but different tissue architectures. As the features do not necessarily follow any specific distribution, for each block, we compute the distribution of each feature on that block, as opposed to the mean, median, etc. Similar blocks are then found by computing the chi-sqared [χ^2] distance between the distributions.

feature class	feature	description
Morphometry (3)	axes ratio	distribution of ratios of minor to major axis length
	size	distribution of nucleus sizes
	small nuclei size	distribution of nucleus sizes for small nuclei
Appearance (3)	exterior R, G, B	distribution of intensities in R, G, and B channels in 2-pixel wide ring around nuclei
Architecture (6)	distance	distributions of all distances between every nucleus pair
	minimum distance	distributions of distances between nuclei and their closest neighbor

	median distance	distributions of median distances between each nucleus and all other nuclei in block
	mean distance	distributions of mean distances between each nucleus and all other nuclei in block
	maximal distance	distributions of distances between nuclei and their furthest neighbor
	standard deviation of distances	distribution of standard deviation of distances between each nucleus and all other nuclei in block
Spread (4)	Location Distributions	distribution of nucleus centroid locations across block. Block is divided into 4,9, 15, and 25 spatial bins on which distribution is approximated.
Block Texture	Histogram of Oriented Gradients (HoG) [90]	texture features, describe edges within block
	Haralick features [88]	texture features, describe gray-level patterns within block
Total = 18 features		

Table 5: Features selected to identify histologically similar tissue components.

We perform a simple rough nuclei segmentation by thresholding each block for pixels with intensity below an empirical threshold. While this method is not capable of separating tightly packed nuclei, as we are seeking to group blocks with similar nuclei architectures, the errors in nuclei segmentation are somewhat irrelevant, as long as the same errors are made consistently.

For example, if the nuclei segmentation always identifies a chain of epithelial nuclei as a single extremely elongated nucleus, then blocks containing this shape will be grouped, resulting in blocks containing epithelial chains being grouped. The features used are designed to accommodate the approximate nuclei segmentation; we do not look at high-resolution nucleus descriptors, but focus on rough morphometric descriptors and relationships between nuclei. This rough nuclei segmentation has the benefit of being extremely fast in comparison to segmentation methods that employ successive steps to break up large nuclei.

Microstates containing homogeneous blocks are identified by building a network in which each block is a node, and finding clusters on the network through a random walk. Edges are placed between nodes with similar R,G, B intensities, and edge weights are determined based on similarity between feature distributions. Specifically, for each node b_i, histograms of R,G, and B intensities on pixels containing tissue in the corresponding block are computed. Using χ^2 distances between these histograms, similar nodes are identified. Edge weights between the neighboring nodes are defined as a combination of the χ^2-distance between the features on the corresponding blocks. By performing a random walk on this network 4, clusters intrinsic to the network appear. To obtain highly homogeneous blocks within each microstate, we perform only one round of random walk, which produces around $\frac{N_b}{4}$ microstates for each image.

Specifically, we define the RGB distance between each node pair as a function of the χ^2 distances between the distributions of intensities in the red, green, and blue channels for each node. To obtain sparsity in the network, a threshold r on the χ^2 distances is determined such that each node has at least one neighbor, and neighboring nodes are defined as any pair of nodes whose χ^2 distance is below that threshold. Edges are placed between neighboring nodes. For the pair of neighboring nodes n_i and n_j, let $\mathcal{H}_i$, $\mathcal{H}_j \in \mathbb{R}^s$ be the RGB distributions on each node, where s is the number of histogram bins used to compute the distributions. The texture feature distributions for each node are contained in the matrices $\mathcal{T}_i, \mathcal{T}_j \in \mathbb{R}^{N_f \times s}$, where N_f is the number of features. For each feature $f \in \{1, \ldots, N_f\}$, we define a similarity measure for a χ^2 distributed variable as: $w_f(i,j) = \frac{1}{2\sigma_{ij}^f} e^{-\frac{\chi^2(T_i\vec{(}f), T_j\vec{(}f))}{\sigma_{ij}^f}}$, where $\sigma_{ij}^f = \sqrt{\text{median}(\{T_x\vec{(}f) T_y\vec{(}f) \,|\, x \sim y\})}$ and $x \sim y$ indicates that nodes x and y are connected by an edge. We take the arithmetic mean across all w_f as $w(i,j) = \frac{1}{N_f} \sum_{f=1}^{N_f} w_f(i,j)$ as the weight of the edge between nodes i and j. The

matrix $A = \{w_{ij}\}$ describes the network and is used for clustering as in 4.

Combining microstates into TH-states

For each image I_j, every image block is assigned to a microstate from the set of N^j microstates $\mathcal{M}^j$. As each image has its own set of microstates, at this point TH-state composition can not be compared across images. Thus, the microstates must be grouped across the images into one universal set of TH-states. To obtain these TH-states, we compute representative RGB distributions and feature distributions for each microstate by taking the mean of each intensity/feature distributions over all blocks within that microstate. As the blocks within each microstate are very homogeneous, the mean distribution is an accurate representation of each microstate. For visualization, for each microstate, we assign the block whose distribution is closest to that microstate's mean distribution as the representative block for that microstate. By considering each microstate as a node in the network and defining edges between nodes analogously to above, a network is built between the individual images. We perform hierarchical clustering on this network to find sets of TH-states of increasing coarseness 4.

5.4.2 Architectural Signature

Each image can be represented as a 2D matrix of block TH-state labels at each coarseness level (see Figure 5.4.2D). For an image of size $n_x \times n_y$ pixels, this TH-state label representation has size $\frac{n_x}{t_x} \times \frac{n_y}{t_y}$, and therefore is a significantly coarsened view of the image. As each TH-state captures a histologic pattern (see Results), this TH-state label representation provides a simplified view of the spatial arrangement of histologic patterns in the tissue image. To quantify the spatial arrangement of tissue components at a given coarseness level, we define an **architectural signature matrix** for each image at that coarseness level. Specifically, for coarseness level h with N_h TH-states, we define a $N_h \times N_h$ architectural signature matrix S_j^h for each image I_j. Matrix entry $S_j^h(s_1, s_2)$ gives the probability that a block assigned to TH-state s_1 is adjacent to a block assigned to TH-state s_2 in that image. We add two additional columns to this matrix to account for the likelihood that a TH-state neighbors empty space in the interior of the tissue (air sacs, arteries,..) or empty space exterior to the tissue. These tissue-free spaces are diagnostically significant, as different disorders have different amounts of fibrosis near pleura, interstitium, and

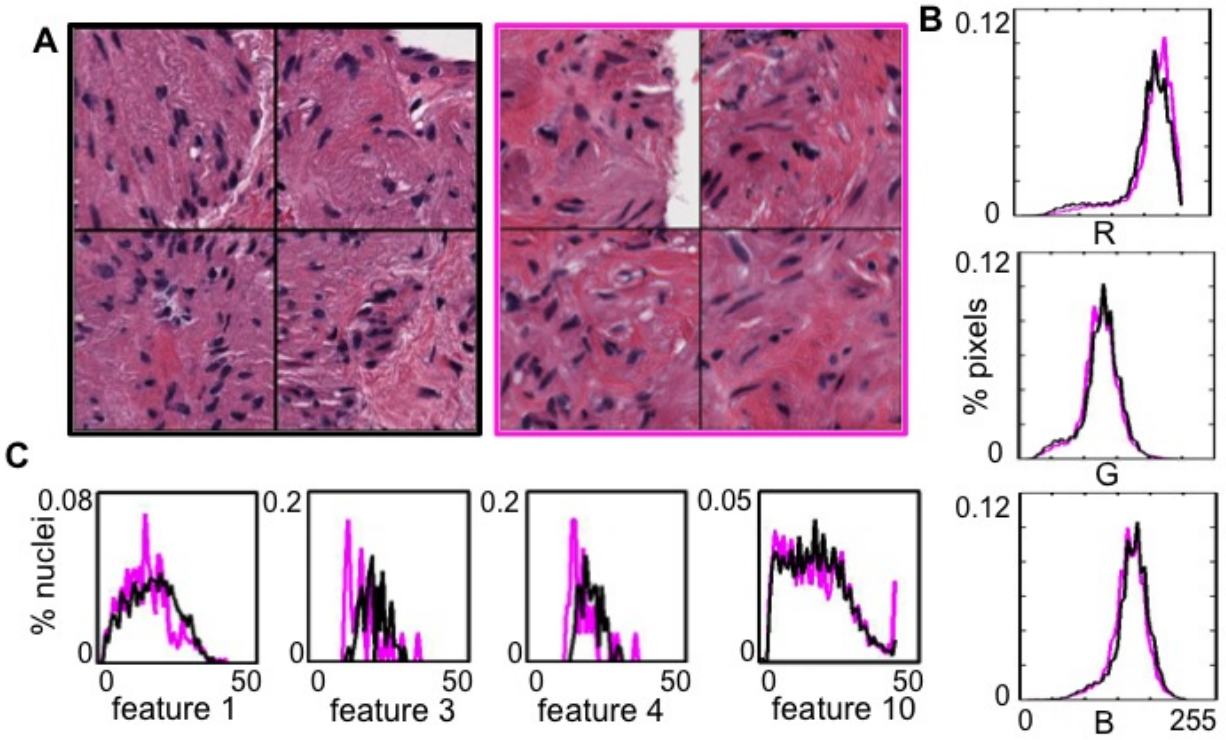

Figure 26: Nuclei features capture distinct patterns in regions with similar RGB distributions. A: Four image blocks each are shown from two microstates.In B and C, distributions computed on the left image are shown in black, and the right image are shown in pink. B: Histograms of mean (across all blocks in that TH-state) R, G, and B distributions on non-white pixels in each block are shown. C: Histograms of mean (across all blocks in that TH-state) feature distributions on nuclei from each block are shown for four features.

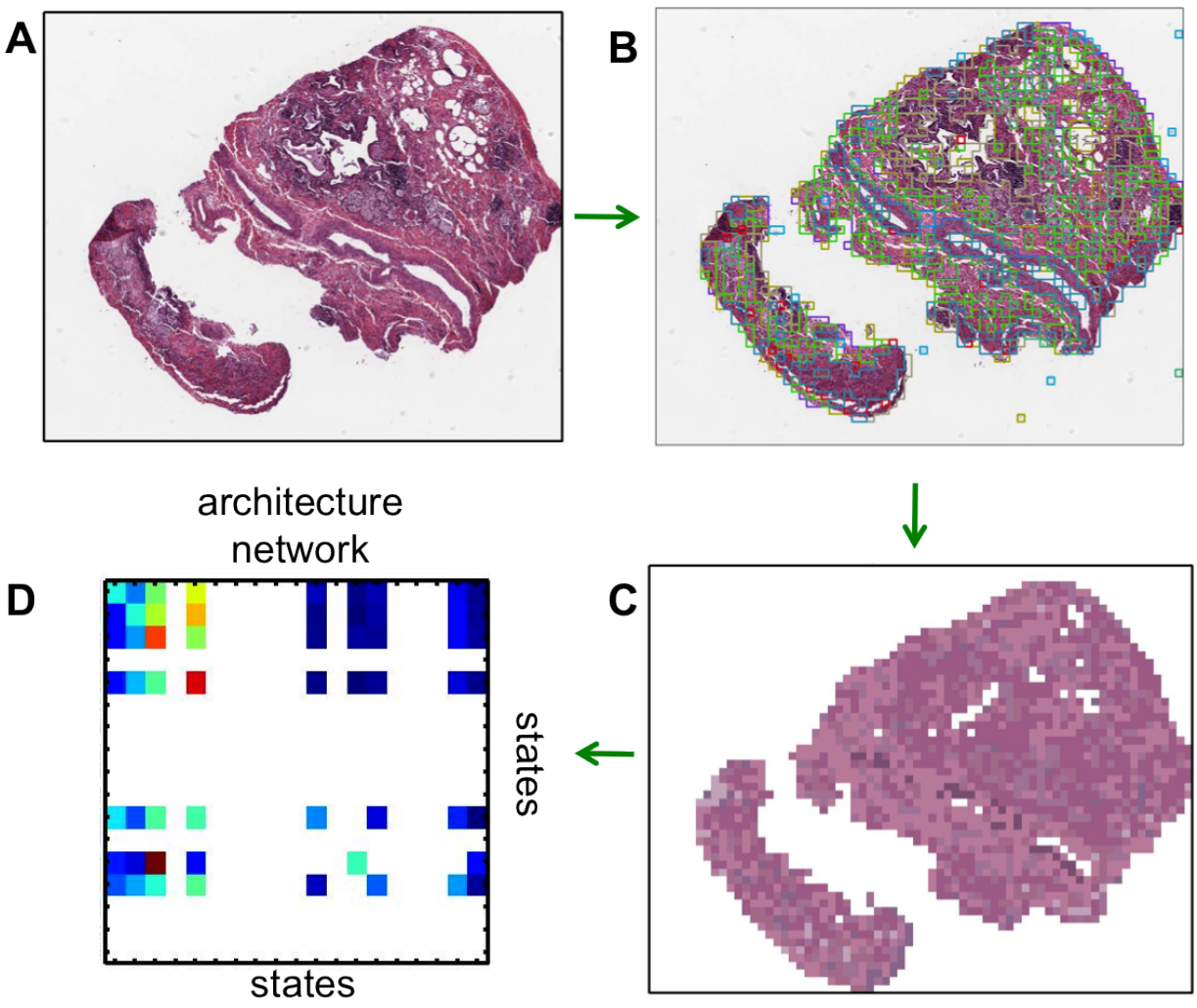

Figure 27: Representation of image through TH-state composition and architectural network at a single coarseness level. A: raw image. B: clustered image, where blocks assigned to the same TH-state are contained within the same color border. C: clustered image equivalent to B, where blocks are painted according to the mean image intensities within their TH-state. This yields a small-scale representation of the spatial layout of tissue components. D: Architectural network computed from C describing the likelihood that a given TH-state is spatially adjacent to every other TH-state. Red color indicates many neighboring blocks, blue color indicates few neighboring blocks, and white indicates no neighboring blocks.

other structural components of lungs. To compute these probabilities, we simply count the number of blocks of label s_2 adjacent to blocks of label s_1 and divide by the total number of blocks neighboring blocks of label s_1. In addition to architectural signature matrices for specific images, we compute analogous matrices for each disease termed $\mathcal{S}_j$ by counting the total number of occurrences of each possible pair of neighboring TH-states across all images of that disease and normalizing by the total number of neighbors of each TH-state.

5.5 RESULTS

Microstates capture histologic patterns

FS reviewed a subset of the largest microstates found from several images to validate the homogeneity of the clusters and ensure grouping of diagnostic features into clusters. To further ensure homogeneity, a second and third round of random walk was performed, and the microstates were again verified by FS 5.5). All microstates were considered homogeneous from the first and second rounds of random walk, while microstates in the third round were determined to be more heterogenous with respect to nuclei architecture. Microstates from the first round of clustering are used to form TH-states.

States are associated with disorders

States at each coarseness level describe increasingly coarse histologic patterns associated with ILD. In Figures 5.5 and 5.5, we show representative blocks from each TH-state along with the TH-state compositions for each disorder. For visibility, we show only show examples from the coarser hierarchy levels. At coarseness level 7 (5.5), images with UIP, NSIP, and fibrosis are dominated by clusters 1-3 and 5, which contain patterns seen in fibrotic tissue. These disorders contain very little tissue of TH-state 4, 6 or 7, which are healthier, and TH-state 12, which is common is environmentally-related diseases. In contrast, the control images have very little tissue from clusters 1-3 and 5, but are dominated by clusters 4,6, and 7, which contain healthy tissue. Both the control and emphysema tissues have high amounts of cluster 12, which is expected as many of the control patients have a diagnosis of carcinoma, and both lung carcinoma

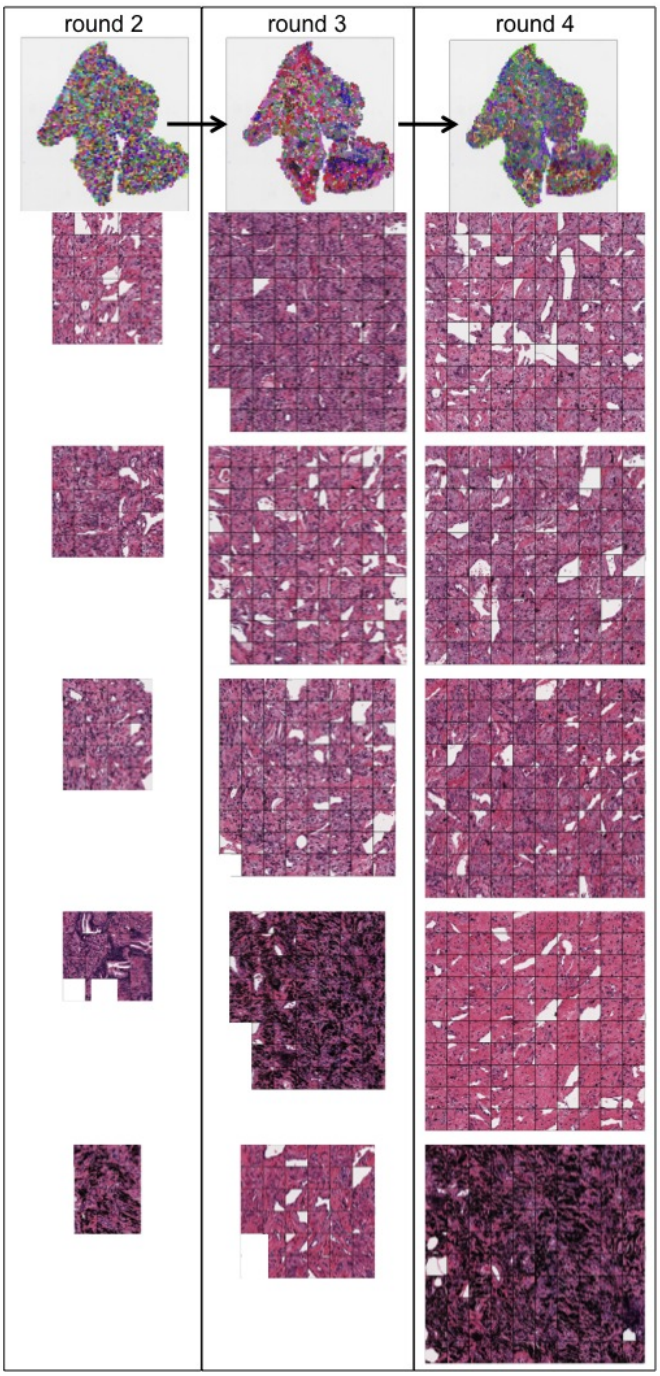

Figure 28: Microstates from a fibrotic lung tissue after one (column one), two (column two), or three (column three) rounds of clustering. The top row shows the full whole slide tissue image, where colored boxes indicate blocks assigned to the same microstate. Microstates from the first and second rounds of clustering were deemed homogeneous by a pathologist. Rows 2-4 show all blocks assigned to the three largest microstates in each round of clustering. Rows 5 and 6 show two other microstates from each clustering round. Note that the clusters become larger and more heterogenous with each round of clustering.

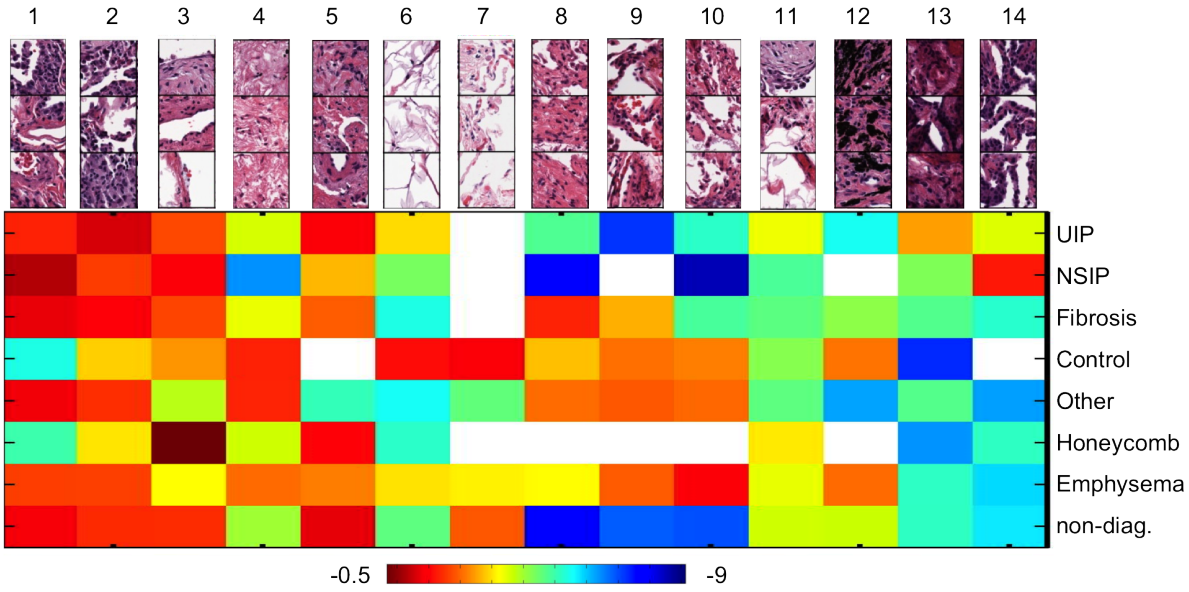

Figure 29: State compositions for images from each disorder. The heat map shows the log-percentage of blocks from images with each disorder (rows) that are assigned to each of the 14 TH-states (columns) at coarseness level 7. For each TH-state, three representative blocks from that TH-state are shown above the corresponding column. Red indicates higher percentages, blue indicates lower percentages.

and emphysema are often related to environmental factors. In coarseness level 6 (5.5), clusters 5, 7, 13, 15, and 17 contain healthy lung tissue are compose a large amount of the control class's tissue. Cluster 20 contains fibrotic tissue and is most common in UIP.

Architectural signatures can better differentiate disorder types than TH-state composition

By taking the χ^2-distance between the TH-state composition vectors for each pair of disorders at any given coarseness level, we find pairs of disorders that are most similar with respect to their TH-state composition. In Figure 5.5 (left column), we show the distances between each image pair, averaged according to image disorders for coarseness level 9, which provides the coarsest representation of the images with only 3 clusters. We see that UIP and fibrosis have the most similar TH-state compositions, while the honeycomb and non-diagnostic images are the most different. NSIP's TH-state composition is more similar to several other disorders than it is to itself, indicating that the TH-state composition among images diagnostic of NSIP can vary greatly. A diagnostic characteristic of NSIP is the presence of large homogeneous regions of fibrosis interspersed with homogeneous healthy regions, thus explaining similarities with both fibrotic disorders and healthy disorders. Control and emphysema also have similar distributions to fibrotic disorders using this metric and this coarseness level. If we look at the Frobenius norm between spatial architecture matrices for each image pair, averaged over disorders, the same associations between diseases can be extracted from the distance matrix (Figure 5.5, middle). Additionally, by comparing spatial architecture matrices, we see that control and emphysema have are not as similar to the fibrotic diseases as those diseases are to each other. However, the spatial architecture matrices alone do not distinguish between NSIP and UIP with simply the frobenius norm. In Figure 5.5 (right panel), we compare a vectorized form of the spatial architecture matrices, weighted by the state composition vectors, using the χ_2 distribution. This hybrid metric establishes the differences between control and emphysema tissue from fibrotic tissue, and also distinguishes NSIP tissue from fibrotic tissue.

Architectural signature matrices capture diagnostic features of lung disease

In Figure 5.6, we show the TH-state labels and architectural signature matrices for images that are clearly diagnostic of UIP (A,C) and images that are clearly diagnostic of NSIP (B,D) at

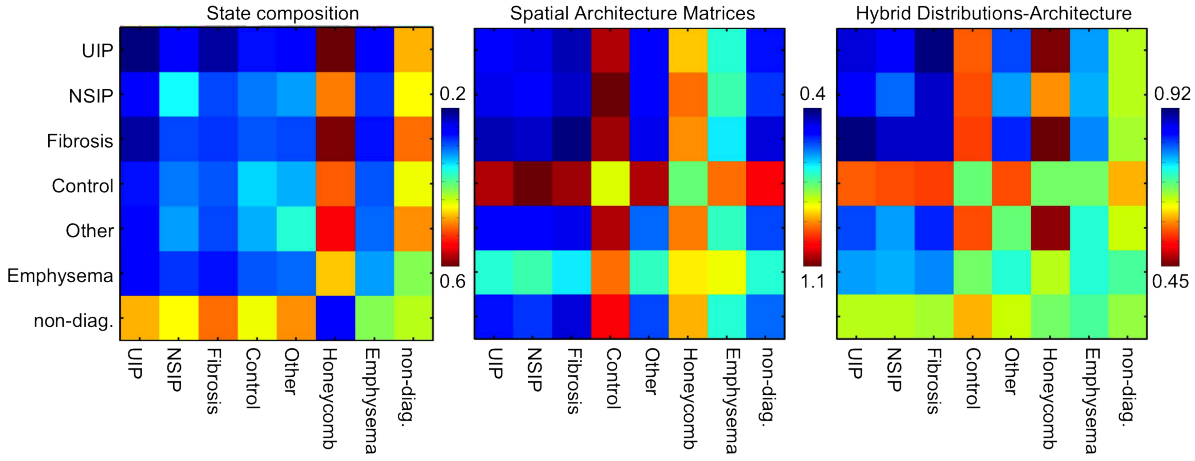

Figure 30: Distances between image pairs, averaged according to diagnosis, at the coarsest coarseness level (9 rounds of clustering). The left panel shows the mean χ^2 distance between TH-state composition vectors for each disorder pair. The middle panel shows the mean Frobenius norm between architectural signature matrices for each pair of disorders. The final column shows the mean χ^2 distance between the architectural signature matrices, weighted and vectorized using the TH-state composition vectors, for each disorder pair. Blue indicates lower distances (more similar) while red indicated higher distances (less similar). Color bars are shown for each heat map, but as each heat map uses a different distance member, only relative comparisons between matrices are intended.

coarseness level 6. In general, the architecture signature matrices of UIP images are densely connected, agreeing with the heterogeneous nature of that disorder. In contrast, the NSIP images mostly have fewer connections, which agrees with the diagnostic description of NSIP as being "homogeneous compared to UIP". This heterogeneity/homogeneity is somewhat apparent in the TH-state images (A,B), however the matrix form quantifies and simplifies this feature.

5.6 DISCUSSION

Potential for Classifier

While the data set is not large enough to develop a classifier, we demonstrate the potential of spatial architecture signatures for classification of ILD. For this task, we use a set of 14 images selected by the pathologist as being clearly diagnostic of UIP (8 images) or NSIP (6 images). In Figure 5.6, we show the spatial architecture matrices at coarseness level six for these images. Let μ_{U}^{l} be the mean UIP spatial architecture matrix at coarseness level l and μ_{N}^{l} be the mean NSIP spatial architecture matrix at coarseness level l. For any image I_j with spatial architecture matrix S_j, we form a simple classifier by computing the distance between S_j and both μ_{U}^{l} and μ_{N}^{l}, and assigning the image to whichever disorder's mean is closer. In Table 6, we show the true positive rate of classifying each of the 14 images as either UIP or NSIP using this basic method. Additionally, we demonstrate the classification ability on the remaining UIP images, which were labeled as somewhat diagnostic of UIP, in the fourth column (TPR somewhat UIP). 4. The overall percentage of data assigned to UIP at each coarseness level is shown in column 4, instead of a false positive rate. Importantly, in levels 1-7, although the majority of images were assigned to UIP, most of the NSIP images were correctly labeled as NSIP. This may indicate that the NSIP architecture signature is distinguishable from the UIP architecture signature using these matrices. Moreover, at each coarseness level, a greater percentage of clearly UIP images were labeled as UIP than the overall percentage of images labeled as UIP, indicating that the UIP pattern is also recognized. We also removed the fifth NSIP image from the set of clearly diagnostic images, as it appears to be an outlier in this set (5.6, before computing the mean NSIP architecture

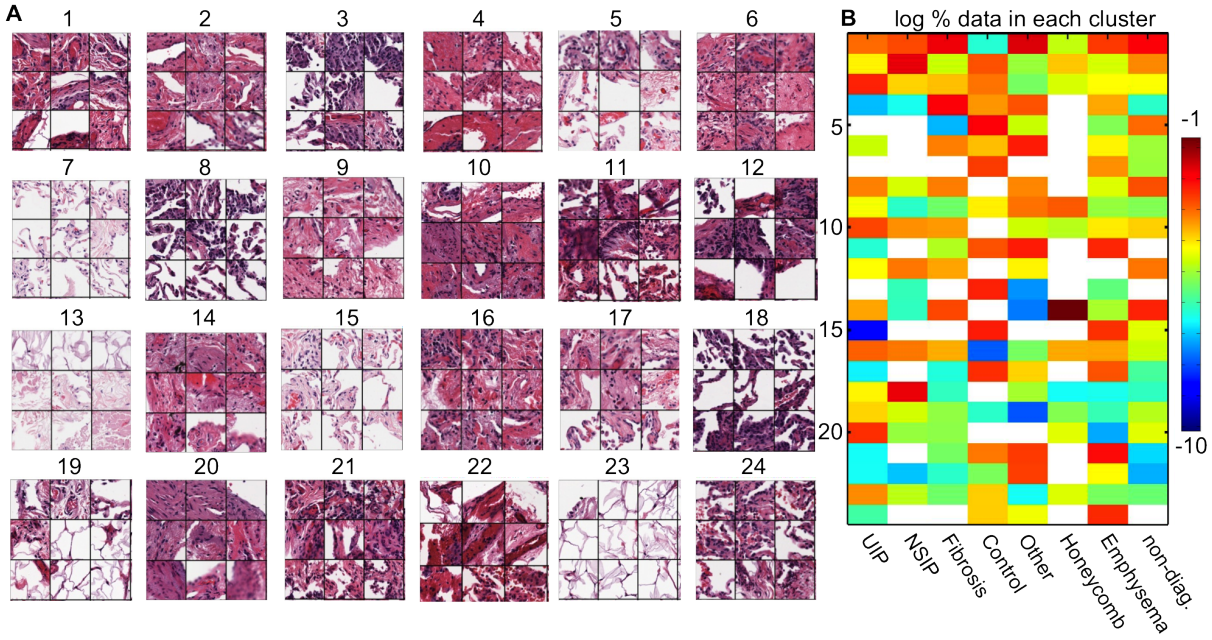

Figure 31: TH-state memberships for images from each disorder at coarseness level 6. For each of the 24 largest TH-states, nine representative blocks from that TH-state are shown in panel A. The heat map (B) shows the log-percentage of blocks from images with each disorder (rows) that are assigned to each of the 25 largest TH-states (columns) at this coarseness level. Red indicates higher percentages, blue indicates lower percentages.

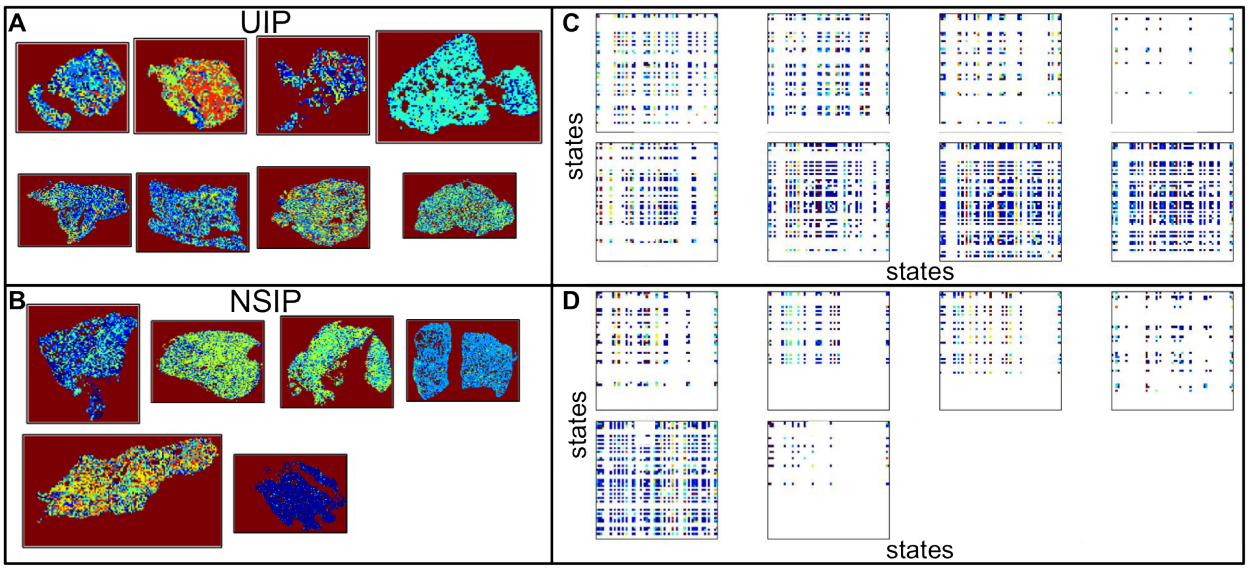

Figure 32: State assignments and spatial architecture matrices for the set of clearly diagnostic UIP and NSIP images at coarseness level 6. Panels A, B: State assignments painted on the whole slide images for UIP (A) and NSIP (B). Color indicates TH-state index. Background is colored maroon, as the empty space is considered a TH-state in the spatial architecture matrices. Panels C,D: Spatial architecture matrices for the corresponding images in A,B. Red indicates greatest number of neighboring blocks, blue indicates least. Each matrix has one row and column for each TH-state, plus additional columns for airways and background.

level	TPR clearly UIP	TPR clearly NSIP	TPR somewhat UIP	% data assigned to UIP
3	100	100	100	90
4	100	100	100	90
5	100	83	83	89
6	100	83	67	83
7	88	83	50	68
8	63	50	67	52
9	75	17	50	41

Table 6: Potential as classifier: True positive rate for assigment of clearly UIP images to UIP (col. 2), clearly NSIP images to NSIP (col. 3), and somewhat UIP images to UIP (col. 4) compared to the overall percent of data assigned to UIP. Rows indicate coarseness level of the spatial architecture matrix.

matrix at each level. After removing this image, NSIP was even more distinguishable from UIP using this metric. However, more data is necessary to determine whether the fifth image truly is an outlier, and to develop a rigorous classifier for labeling images as NSIP or UIP. Such a classifier would be beneficial in the clinical setting to provide unbiased analysis of images that would assist pathologists in distinguishing these two classes, which have a high inter-pathologist disagreement rate. Additionally, the classifier could be used as a prescreening method to filter out images that are clearly diagnostic of a given disorder, so that pathologist time could be devoted to less obvious cases.

Future Work

In addition to interior and exterior air space, more context information could be added to the architectural signature matrix, e.g. interstititium, bronchioles, pleura, etc. Adding more architectural components of lungs would closer replicate tissue analysis by pathologists. These structures may be implicitly described by the clusters, but the possibility that classification would be improved through explicit labeling of these structure should be explored.

Explore use of architectural signatures as classifier. This would require a larger and more balanced data set, but methods for classification of networks could be explored on the current set. Additionally, other metrics for defining neighboring networks should be explored, as well as higher degree neighbor relationships.

For the initial formation of homogeneous microstates, the block size could be further explored, and block boundaries could be adjusted so that microstates were entirely homogeneous and not limited to a square shape. Pathologists use the presence of specific cells in tissue, such as lymphocytes and endothelial cells, as well as their abundance and arrangement while making diagnoses. This information could be incorporated in the feature set through cell-type specific nuclei detection.

6.0 ANALOGOUS METHODS APPLIED TO MOLECULAR DYNAMICS SIMULATIONS

This Chapter was published as
Quasi-Anharmonic Analysis Reveals Intermediate States in the Nuclear Co-Activator Receptor Binding Domain Ensemble Virginia M. Burger, Arvind Ramanathan, Andrej J. Savol, Christopher B. Stanley, Pratul K. Agarwal, and Chakra S. Chennubhotla; Pacific Symposium on Biocomputing 17:70-81(2012)

6.1 INTRODUCTION

Intrinsically disordered proteins (IDPs) play a vital role in regulating cellular processes in eukaryotic cells[91, 92]. Structural studies have revealed that unlike well-folded globular proteins, IDPs exist as highly dynamic ensembles even under equilibrium conditions, with diverse and constantly fluctuating secondary/tertiary structure[93]. The ability of IDPs to adapt their binding surface to recognize various binding partners provides a novel means of regulating various cellular activities[94]. Given the abundance of IDPs in the human genome and their involvement in neurodegenerative, cardiovascular, and amyloid-related diseases[95, 96], there is tremendous interest in understanding the basic molecular mechanisms by which IDPs recognize their binding partners and facilitate their specific functions. For example, some IDPs possess the remarkable ability to undergo synergistic folding upon recognizing their binding partners[97]. The contrasting ability of IDPs to achieve a high degree of structural plasticity while retaining binding specificity presents a serious challenge in characterizing their sequence-structure-function rela-

tionships.

The intrinsically disordered nuclear co-activator binding domain (NCBD) of the CREB binding protein (CBP) interacts with numerous transcription co-activator proteins (TCA), including the steroid receptor co-activators (SRC)[98], p53[99], p73[100], interferon regulatory factors (IRF)[101] and the viral protein Tax[102]. As NCBD aids recruitment of the transcriptional machinery, its dysfunction (and that of its binding partners) is implicated in several forms of leukemia[103] and lung cancer[104]. Circular dichroism (CD) and ultra-violet (UV) spectroscopic studies reveal that native NCBD adopts a compact structure with a high degree of helicity but lacks the sigmoid unfolding curve characteristic of folded proteins[105]. Structural studies using nuclear magnetic resonance (NMR) and X-ray crystallography indicate that NCBD adopts unique conformations when complexed with specific partners[105, 106] and that synergistic folding facilitates the interdigitation of three helices, a feature common in NCBD's bound topology (identified by $\alpha_1 - \alpha_3$; see Fig. 33)[107, 108]. Increasingly, the specific orientations of these three α-helices are thought to confer the specificity inherent to NCBD:TCA intermolecular recognition.[105, 106, 107, 108]

While a number of studies point to the behavior and structure of NCBD in its bound state[105, 106, 107, 108], the conformational heterogeneity of *apo*-form NCBD has been challenging to characterize. Emerging evidence from NMR experiments[106] suggest that native NCBD can adopt conformations that largely resemble the SRC/ACTR-bound conformation. However, that study also revealed that ligand-free NCBD does not sample states that resemble the IRF-bound conformations. Moreover, Fraenkel et al.[109] have determined the *apo*-form of NCBD to be quite different from Poulsen et al[106]. Based on the current insights gained from experimental studies, the biophysical mechanisms underlying NCBD:TCA recognition process remain unclear. Likewise, a quantitative description of *disorder-to-order* transitions between the ligand-free or ligand-bound NCBD ensembles is lacking.

In this paper, we address the aforementioned issues and outline an integrated experimental and computational strategy to analyze disorder-to-order transitions in NCBD's conformational landscape. Our aims are to: (a) obtain insights into the nature of intrinsic fluctuations accessible to ligand-free NCBD, (b) identify regions within NCBD that are implicated in its disorder-to-

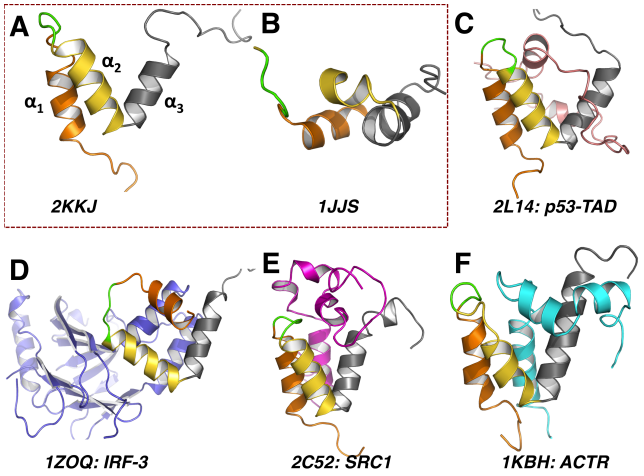

Figure 33: **Bound and unbound forms of NCBD.** NMR ensembles of the ligand-free structures: 2KKJ (A) and 1JJS (B); NCBD in complex with (C) p53 trans-activation domain (TAD) (2L14: TAD in pink); (D) interferon regulatory factor 3 (IRF3) (1ZOQ: IRF3 in pale blue); (E) steroid receptor coactivator 1 (SRC1) (2C52: SRC1 in magenta); (F) interaction domain of activator for thyroid hormone and retinoid receptors (ACTR) (1KBH: ACTR in cyan). In all panels, the three helix bundle of NCBD is highlighted in orange (α_1), yellow (α_2) and gray (α_3), while the specificity loop (PSSP) is in green.

order transitions and (c) elucidate whether ligand-free NCBD can access conformations that resemble the ligand-bound conformations. To this end, we exploit recent advances in molecular simulation technologies to extensively sample ligand-free NCBD. Using graphics processing units (GPUs), we accelerate conventional all-atom explicit solvent molecular dynamics simulations to microsecond time-scales. The aggregate dataset constitutes 40μs of MD simulation and required approximately two months of total clock-time.

Long time-scale simulations challenge conventional trajectory analysis methods. In particular, biophysically relevant events within such trajectories are often difficult to detect[110, 111]. Likewise, experimental techniques also present modeling challenges; results from small-angle neutron scattering (SANS) experiments on NCBD suggest a distinctly long-tail (or anharmonic) behavior in the distributions of radius of gyration and end-to-end distance values[112]. This long-tailed behavior implies that atomic fluctuations in NCBD involve significant higher-order correlations, which are commonly overlooked with typical trajectory analysis tools[113]. Recently, we introduced quasi-anharmonic analysis (QAA) as an effective computational model to quantify these higher-order correlations which emerge prominently within long simulations[114]. QAA provides insights into the inherent *anharmonicity* in atomic fluctuations and is thus ideal for quantifying the disorder-to-order transitions in NCBD observed from both experiments and simulations. Furthermore, QAA organizes the conformational heterogeneity in NCBD fluctuations into a small set of conformational sub-states that share structural and energetic homogeneity.

Markov state models (MSMs) and their variants also provide organizational principles for molecular simulations. These methods exploit the kinetic connectivities[115] or structural similarities[116] between conformational sub-states and have been useful for determining transition pathways between conformational sub-states[110, 117]. As a comparison to QAA, MSMs discretize conformation space into a network or graph of sub-states rather than projecting it into a low-dimensional, continuous representation. A central contribution of the work here is an approach which exploits both the dimensionality reduction (and visual interpretability) of QAA and rigorous graph theoretic methods to determine a hierarchy of transitions between sub-states. With this integrated approach, we determined that ligand-free NCBD can indeed access conformations representative of the ligand-bound form. Within our simulations, NCBD's α_1 and α_2

helices in the ligand-free and ligand-bound conformations are largely similar; α_3 however, can exhibit a wide degree of flexibility and does not generally sample conformations that are similar to the ligand-bound state.

6.2 APPROACH

We performed ten $4\mu s$ all-atom explicit solvent MD simulations of apo-NCBD (Section 6.3). To identify biophysically relevant motions within these simulations, we developed a novel, broadly extensible, dimensionality reduction framework based on quasi-anharmonic analysis in the dihedral angle space, called *dihedral QAA* or *dQAA* (Section 6.4). To validate our simulations we used two order parameters: radius of gyration (R_g) and helicity (H; defined here as the percentage of NCBD that adopts α-helical structure as assigned by STRIDE[118]), which can be measured experimentally via SANS[112] and circular dichroism[106] experiments respectively.

To determine meta-stable conformational states, we invoke a multi-scale Markov diffusion approach (Section 6.5) to group similar conformations in the dQAA space. Iterative diffusion-based clustering in the dQAA space results in a hierarchical description of the NCBD conformational landscape. Each level of the hierarchy provides a set of increasingly broad (or inclusive) meta-stable states, allowing the conformational landscape of NCBD to be viewed as a collection of nested sub-states. As we demonstrate, dQAA coordinates provide a natural framework for organizing the conformational heterogeneity of the apo-NCBD ensemble and help identify disordered or compact conformational states. In addition, the Markov diffusion approach captures meta-stable states that provide insight into the nature of structural changes that NCBD must undergo in order to sample conformations close to the ligand-bound state (Section 6.6).

6.3 MOLECULAR SIMULATIONS FOR NCBD

A total of six NMR and X-ray NCBD structures are available in ligand-free and ligand-bound form. Fig. 33 shows the variation in the orientation of the three α-helices between these structures. While NCBD adopts very similar helical orientations when binding ACTR, SRC1 and p53, the interfaces and helical turns of NCBD when complexed to each ligand are quite different. Furthermore, NCBD adopts a radically different orientation for interacting with IRF3; α_3 twists and rests on a very different axis from that in the ACTR interaction.

In the interest of sampling the large conformational space of ligand-free NCBD, we initiated a 4μs long simulation for each of the 10 conformations in the NMR ensemble (2KKJ) that is representative of the ligand-free state. We used the AMBER suite of tools[119] and the ff99SB[120] force-field to model the proteins. Each of the ten conformations was immersed in a cubic box of SPC water molecules such that the solvent box boundary was never less than 10Å from the protein. Counter-ions consisting of 10 Cl^- were added to ensure system neutrality. The box sizes were approximately 90 $\times$ 90 $\times$ 90 $Å^3$ (with slight variations for each of the ten simulations). Using the protocol highlighted in our previous work [121], each of the simulation systems was subjected to energy minimization and equilibration. A final MD equilibration of 1.0ns duration was run to ensure the systems reached a stable conformation. All the simulations were carried out at 300K using the NVE ensemble. Each of the ten systems had between 9,000 and 12,000 water molecules, resulting in system sizes varying between 18,000 and 22,000 atoms.

Production runs were carried out using the recently developed ACEMD (accelerated MD) code specifically for graphics processing unit (GPU) systems[122]. In order to accelerate the MD simulations to reach microsecond time-scales, the systems were simulated using a time-step of 4fs using a hydrogen mass-partitioning scheme[123]. The alteration to the dynamics due to the mass-partitioning scheme is minimal since individual atom masses do not appear explicitly in the equilibrium distribution[122]. Ten production runs sampling 4μs per simulation were performed. Coordinates were saved every 200 ps, resulting in about 20,000 conformations per simulation or an aggregate total of 200,000 conformations for all simulations (40μs total).

Comparison with NMR: To compare our production runs with NMR data, we used

SPARTA[124] to predict the 1H, 13C, and 15N chemical shifts for the ensembles generated from MD simulations. SPARTA uses backbone ϕ and ψ torsion angles, side-chain χ_1 angles, and sequence information to predict backbone chemical shifts of protein structures [124]. We found that the simulations show reasonable agreement with the chemical shifts from the experimental ensembles (2L14, 1KBH and 2KKJ). In particular, the correlation coefficients between the mean MD and the experimental 15N shifts are 0.74, 0.78, and 0.88, respectively, for the 2L14, 1KBH and 2KKJ data. We note that computed 1H and 13C chemical shifts are less consistent with respective experiments presumably due to force-field inaccuracies and the 4 fs MD integration time-step[125]. While the agreement between experiments and computations is a cursory check on the quality of data obtained, we must also note that the chemical shifts from the experimental ensembles may not be fully representative of the conformational heterogeneity of apo-NCBD.

Comparison with SANS: We next compare simulation results with experimentally derived R_g values from small-angle neutron scattering (SAS) experiments. The distribution of R_g values from MD simulations is observed to be more constrained than that obtained from SANS, possibly due to MD sampling deficits(Fig. **??** panel B, blue: aggregate simulations; red dash: single simulation; red: SANS data). This is in part because MD trajectories are strongly biased by the chosen starting pose, which is commonly an energy-minimized X-ray or NMR ensemble structure [126, 127]. We note that the range of SANS-derived R_g values suggests that NCBD may undergo disorder-to-order motions on a larger scale than observed in the present simulations.

From a molten globule state to a near ACTR-bound form: To quickly overview significant conformational events in the MD trajectory, we track R_g on-line along a subset of one of the simulation trajectories using two different exponential window smoothing timescales (Fig. 6.3). We observe that NCBD changes from a molten-globule form (high R_g) to a near ACTR-bound form (gray cartoon for comparison, shown along with RMSDs). The pathway chosen by this trajectory is highly dynamic, involving several significant rearrangements of the α_1-α_2 (PSSP) loop and α_3. Interestingly, the conformational changes persist across the timescales of the exponential window, confirming the evolution of NCBD from a molten globule state to a near ACTR-bound form. In this particular trajectory, generated from model 2 of the NMR ensemble (2KKJ), NCBD adopts a form that is about 4.27 Å (C^{α}-RMSD) from the bound form; however, other trajectories

adopt conformations that are much closer to the ACTR-bound form (see Section 6.5).

6.4 DQAA: QUASI-ANHARMONIC ANALYSIS IN THE DIHEDRAL ANGLE SPACE

The conformational heterogeneity we observed in long timescale simulations of NCBD motivated us to eliminate the sensitivity to Cartesian alignment by analyzing the NCBD ensemble in the dihedral angle space. For a N residue protein there are a total of $2N$ backbone ϕ and ψ angles, $\phi = \{\phi_i\}_{1,\ldots,N}$, $\psi = \{\psi_i\}_{1,\ldots,N}$. Each backbone dihedral angle pair (ϕ_i, ψ_i) can be converted into a Euclidean representation by $x_{i-3} = \cos(\phi_i);\ x_{i-2} = \sin(\phi_i);\ x_{i-1} = \cos(\psi_i);\ x_i = \sin(\psi_i)$, yielding a $4N$ vector x. We first considered dihedral PCA (dPCA), where a covariance matrix is generated from this data and is diagonalized to obtain a low-dimensional representation of the conformational ensemble[128, 129, 130]. We observed that NCBD conformers projected into low-dimensional dPCA space lacked coherency (or homogeneity) with respect to the R_g values, indicating that dPCA is unable to fully describe the disorder-to-order motions of NCBD (data not shown).

Protein motions are anharmonic; therefore, capturing the conformational diversity of protein fluctuations requires effective models that quantify anharmonic motional signatures[113, 131, 132, 133, 134, 135]. Anharmonicity is best summarized by higher-order statistics[131, 132]. Our previously developed framework, quasi-anharmonic analysis (QAA), exploits these higher-order statistical signatures of protein motions [114]. When applied to μs time-scale simulation data of proteins involved in molecular recognition and enzyme catalysis, QAA revealed (i) functionally relevant, hierarchically-organized conformational sub-states and (ii) a set of on-pathway intermediates between these sub-states. This result is consistent with the understanding that proteins sample from a hierarchical, multilevel energy landscape with minima and maxima separated by energy barriers [136, 137]. We observed that the sub-states determined with QAA were energetically coherent, indicating that our low-dimensional representation appropriately depicts energetically-related conformers as neighbors. We emphasize, however, that the resultant energy coherence within observed sub-states is an emergent property of QAA, indicating that our

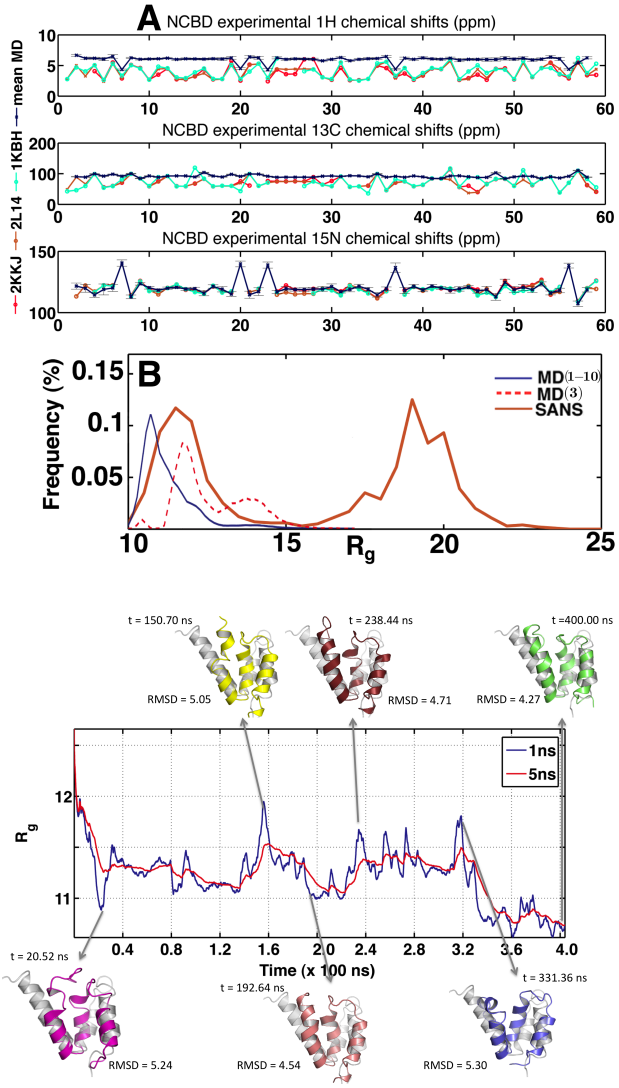

Figure 34: **Disorder-to-order transitions in NCBD ligand-free ensemble** (a) A comparison of simulated NCBD ensembles with NMR (A) and SAS (B) experimental data, illustrating qualitative agreement. Chemical shift data is taken from three ensembles, 2KKJ (16363cat.bmrb, red), 2L14 (17071cat.bmrb, brown), 1KBH (5228cat.bmrb, cyan), and compared to computed mean chemical shifts from the simulations. (B) R_g is shown for SANS data (tan, solid), aggregated MD data (blue, normalized), and a single MD trajectory (2KKJ, model 3)(dashed red, normalized). Not all of the conformational landscape is sampled by MD, as is evident from the second SANS peak. (b) R_g during first 400ns of a single MD trajectory (2KKJ, model 2), with 1ns (blue) and 5ns (red) exponential smoothing showing disorder-to-order transitions. Conformations at six timepoints are aligned to crystal structure 1KBH.

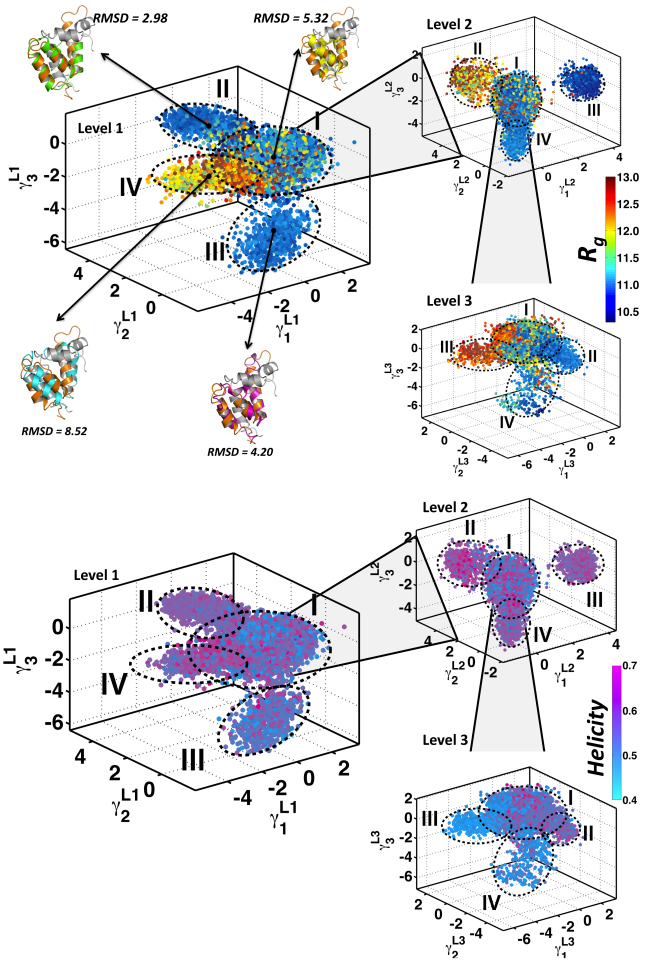

Figure 35: **dQAA identifies a hierarchy of disorder-order promoting motions and homogeneous clusters in 2KKJ μs timescale ensemble**. MD trajectory frames are projected along the top three dQAA modes and colored by (a) R_g and (b) Helicity. (a) Level 1 of the dQAA hierarchy reveals two compact, low R_g clusters (II and III). Cluster IV has high R_g values (red) indicating a more open conformation. Mean conformers in each cluster (I: yellow, II: green, III: maroon, IV: blue) are superimposed on the bound conformer of NCBD-ACTR (orange) and the respective RMSDs are given. Successive application of the dQAA analysis to heterogenous clusters (Level 2 and 3) highlight a rich conformational diversity when painted with R_g values values. (b) In level 1, dQAA clusters I and III are predominantly low in helicity (blue) and dQAA clusters II and IV are predominantly high in helicity (pink). The ability to separate ordered (high helicity) from disordered (low helicity) conformers improves as dQAA is applied recursively to subsets of conformers.

higher-order statistical approach selects meaningful reaction coordinates.

With the intention of capturing anharmonic disorder-to-order motions, we pursued anharmonicity as an informative statistic in the form of dihedral QAA (dQAA), basing our technique on the diagonalization of a tensor of fourth-order statistics in the dihedral angle space. This tensor describes dihedral angle fluctuations and their couplings and can be efficiently diagonalized with a technique called joint-diagonalization of cumulant matrices (JADE), a well known machine learning algorithm for analyzing multi-variate data [138]. To begin with, second-order correlations are removed from the dihedral angle fluctuation data. Next, a fourth order cumulant tensor $\mathcal{K}$ is computed consisting of both auto- and cross-cumulants. The cumulant tensor will have a total $4N \times (4N+1)/2$ matrices each of size $4N \times 4N$ accounting for auto- and cross-cumulant terms. Finally, the fourth order dependencies denoted by the sum of the cross-cumulant terms are minimized, a procedure equivalent to diagonalizing $\mathcal{K}$. No closed form solution exists for diagonalizing a tensor, however an approximate solution can be found using efficient algebraic techniques such as Jacobi rotations [139]. Just as an eigenbasis diagonalizes a covariance matrix, a matrix U is found to approximately diagonalize the cumulant tensor. The basis matrix U represents anharmonic modes of motion derived by minimizing the fourth-order dependencies in dihedral angle fluctuations, in addition to eliminating the second-order correlations as is the case with dPCA. Unlike in dPCA, the column vectors of U (sorted decreasingly by amplitude ($\|U_i\|$)) can be non-orthogonal and hence intrinsically coupled.

Results: Using 40 μs simulations of NCBD, we performed dQAA to reduce 232-dimensional input data (from 58 dihedral angles in each conformer) to a 50-dimensional subspace. For visualization, we projected the conformers along the top three QAA modes as shown in Fig. 35. To assess if the projected conformers share any structural similarities, we colored the conformations using two biophysically relevant order parameters: (a) R_g and (b) H (helicity). The dQAA space colored with R_g revealed two compact (homogeneous) clusters with low R_g values, one open conformation cluster with high R_g and one heterogeneous cluster. Thus, dQAA modes can reveal disorder-to-order motions, an ability that can be further tested by recursively applying dQAA on the heterogeneous cluster. The results from a recursive decomposition highlight the rich conformational diversity present in the simulated NCBD ensemble and illustrate the ability

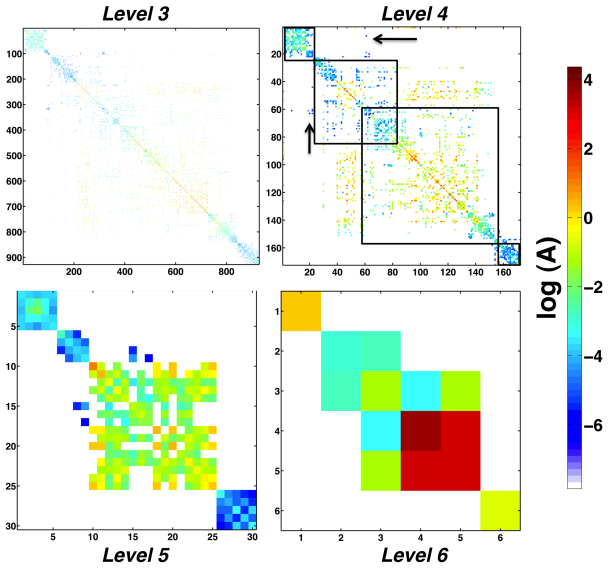

Figure 36: **A hierarchy of conformational sub-states in the disorder-to-order transitions of NCBD conformational landscape.** A total of 6 levels are found by the hierarchical clustering. For hierarchy levels 3-6, the log of the affinity between each sub-state pair is shown.

of dQAA to capture meaningful conformational transitions. Although dQAA cannot directly compensate for the deficiencies of MD sampling, the determined anharmonic modes suggest functionally relevant disorder-to-order transitions. Similar results can be seen by coloring the dQAA space with helicity values, showing that the sub-states involve transitions in NCBD from a more extended form to a more helically compact form. This emergent homogeneity in dQAA space suggests a new strategy to identify metastable states in the MD trajectory, which we discuss next.

6.5 HIERARCHICAL CLUSTERING IN THE DQAA-SPACE TO IDENTIFY META-STABLE STATES

Observing that neighboring conformers in dQAA-space have similar R_g and H values, and noting that this coherence is an emergent property of dQAA representation, we hypothesize that

nearest neighbors in dQAA-space are dynamically and kinetically related. We use the conformational coordinates returned by dQAA to build long-lived metastable states using graph-theoretic spectral clustering approaches. To this end, we consider each frame in the trajectory as a node in an undirected graph and connect each node to 10 of its nearest Euclidean neighbors in the three-dimensional dQAA space. The edges are assigned weights inversely proportional to the difference in their radius of gyration values, thus merging both the dynamic and emergent properties of the dQAA space into the edge weights. We then cluster this network using a hierarchical Markov diffusion framework[74]. This approach is an adaptation of our earlier work developing spectral graph partitioning algorithms for segmenting natural images[74], understanding protein dynamics and allosteric propagation[76], relating signal propagation on a protein structure to its equilibrium dynamics [140], and finally discovering metastable states in MD trajectories[78].

We begin hierarchical clustering by constructing a Markov transition matrix using an affinity matrix of edge weights between conformer pairs in the dQAA space. We then initiate a Markov chain (or random walk) on the weighted undirected network. As Markov transition probabilities homogenize through diffusion, an implicit clustering emerges from the network. First, a set of nodes representing the putative clusters are identified. The number of clusters chosen is determined by the algorithm so that every node in the network has some Markov probability of transitioning into at least one of the clusters. Then, a Markov transition matrix is newly constructed using this reduced representation. The important principle behind this construction is that upon reaching a stationary distribution at the coarsest hierarchy level, the Markov chain has also converged at finer (more local) network levels. This consistency regulates the overall topology of the network and helps build a multi-resolution representation of metastable states.

We expect that fine-grained hierarchy levels will produce many small clusters containing close neighbors in the QAA space; that is, within each such cluster most members will be drawn from the same, narrow time-window. As Markov diffusion progresses (fine-grained to coarse-grained), conformers that are more distant neighbors will be connected by edges in the diffused network and will therefore be assigned to the same cluster. Thus, the hierarchical clustering can highlight dynamical connections between conformers at different timescales.

Results: The affinity matrix hierarchy derived by the clustering algorithm is shown in

Fig. 36. The affinity matrices show several regions of high cross-talk at lower levels of the hierarchy. Iterative diffusion of the Markov chain derived from the initial affinity matrix (200000 $\times$ 200000), results in six hierarchy levels (Table 1). The mean C^{α}-RMSD to cluster center at the bottom hierarchy level is 3.2Å, indicating that clustering in dQAA-space also captures structural similarity between trajectory frames in Cartesian-space. Clusters with low mean RMSDs to the four experimental bound conformations and the two experimental unbound conformations occur at each hierarchy level. At the finest level of the hierarchy, the clusters representing the bound conformations are very small, but as the hierarchy progresses, they are found in more dominant sub-states, indicating that the bound conformations are energetically accessible. As seen in Table 1, the alignment to 1ZOQ is poor. However, if only helices α_2 and α_3 are considered, the RMSD is very low (data not shown). In contrast, for the three other ligand bound states, α_1 and α_2 align well to the simulations. Thus, a barrier involving the repositioning of this helix may need to be crossed in order to access the IRF-3 bound state.

PDB	ligand-free 1JJS	ACTR 1KBH	IRF3 1ZOQ	SRC1 2C52	ligand-free 2KKJ	p53 2L14	
Level	rank/ RMSD(Å)	rank/ RMSD(Å)	rank/ RMSD(Å)	rank/ RMSD(Å)	rank/ RMSD(Å)	rank/ RMSD(Å)	Total number of clusters
3	895/5.3	928/1.8	313/7.3	928/1.9	928/1.4	910/5.2	928
4	49/6	110/1.9	122/7.3	168/2.0	81/1.5	132/5.2	172
5	10/6.3	30/1.9	25/7.4	30/2.1	30/1.5	30/5.3	30
6	1/6.4	3/2.0	5/7.4	3/2.2	3/1.6	3/5.3	6

Table 7: Conformational similarity between determined sub-states and extant structural models. Sub-states are ranked according to membership, 1 being the largest. For the coarsest hierarchy levels, sub-state rank and RMSD from sub-state center to experimental conformation is given for the sub-state with lowest RMSD to the experimental conformation.

6.6 INTERMEDIATE STATES OF LIGAND-FREE NCBD ACCESS LIGAND-BOUND CONFORMATIONS

The organization of the ligand-free NCBD ensemble indicates the presence of six large conformational sub-states that interconvert between each other. One can visualize the six sub-states from the coarsest hierarchy level as illustrated in Fig. 37(a). Of the six sub-states, sub-states 4 and 5 constitute over 88% of the entire ligand-free ensemble, consisting of 98,143 and 79,672 conformers respectively. The remaining sub-states (1, 2, 3 and 6) represent rare transitions in the landscape. It is interesting to observe that sub-states 1 and 6 are somewhat isolated from the conformational states, however a sizable population of conformations exist in each state (see affinity map in 37(a)). Although one may attribute the isolation to the MD sampling protocol, it is important to note that descending through the various levels of the hierarchy (Level 5 through Level 2) indicates that both sub-states 1 and 6 are connected via extremely lowly populated states (see Fig. 36), indicating that multiple paths exist through which states 1 and 6 can be reached. We also note that while certain pairs of sub-states (such as [2,3] and [4,5]) freely interconvert between each other, sub-state 3 alone can access conformations that are similar to that of sub-state 5. Therefore, sub-state 3 acts as an intermediate state from which conformations in sub-states 2, 4 and 5 interconvert.

Sub-state 1 (rank 3) represents the state closest to the bound conformations observed experimentally (Table 1). As illustrated in Fig. 37(b), a representative structure from sub-state 1 is compared with two ligand-bound structures, namely 1KBH (panel A) and 2C52 (panel B). Sub-state 1 represents the third least populated state of all sub-states (9,488 or 4.7% of conformers). However, when compared with the bound structures, on an average, it exhibits smaller RMSD values to the bound 1KBH (RMSD: 2.0 Å) and 2C52 (RMSD: 2.2 Å) conformers. This observation indicates that the ligand-free state of NCBD can access sub-states resembling the bound state.

It may be tempting to conclude that sub-state 1 is isolated from other conformational sub-states. However, as noted above, closer examination of the cluster hierarchy (Fig. 36, Level 4) reveals that concerted structural changes along a complex pathway are required for NCBD to

adopt a binding competent conformation. By descending through the hierarchy, one can observe from Level 4 that a small subset of states (indicated by arrows on Fig. 36) closely resemble conformations in sub-state 1. This conformational state arises out of a rare state mostly consisting of conformers similar to sub-states 2 and 3 in level 6 of the hierarchy. Note that sub-state 2 in level 6 of the hierarchy consists of just 938 (or less than 0.05%) of the overall conformers, representing a rare transition. In this sub-state, the α_3 helix adopts a conformation that is more extended and hence represents an intermediate state that mediates a transition from sub-states 4 and 5 to the bound sub-state 1.

The observed clusters and conformational changes also provide a hypothesis for inter-conversions necessary for facilitating NCBD-ligand binding. For one, if NCBD is relatively compact, as in sub-states 4 and 5, then α_3 must initially undergo partial unfolding, seen in sub-states 2 and 3, to allow for the ligand to bind. Only then can α_3 adapt itself to form a full α-helix, as seen from experimental ensembles. Since we have not performed a comparison of our simulations with the ligand-bound state of either 1KBH or 2C52, we cannot provide a quantitative picture about the nature of changes that are required. However, based on the structural information available from experiments, such a partial unfolding-refolding pathway may indeed be responsible for facilitating NCBD's recognition of its binding partners. A similar scenario can also be proposed for α_1, which twists when binding with IRF3 (seen in Fig. 33D), although these experiments will be pursued in the future.

6.7 CONCLUSIONS AND FUTURE WORK

As part of pursuing further work in the area, we propose to incorporate simulations from a second NMR ensemble (1JJS) as well as several ligand-bound conformations to map out the conformational landscape of NCBD. Furthermore, by extending the Markov diffusion framework, we will elucidate the kinetic rates of significant conformational transitions.

The methodologies we have put forward yield the following insights: (a) ligand-free NCBD can indeed access conformations representative of the ligand-bound form and (b) structural

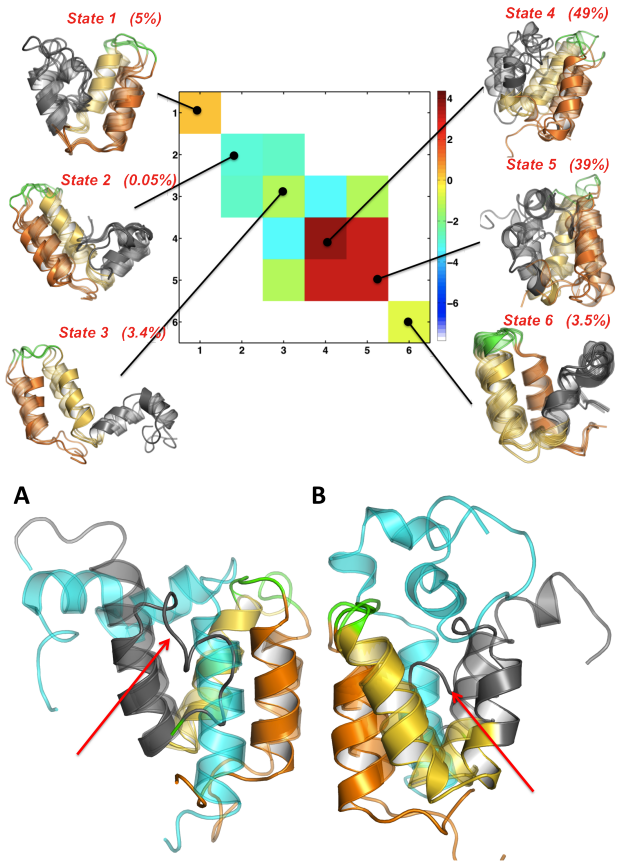

Figure 37: **Intermediate states of ligand-free NCBD enable access to ligand-bound conformations** Intermediate states of ligand-free NCBD enable access to ligand-bound conformations (a) Log affinities between sub-states at hierarchy level 6 are shown. For each of the 6 clusters, an ensemble of random conformers within that cluster are shown, and the percent of total frames within the cluster is given. High affinity (red) between two clusters indicate that those clusters are similar in dQAA space. Low affinity (blue - white) indicates that clusters have low similarity in dQAA space. (b) Comparing NCBD ensembles with the bound ligands (A) ACTR (1KBH; cyan) and (B) SRC1 (2C52; cyan) showing the orientations of α_3 indicated by red arrows.

changes required for ligand-free NCBD to access states that resemble ligand-bound conformations require concerted changes throughout the protein. We show that within our simulations, ligand-free α_1 and α_2 orientations largely resemble those of ligand-bound conformations; α_3 however, can exhibit a wide degree of flexibility and does not generally sample conformations that are similar to ligand-bound states.

7.0 CONCLUSION

Since cytologists began employing computers for automated screening in the 1950s [141], computers have been assisting disease diagnosis and prognosis. In addition to performing simple tasks in place of cytologists and pathologists, algorithms can identify novel disease features, extending current knowledge of disease [13]. Consider the CT-scan in Figure 7 that was used in an attention study by Drew, et al [5]. When 24 radiologists examined the image for lung nodules, 83% of them did not notice the gorilla in the slide. A non-specialist observer, who is not trained to search for nodules, might notice the gorilla right away. However, the presence of the gorilla was not an important factor for the radiologists' test. A computer algorithm specifically designed to look for lung nodules would function more like the radiologists, and quickly scan each image region for nodules, without taking in the image as a whole. However, mimicking the search method used by experts exactly may not make use of the strengths of a computer. That is, computers can analyze more information from the image simultaneously than a human, and they are capable of picking up patterns not apparent to humans [7]. A recent study showed how machine learning could uncover novel diagnostic/prognostic features in stained images that had been analyzed in the same way for the past century [13].

In addition to uncovering new features, computers can simply assist experts in their tasks. For example, while a pathologist (user) observed an image, the computer could also detect salient features in the image, and learn each user's strengths and weaknesses. If a certain user were known for missing nodes that were smaller than average, the computer could remind the user of those nodes before the user assigned a diagnosis. If the user neglected an unfamiliar shape in the image, such as the gorilla in Figure 7, and the computer identified it as diagnostically significant, the computer could highlight it for the user. Additionally, in ambiguous cases, the computer

Figure 38: CT scan [5]

could rapidly pull up examples of previous similar cases, which the pathologist could use to help make a diagnosis, or even identity names of other pathologists who are more confident with similar cases, to provide additional expertise. There is also much potential for computational pathology to be used in teaching. While medical students and residents analyze test slides, the computer would be able to point out features that they missed or overemphasized in their analysis, as well as provide the student with additional cases in their trouble areas.

Unlike humans, computers do not get tired or lose focus, thus their test results are consistently equally reliable, whereas an expert may have varied results, depending on times of day and other factors. While computational models may not be as accurate as expert analyses, a reasonably accurate algorithm should be able to screen through images and assign less challenging cases reliable, unbiased diagnostic labels, while marking more ambiguous images for expert analyses, as is the case with pap smears [141]. Such a system would allow experts to focus their analysis on difficult cases at times when they are most alert.

Disagreements between pathologists are especially common in critical cases where a patient is between two grades of cancer requiring different treatment plans [63]. In such cases, a computer algorithm could analyze an image and display an unbiased list of factors in favor of each diagnosis as a means of moderating the disagreement. A recent breast cancer study among eight experienced laboratories showed an inter-lab intraclass correlation coefficient of 0.71 in scoring of a common biomarker used to assess cancer proliferation, compared to an intra-lab correlation coefficient of 0.94 [142]. They found that the inter-lab discrepancy was contributed to by factors such as selection of tumor region for analysis, methods for quantifying the biomarker, and subjective assessment of biomarker values. Employing a standard algorithm for any one, if not all three, of these tasks would remove the inter-lab variability in that task, allowing for a standardized, unbiased methodology for scoring of this biomarker. Such unbiased analyses in all fields of pathology are necessary for laboratories to be able to communicate effectively with each other, and thus unencumber technological advancement.

Review of contributions

Our epithelial classification method for Barrett's Esophagus images enables rapid identification of epithelial nuclei in tissue images, on which phase can be computed to detect pre-cancerous changes in cell nuclei. If these optical biomarkers are shown to be effect on a larger scale, SL-QPM imaging could be implemented on endoscopes for live scanning of tissue for pre-cancerous lesions without necessity of biopsy. Moreover, as recent work has shown the benefit of analyzing diverse cell types individually, epithelial nuclei segmentation has widespread use among computational biology efforts [13, 66, 143]. Similarly, online epithelial nuclei segmentation may benefit new imaging technologies for assessing the effectiveness of cancer therapies [144].

We have presented a novel quantitative model of whole slide lung tissue images through the spatial arrangement of diagnostically significant tissue histologies, and have shown that these models relate to disease and have potential to be used for computational diagnosis. This model need not be limited to interstitial lung disease, but could also be applied to any process which affects tissue architecture, such as development, cancer, and aging, as well as other diseases [145].

On the molecular level, we have explored the conformational landscape of an intrinsically

disordered protein implicated in leukemia. The described method is able to identify bottleneck states in the protein's landscape that could be targeted by medical therapies in order to lock the protein into a certain state. As intrinsically disordered proteins are involved in around 50% of cancers, methods such as these are needed for computational drug design, so that early disease detection can be complemented by optimal drug therapy.

Future Work

The projects described here could be expanded further. Pathologists pay heed to specific nuclei types when diagnosing disease - for example, location and density of lymphocytes is used for diagnosing ILD and some cancers. The feature set used in Aim II could be improved by cell-type specific features, which would require cell classification methods such as the epithelial nuclei classification algorithm presented in Aim I. In turn, Aim I could be expanded to assign multi-class labels to all tissue components, instead of only nuclei classification. A markov random field designed to label stroma and lumen as well as nuclei has potential to classify nuclei with even higher accuracy, as location and orientation of nuclei with respect to lumen is an important characteristic for identifying epithelial nuclei, and in the current implementation this characteristic is only incorporated implicitly in the feature set. In Figure 7, an example of a hybrid form of Aims I and II designed to form tissue histology states not simply using features averaged over tissue blocks, but over superpixels containing single cells, is presented.

Additionally, location of architectural structures such as interstitial septum and lung pleura would further improve the spatial architecture matrices presented in Aim II, as location of inflammation with respect to these structures is an important diagnostic factor. Furthermore, the classifier would benefit from clinical information not included in the current implementation, such as smoking status, age, gender, and breathing ability.

The current GUI presented in Aim I for correcting nuclei labels automatically predicts putative nuclei and classifies a subset of these nuclei as epithelia, which are optionally presented to the user for verification. This system could be improved by incorporating a learning aspect to the GUI, so that the epithelial selection is adjusted to the user's preference after each examined image. If features were added to capture other common nuclei architectures from various diseases, such as rings and clusters, this method could be used to learn to predict a selection of nuclei

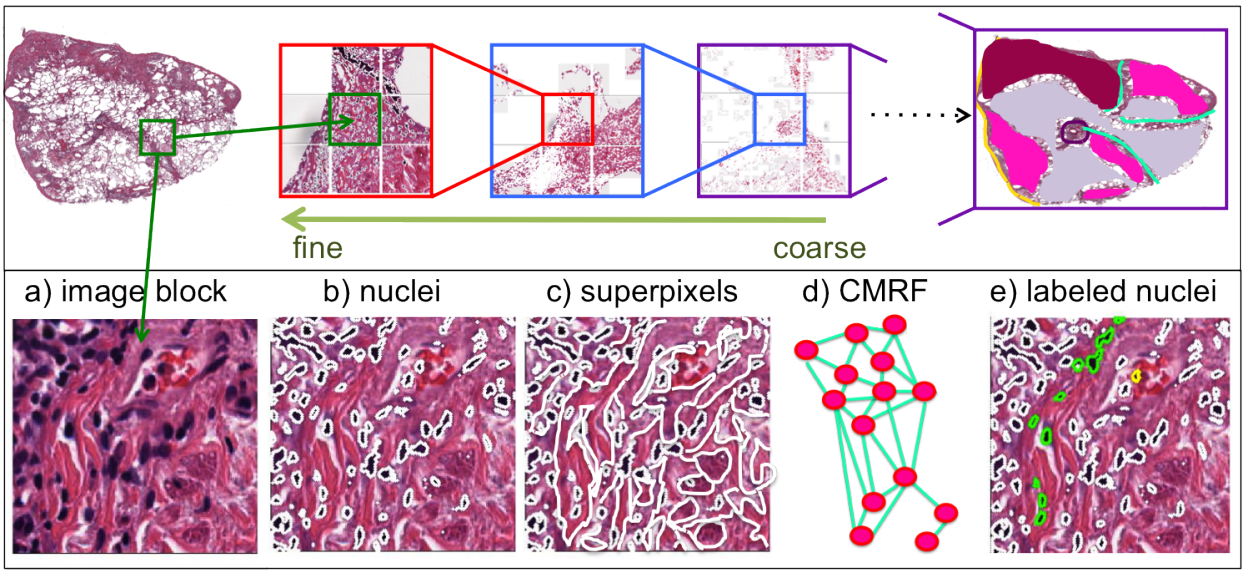

Figure 39: Hybrid version of Aims I & II: Labeling cells within blocks according to cell-type using an MRF, as in Aim I, would allow improved characterization of blocks according to tissue type, and a more accurate representation of the tissue as a whole. Computational efficiency could be maintained by performing this analysis hierarchically initializing with the coarsest level and biasing cell-level labels according to block-level labels. Additionally, an MRF would be used on the blocks to smooth tissue labels across neighboring block labels

in any disease type, based on a user's preference on an initial image set. Preliminary work has shown that the putative nuclei segmentation is reasonable for breast and lung tissue (not shown), so the future work could focus on gathering architectural features from a wide range of diseases and developing a GUI that can learn on the fly.

On the level of proteins, the algorithm for determining a set of conformational states for intrinsically disordered proteins could be demonstrated on a larger set of proteins and compared to more experimental data to establish its ability to model the landscape of these flexible proteins. Furthermore, the algorithm could be expanded to determine time-scales of transitioning between states.

BIBLIOGRAPHY

[1] Wayne A Phillips, Reginald V Lord, Derek J Nancarrow, David I Watson, and David C Whiteman. Barrett's esophagus. *Journal of Gastroenterology and Hepatology*, 26(4):639–648, 2011.

[2] RK Bista, S Uttam, DJ Hartman, W Qiu, J Yu, L Zhang, RE Brand, and Y Liu. Investigation of nuclear nano-morphology marker as a biomarker for cancer risk assessment using a mouse model. *J Biomed Opt.*, 17(6), 2012.

[3] John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679 –698, nov. 1986.

[4] Arun Devakonda, Suhail Raoof, Arthur Sung, William D. Travis, and David Naidich. Bronchiolar disorders: A clinical-radiological diagnostic algorithm. *CHEST Journal*, 137(4):938–951, 2010.

[5] Trafton Drew, Melissa L.H. Voe, and Jeremy M. Wolfe. The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science*, 24(9):1848–1853, 2013.

[6] M Bibbo, P H Bartels, H E Dytch, and G L Wied. Computed cell image information. *Monogr Clin Cytol*, 9:62–100, 1984.

[7] P Bartels, M Bibbo, G Olson, and G Wied. *Cell Image Analysis in Quantitative Cytology*, pages 59–90. Springer Netherlands, 1984.

[8] Jason Hipp, Jerome Cheng, Mehmet Toner, Ronald Tompkins, and Ulysses Balis. Spatially Invariant Vector Quantization: A pattern matching algorithm for multiple classes of image subject matter including pathology. *Journal of Pathology Informatics*, 2(1):13, 2011.

[9] F Ghaznav, A Evans, A Madabhushi, and M Feldman. Digital imaging in pathology: Whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease*, 8:331–359, 2012.

[10] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, and B. Yener.

Histopathological image analysis: A review. *Biomedical Engineering, IEEE Reviews in*, 2:147 –171, 2009.

[11] Thomas Bauer, Lynn Schoenfield, Renee Slaw, Lisa Yerian, Zhiyuan Sun, and Walter Henricks. Validation of whole slide imaging for primary diagnosis in surgical pathology. *Archives of Pathology & Laboratory Medicine*, Jan 2013. doi: 10.5858/arpa.2011-0678-OA.

[12] Yinyin Yuan, Henrik Failmezger, Oscar M. Rueda, H. Raza Ali, Stefan GrŁf, Suet-Feung Chin, Roland F. Schwarz, Christina Curtis, Mark J. Dunning, Helen Bardwell, Nicola Johnson, Sarah Doyle, Gulisa Turashvili, Elena Provenzano, Sam Aparicio, Carlos Caldas, and Florian Markowetz. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science Translational Medicine*, 4(157), 2012.

[13] Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Robert J. Marinelli, Torsten O. Nielsen, Marc J. van de Vijver, Robert B. West, Matt van de Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113, 2011.

[14] Wilma E Mesker. The intra-tumor stroma as simple parameter for prognostication. *Science Translational Medicine*, Response, 2012.

[15] Ji-Yeon Yang, Kosuke Yoshihara, Kenichi Tanaka, Masayuki Hatae, Hideaki Masuzaki, Hiroaki Itamochi, Masashi Takano, Kimio Ushijima, Janos L. Tanyi, George Coukos, Yiling Lu, Gordon B. Mills, and Roel G.W. Verhaak. Predicting time to ovarian carcinoma recurrence using protein markers. *The Journal of Clinical Investigation*, 123(9):3740–3750, 9 2013.

[16] Hong Chai and Robert E. Brown. Field effect in canceran update. *Annals of Clinical & Laboratory Science*, 39(4):331–337, 2009.

[17] Danely P. Slaughter, Harry W. Southwick, and Walter Smejkal. field cancerization in oral stratified squamous epithelium. clinical implications of multicentric origin. *Cancer*, 6(5):963–968, 1953.

[18] A Laurinavicius, A Laurinaviciene, D Dasevicius, N Elie, B Plancoulaine, C Bor, and P Herlin. Digital image analysis in pathology: Benefits and obligation. *Analytical Cellular Pathology*, 35:75–58, 2012.

[19] Thomas J. Fuchs and Joachim M. Buhmann. Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics*, 35:515 – 530, 2011.

[20] KE Fasanella, RK Bista, K Staton, S Rizvi, S Uttam, C Zhao, A Sepulveda, RE Brand, K McGrath, , and Y Liu. Nuclear nano-architecture markers of gastric cardia and upper

squamous esophagus detect esophageal cancer "field effect". *Journal of Cancer*, 4(8):626–634, 2013.

[21] Hang Chang, Ju Han, A. Borowsky, L. Loss, J.W. Gray, P.T. Spellman, and B. Parvin. Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. *Medical Imaging, IEEE Transactions on*, 32(4):670–682, 2013.

[22] Qixing Huang, Mei Han, Bo Wu, and S. Ioffe. A hierarchical conditional random field model for labeling and segmenting images of street scenes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1953–1960, 2011.

[23] K. Isse, A. Lesniak, K. Grama, B. Roysam, M.I. Minervini, and A.J. Demetris. Digital transplantation pathology: combining whole slide imaging, multiplex staining and automated image analysis. *Am J Transplant*, 12(1):27–37, 2012.

[24] S.J.D. Prince. *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.

[25] Andrew Blake and M. Isard. *Active Contours: The Application of Techniques from Graphics,Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition, 1998.

[26] P.F. Felzenszwalb and R. Zabih. Dynamic programming and graph algorithms in computer vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):721 –740, april 2011.

[27] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[28] S. Ali and A. Madabhushi. Graphical processing unit implementation of an integrated shape-based active contour: Application to digital pathology. *J Pathol Inform.*, 2, 2011.

[29] G. Srinivasa, M. Fickus, M.N. Gonzalez-Rivero, S.Y. Hsieh, Yusong Guo, A.D. Linstedt, and J. Kovacevic. Active mask segmentation for the cell-volume computation and golgi-body segmentation of hela cell images. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 348 –351, may 2008.

[30] Peter J. Schüffler, Thomas J. Fuchs, Cheng Soon Ong, Volker Roth, and Joachim M. Buhmann. Computational tma analysis and cell nucleus classification of renal cell carcinoma. In *Proceedings of the 32nd DAGM conference on Pattern recognition*, pages 202–211, Berlin, Heidelberg, 2010. Springer-Verlag.

[31] H Digabel and C Lantuejoul. *Iterative algorithms*. J.-L. Chermant, Ed., Riederer Verlag, Stuttgart, 1978.

[32] Elena Bernardis and Stella X. Yu. Pop out many small structures from a very large microscopic image. *Medical Image Analysis*, 15(5):690 – 707, 2011.

[33] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *Biomedical Engineering, IEEE Transactions on*, 57(4):841–852, 2010.

[34] Stephan Wienert, Daniel Heim, Kai Saeger, Albrecht Stenzinger, Michael Beil, Peter Hufnagl, Manfred Dietel, Carsten Denkert, and Frederick Klauschen. Detection and segmentation of cell nuclei in virtual microscopy images: A minimum-model approach. *Scientific Reports*, 2(503), 2012.

[35] N Otsu. A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66, 1979.

[36] Xian Du and Sumeet Dua. Segmentation of fluorescence microscopy cell images using unsupervised mining. *Open Med Inform J.*, 3:4149, 2010.

[37] L.P. Coelho, A. Shariff, and R.F. Murphy. Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, pages 518–521, 2009.

[38] D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995.

[39] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):629 –639, jul 1990.

[40] Markus Grasmair and Frank Lenzen. Anisotropic total variation filtering. *Applied Mathematics & Optimization*, 62(3):323–339, 2010.

[41] Hui Han Chin, Aleksander Madry, Gary L. Miller, and Richard Peng. Runtime guarantees for regression problems. In *ITCS*, pages 269–282, 2013.

[42] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[43] M Foracchia, E Grisan, and A Ruggeri. Luminosity and contrast normalization in retinal images. *Med Image Anal.*, 9(3):179–90, 2005.

[44] Serge Beucher and Christian Lantujoul. Use of watersheds in contour detection. *International workshop on image processing, real-time edge and motion detection*, 1979.

[45] Norberto Malpica, Carlos Ortiz de Solrzano, Juan Jos Vaquero, Andrs Santos, Isabel Vallcorba, Jos Miguel Garca-Sagredo, and Francisco del Pozo. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, 28(4):289–297, 1997.

[46] M.E. Plissiti, C. Nikou, and A. Charchanti. Watershed-based segmentation of cell nuclei boundaries in pap smear images. In *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*, pages 1–4, 2010.

[47] D. Iacoviello and U. Andreaus. *Biomedical Imaging and Computational Modeling in Biomechanics*. Lecture Notes in Computational Vision and Biomechanics. Springer, 2012.

[48] Varun Oswal, Ashwin Belle, Robert Diegelmann, and Kayvan Najarian. An entropy-based automated cell nuclei segmentation and quantification: Application in analysis of wound healing process. *Computational and Mathematical Methods in Medicine*, 2013, 2013.

[49] J Yao, S Fidler, and R Urtasun. Describing the scene as a whole: joint object detection, scene classification, and semantic segmentation. *CVPR*, 2012.

[50] P Arbelarz, M Maire, C Fowlkes, and J Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.

[51] Roozbeh Mottaghi, Sanja Fidler, Jian Yao, and Raqul Urtasun. Analyzing semantic segmentation using hybrid human-machine crfs. *CVPR*, 2013.

[52] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. *CVPR (oral)*, 2013.

[53] Corinna Cortes and Vladimir N Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[54] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:2006, 2006.

[55] Kumar V, A. K Abbas, J. C Aster, and S. L. Robbins. *Robbins basic pathology. 9th ed.* Elsevier/Saunders, 2013.

[56] Nicholas J Shaheen and Joel E Richter. Barrett's oesophagus. *The Lancet*, 373(9666):850 – 861, 2009.

[57] Emily G. Barr Fritcher, Shannon M. Brankley, Benjamin R. Kipp, Jesse S. Voss, Michael B. Campion, Larry E. Morrison, Mona S. Legator, Lori S. Lutzke, Kenneth K. Wang, Thomas J. Sebo, and Kevin C. Halling. A comparison of conventional cytology, dna ploidy analysis, and fluorescence in situ hybridization for the detection of dysplasia and adenocarcinoma in patients with barrett's esophagus. *Human pathology*, 39(8):1128–1135, 2008.

[58] TJ Hayeck, CY Kong, SJ Spechler, GS Gazelle, and C. Hur. The prevalence of barrett's esophagus in the us: estimates from a simulation model confirmed by seer data. *Dis Esophagus*, 2010.

[59] Shanmugarajah Rajendra and Prateek Sharma. Management of barrett's oesophagus and intramucosal oesophageal cancer, a review of recent development. *Therapeutic Advances in Gastroenterology*, 5(5):285–299, 2012.

[60] Alan J. Cameron. Diagnosis and treatment of barrett's esophageal adenocarcinoma. *Esophagus*, 1:31–35, 2003.

[61] James Mueller, Martin Werner, and Manfred Stolte. Barrett's esophagus: Histopathologic definitions and diagnostic criteria. *World Journal of Surgery*, 28:148–154, 2004.

[62] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, 5th Edition*. Garland Science. Kindle Edition., 2012.

[63] David Fleischer, Amitabh Chak, and Gary Falk. Low-grade dysplasia in patients with barrett's esophagus - ablate? *AGSE Leading Edge*, 1(3), 2011.

[64] RajanK. Bista, Pin Wang, Rohit Bhargava, Shikhar Uttam, DouglasJ. Hartman, RandallE. Brand, and Yang Liu. Nuclear nano-morphology markers of histologically normal cells detect the field effect of breast cancer. *Breast Cancer Research and Treatment*, 135:115–124, 2012.

[65] P Wang, R Bista, R Bhargava, RE Brand, and Y Liu. Spatial-domain low-coherence quantitative phase microscopy for cancer diagnosis. *Opt Lett.*, 35(17):2840–2842, 2010.

[66] N Linder, J Konsti, R Turkki, E Rahtu, M Lundin, S Nordling, C Haglund, T Ahonen, M Pietikaeinen, and J. Lundin. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagn Pathol.*, 7(22), 2012.

[67] S Gould. *Probabilistic models for region-based scene understanding*. PhD thesis, Stanford University, 2010.

[68] Hui Han Chin. Applications of spectral algorithms. Bachelor thesis, Carnegie Mellon University, 2012.

[69] Ioannis Koutis and Gary L. Miller. A linear work, o(n1/6) time, parallel algorithm for solving planar laplacians. *Proc. 18th ACM-SIAM Symposium on Discrete Algorithms (SODA 2007)*, 2007.

[70] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

[71] Marius Leordeanu and Martial Hebert. Efficient map approximation for dense energy functions. In *International Conference on Machine Learning 2006*, May 2006.

[72] R. von Mises and H. Pollaczek-Geiringer. Praktische verfahren der gleichungsauflsung. *ZAMM - Zeitschrift fr Angewandte Mathematik und Mechanik*, 9:152–164, 1929.

[73] C. S. Chennubhotla and A. Jepson. Eigencuts: Half-lives of eigenflows for spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 689–696, 2003.

[74] Chakra Chennubhotla and Allan Jepson. Hierarchical eigensolver for transition matrices in spectral methods. *Neural Info. Proc. Sys.*, 17:273–280, 2005.

[75] C Chennubhotla and I Bahar. Markov methods for hierarchical coarse graining of large protein dynamics. *Journal of Computational Biology. Also appeared in Proceedings of the Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB), Venice, Italy (April 2-5, 2006)*, 14(6):765–776, 2007.

[76] C. Chennubhotla and I. Bahar. Markov propagation of allosteric effects in biomolecular systems. *Mol. Sys. Biol.*, 2:36, 2006.

[77] I. Bahar, C. Chennubhotla, and D. Tobi. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Cur. Op. Struct. Biol.*, 17:633–640, 2007.

[78] Andrej Savol, Virginia Burger, Pratul Agarwal, Arvind Ramanathan, and Chakra Chennubhotla. QAARM: Quasi-anharmonic auto-regressive model reveals molecular recognition pathways in ubiquitin - SUBMITTED. *Bioinformatics*, 27(13):i52–i60, 2011.

[79] Virginia Burger and Chakra Chennubhotla. Nhs: Network-based hierarchical segmentation for cryo-electron microscopy density maps. *Biopolymers*, 97(9):732–741, 2012.

[80] Isabelle Couillin, Virginie Peetrilli, and Fabio Martinon. *The Inflammasomes*. Springer, 2011.

[81] Adrien Depeursinge, Jimison Iavindrasana, Asmaea Hidki, Gilles Cohen, Antoine Geissbuehler, Alexandra Platon, Pierre-Alexandre Poletti, and Henning Mller. Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization. *J. Digital Imaging*, 23(1):18–30, 2010.

[82] K R Flaherty, E L Thwaite, E A Kazerooni, B H Gross, G B Toews, T V Colby, W D Travis, J A Mumford, S Murray, A Flint, J P Lynch, and F J Martinez. Radiological versus histological diagnosis in uip and nsip: survival implications. *Thorax*, 58(2):143–148, 2003.

[83] Sonal Kothari, John H Phan, Todd H Stokes, and May D Wang. Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, 20(6):1099–1108, 2013.

[84] MN Gurcan, J Kong, O Sertel, BB Cambazoglu, J Saltz, and U Catalyurek. Computerized pathological image analysis for neuroblastoma prognosis. *AMIA Annu Symp Proc.*, pages 304–308, 2007.

[85] L. Gorelick, O. Veksler, M. Gaed, J.A. Gomez, M. Moussa, G. Bauman, A. Fenster, and A.D. Ward. Prostate histopathology: Learning tissue component histograms for cancer detection and classification. *Medical Imaging, IEEE Transactions on*, 32(10):1804–1818, 2013.

[86] Chao Wang, Thierry Pcot, Debra L Zynger, Raghu Machiraju, Charles L Shapiro, and Kun Huang. Identifying survival associated morphological features of triple negative breast cancer using multiple datasets. *Journal of the American Medical Informatics Association*, 20(4):680–687, 2013.

[87] Christina Mueller-Mang, Claudia Grosse, Katharina Schmid, Leopold Stiebellehner, and Alexander A. Bankier. What every radiologist should know about idiopathic interstitial pneumonias. *RadioGraphics*, 27(3):595–615, 2007.

[88] Robert M. Haralick, K. Shanmugam, and Its'hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(6):610 – 621, nov. 1973.

[89] Markus B. Huber, Kerstin Bunte, Mahesh B. Nagarajan, Michael Biehl, Lawrence A. Ray, and Axel Wismller. Texture feature ranking with relevance learning to classify interstitial lung disease patterns. *Artificial Intelligence in Medicine*, 56(2):91 – 97, 2012.

[90] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[91] Monika Fuxreiter, Peter Tompa, István Simon, Vladimir N Uversky, Jeffrey C Hansen, and Francisco J Asturias. Malleable machines take shape in eukaryotic transcriptional regulation. *Nat. Chem. Biol.*, 4(12):728–737, 2008.

[92] M Madan Babu, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. Intrinsically disordered proteins: regulation and disease. *Curr. Opi. Struct. Biol.*, 21(3):432 – 440, 2011.

[93] H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3):197–208, 2005.

[94] A Keith Dunker, Israel Silman, Vladimir N Uversky, and Joel L Sussman. Function and structure of inherently disordered proteins. *Curr. Opi. Struct. Bio.*, 18(6):756–764, 2008.

[95] Massimo Stefani. Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1739(1):5–25, 2004.

[96] Q C Zhang, T l Yeh, A Leyva, L G Frank, J Miller, Y E Kim, R Langen, S Finkbeiner, M L Amzel, C A Ross, and M A Poirier. A Compact Model of huntingtin Toxicity. *J. Biol. Chem.*, 286(10):8188–8196, 2011.

[97] Mari Luz Acevedo and W. Lee Kraus. Mediator and p300/cbp-steroid receptor coactivator complexes have distinct roles, but function synergistically, during estrogen receptor alpha-dependent transcription with chromatin templates. *Mol. Cell. Biol.*, 23(1):335–348, 2003.

[98] Hui Li and J. Don Chen. The receptor-associated coactivator 3 activates transcription through creb-binding protein recruitment and autoregulation. *J. Biol. Chem.*, 273(10):5948–5954, 1998.

[99] Wei Gu, Xiao-Lu Shi, and Robert G. Roeder. Synergistic activation of transcription by cbp and p53. *Nature*, 387(6635):819–823, 1997.

[100] Sarah Burge, Daniel P. Teufel, Fiona M. Townsley, Stefan M. V. Freund, Mark Bycroft, and Alan R. Fersht. Molecular basis of the interactions between the p73 n terminus and p300: Effects on transactivation and modulation by phosphorylation. *Proc. Natl. Acad. Sci. USA*, 106(9):3142–3147, 2009.

[101] Rongtuan Lin, Christophe Heylbroeck, Paula M. Pitha, and John Hiscott. Virus-dependent phosphorylation of the irf-3 transcription factor regulates nuclear translocation, transactivation potential, and proteasome-mediated degradation. *Mol. Cell. Biol.*, 18(5):2986–2996, 1998.

[102] Kirsten E. S. Scoggin, Aida Ulloa, and Jennifer K. Nyborg. The oncoprotein tax binds the src-1-interacting domain of cbp/p300 to mediate transcriptional activation. *Mol. Cell. Biol.*, 21(16):5520–5530, 2001.

[103] Charles G Mullighan, Jinghui Zhang, Lawryn H Kasper, Stephanie Lerach, Debbie Payne-Turner, Letha A Phillips, Sue L Heatley, Linda Holmfeldt, J Racquel Collins-Underwood, Jing Ma, Kenneth H Buetow, Ching-Hon Pui, Sharyn D Baker, Paul K Brindle, and James R Downing. CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature*, 471(7337):235–239, 2011.

[104] Maria I Torres-Arzayus, Jaime Font de Mora, Jing Yuan, Francisca Vazquez, Roderick Bronson, Montserrat Rue, William R Sellers, and Myles Brown. High tumor incidence and activation of the PI3K/AKT pathway in transgenic mice define AIB1 as an oncogene. *Cancer Cell*, 6(3):263–274, 2004.

[105] Chul Won Lee, Maria A Martinez-Yamout, H Jane Dyson, and Peter E Wright. Structure of the p53 Transactivation Domain in Complex with the Nuclear Receptor Coactivator Binding Domain of CREB Binding Protein. *Biochemistry*, 49(46):9964–9971, 2010.

[106] Magnus Kjaergaard, Kaare Teilum, and Flemming M Poulsen. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc. Natl. Acad. Sci. USA*, 107(28):12535–12540, 2010.

[107] Stephen J Demarest, Maria Martinez-Yamout, John Chung, Hongwu Chen, Wei Xu, H Jane Dyson, Ronald M Evans, and Peter E Wright. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*, 415(6871):549–553, 2002.

[108] S J Demarest, S Deechongkit, H. J. Dyson, R. M Evans, and P. E. Wright. Packing, specificity, and mutability at the binding interface between the p160 coactivator and CREB-binding protein. *Protein Sci.*, 13(1):203–210, 2004.

[109] Charles H. Lin, Brian J. Hare, Gerhard Wagner, Stephen C. Harrison, Tom Maniatis, and Ernest Fraenkel. A small domain of cbp/p300 binds diverse proteins: Solution structure and functional studies. *Mol. Cell*, 8(3):581 – 590, 2001.

[110] G. Bowman, K. Beauchamp, G. Boxer, and Vijay S. Pande. Progress and challenges in the automated construction of markov models for full protein systems. *J. Chem. Phys.*, 131:124101, 2009.

[111] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal component analysis and long time protein dynamics. *J. Phys. Chem.*, 100(7):2567–2572, 1996.

[112] G Zaccai and B Jacrot. Small angle neutron scattering. *Annu. Rev. Biophysics. Bioeng.*, 12(1):139–157, 1983.

[113] Martin Kurylowicz, Ching-Hsing Yu, and Régis Pomès. Systematic study of anharmonic features in a principal component analysis of gramicidin a. *Biophys. J.*, 98(3):386 – 395, 2010.

[114] Arvind Ramanathan, Andrej Savol, Christopher Langmead, Pratul Agarwal, and Chakra Chennubhotla. Discovering conformational sub-states relevant to protein function. *PLoS ONE*, 6(1):e15827, 2011.

[115] NJ Deng, Weihua Zheng, Emillio Gallicchio, and Ronald M. Levy. Insights into the dynamics of hiv-1 protease: A kinetic network model constructed from atomistic simulations. *J. Am. Chem. Soc.*, 133(24):9387–9394, 2011.

[116] Faruck Morcos, Santanu Chatterjee, Christopher L. McClendon, Paul R. Brenner, Roberto López-Rendón, John Zintsmaster, Maria Ercsey-Ravasz, Christopher R. Sweet, Matthew P. Jacobson, Jeffrey W. Peng, and Jesús A. Izaguirre. Modeling conformational ensembles of slow functional motions in pin1-ww. *PLoS Comput. Biol.*, 6(12):e1001015, 2010.

[117] Gregory R. Bowman G. and Vijay S. Pande. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. USA*, 107(24):10890–10895, 2010.

[118] M. Heinig and D. Frishman. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucl. Acids Res.*, 32:W500–W502, 2004.

[119] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. The amber biomolecular simulation programs. *J. Comp. Chem.*, 26(16):1668–1688, 2005.

[120] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct. Func. Bioinfo.*, 65(3):712–725, 2006.

[121] A. Ramanathan and P. K. Agarwal. Computational identification of slow conformational fluctuations in proteins. *J. Phys. Chem. B*, 113:16669–16680, 2009.

[122] M. J. Harvey, G. Giupponi, and G. De Fabritiis. Acemd: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.*, 5(6):1632–1639, 2009.

[123] K. Anton Feenstra, Berk Hess, and Herman J. C. Berendsen. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comp. Chem.*, 20(8):786–798, 1999.

[124] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, 38:289–302, 2007.

[125] Jeetain Mittal and Robert B. Best. Tackling force-field bias in protein folding simulations: Folding of villin hp35 and pin ww domains in explicit water. *Biophys. J.*, 99(3):L26 – L28, 2010.

[126] L. S. Caves, J.D. Evanseck, and M. Karplus. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.*, 7(3):649–666, 1998.

[127] Alan Grossfield, Scott E. Feller, and Michael C. Pitman. Convergence of molecular dynamics simulations of membrane proteins. *Proteins: Struc. Func. Bio.*, 67(1):31–40, 2007.

[128] A. Altis, P. Nguyen, R. Hegger, and G. Stock. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.*, 126(24):244111, 2007.

[129] Y. Mu, P.H. Nguyen, and G. Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Struct. Func. Bio.*, 58(1):45–52, 2004.

[130] Gia G. Maisuradze and David M. Leitner. Free energy landscape of a biomolecule in dihedral principal component space: Sampling convergence and correspondence between structures and minima. *Proteins: Struct. Func. Bio.*, 67(3):569–578, 2007.

[131] B. Mao, M. R. Pear, J. A. McCammon, and S. H. Northrup. Molecular dynamics of ferrocytochrome c: anharmonicity of atomic displacements. *Biopolymers*, 21:1979–1989, 1982.

[132] T. Ichiye and M. Karplus. Anisotropy and anharmonicity of atomic fluctuations in proteins: implications for x-ray analysis. *Biochemistry*, 27(9):3487–3497, 1988.

[133] F. Pontiggia, G. Colombo, C. Micheletti, and H. Orland. Anharmonicity and self-similarity of the free energy landscape of protein g. *Phys. Rev. Lett.*, 98(4):048102, 2007.

[134] K. N. Woods. Solvent-induced backbone fluctuations and the collective librational dynamics of lysozyme studied by terahertz spectroscopy. *Phys. Rev. E*, 81(3):031915, 2010.

[135] Giorgio Schiró, Chiara Caronna, Francesca Natali, and Antonio Cupane. Direct evidence of the amino acid side chain and backbone contributions to protein anharmonicity. *J. Am. Chem. Soc.*, 132(4):1371–1376, 2010.

[136] H. Frauenfelder, F. Parak, and R. D. Young. Conformational substates in proteins. *Ann. Rev. Biophys. Biophys. Chem.*, 17:451–479, 1988.

[137] Mark A. Miller and David J. Wales. Energy landscape of a model protein. *J. Chem. Phys.*, 111(6610), 1999.

[138] Jean-Francois Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

[139] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[140] Chakra Chennubhotla and Ivet Bahar. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput Biol*, 3(9):1716–26, 2007.

[141] Robert C. Mellors, Adele Glassman, and G. N. Papanicolaou. A microfluorometric scanning method for the detection of cancer cells in smears of exfoliated cells. *Cancer*, 5(3):458–468, 1952.

[142] Mei-Yin C. Polley, Samuel C. Y. Leung, Lisa M. McShane, Dongxia Gao, Judith C. Hugh, Mauro G. Mastropasqua, Giuseppe Viale, Lila A. Zabaglo, Frdrique Penault-Llorca, John M.S. Bartlett, Allen M. Gown, W. Fraser Symmans, Tammy Piper, Erika Mehl, Rebecca A. Enos, Daniel F. Hayes, Mitch Dowsett, Torsten O. Nielsen, on behalf of the International Ki67 in Breast Cancer Working Group of the Breast International Group, and North American Breast Cancer Group. An international ki67 reproducibility study. *Journal of the National Cancer Institute*, 2013.

[143] J Kwak, S Hewitt, S Sinha, and R Bhargava. Multimodal microscopy for automated histologic analysis of prostate cancer. *BMC Cancer*, 11, 2011.

[144] Alex J. Walsh, Rebecca S. Cook, H. Charles Manning, Donna J. Hicks, Alec Lafontant, Carlos L. Arteaga, and Melissa C. Skala. Optical metabolic imaging identifies glycolytic levels, subtypes, and early-treatment response in breast cancer. *Cancer Research*, 73(20):6164–6174, 2013.

[145] Celeste M. Nelson and Mina J. Bissell. Of extracellular matrix, scaffolds, and signaling: Tissue architecture regulates development, homeostasis, and cancer. *Annual Review of Cell and Developmental Biology*, 22:287–309, 2006.