



HAL
open science

Characterization of fibronectin networks using graph-based representations of the fibers from 2D confocal images

Anca-Ioana Grapa

► **To cite this version:**

Anca-Ioana Grapa. Characterization of fibronectin networks using graph-based representations of the fibers from 2D confocal images. Signal and Image processing. Université Côte d'Azur, 2020. English. NNT : 2020COAZ4031 . tel-03052167

HAL Id: tel-03052167

<https://theses.hal.science/tel-03052167>

Submitted on 10 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Caractérisation des réseaux de fibronectine représentés par des graphes de fibres à partir d'images de microscopie confocale 2D

Anca-Ioana GRAPĂ

Equipe MORPHEME INRIA Sophia Antipolis – Méditerranée/IS5/IBV

**Présentée en vue de l'obtention
du grade de docteur en** Automatique,
Traitement du Signal et des Images
d'Université Côte d'Azur

Dirigée par : Xavier Descombes / Ellen
Van Obberghen-Schilling

Soutenue le : 11 juin 2020

Devant le jury, composé de :

Alin Achim, PR, University of Bristol
Laure Blanc-Féraud, DR, CNRS, Université Côte d'Azur
Isabelle Bloch, PR, LTCI, Télécom Paris
Jean-Pierre Da Costa, PR, Bordeaux Sciences Agro
Xavier Descombes, DR, INRIA, Université Côte d'Azur
Ellen Van Obberghen-Schilling, DR, IBV, Université
Côte d'Azur

CARACTÉRISATION DES RÉSEAUX DE FIBRONECTINE
REPRÉSENTÉS PAR DES GRAPHES DE FIBRES À PARTIR
D'IMAGES DE MICROSCOPIE CONFOCALE 2D

CHARACTERIZATION OF FIBRONECTIN NETWORKS
USING GRAPH-BASED REPRESENTATIONS OF THE
FIBERS FROM 2D CONFOCAL IMAGES

Jury:

Alin Achim, PR, University of Bristol	Rapporteur
Laure Blanc-Féraud, DR, CNRS, Université Côte d'Azur	Examineur
Isabelle Bloch, PR, LTCL, Télécom Paris	Examineur
Jean-Pierre Da Costa, PR, Bordeaux Sciences Agro	Rapporteur
Xavier Descombes, DR, INRIA, Université Côte d'Azur	Directeur de thèse
Ellen Van Obberghen-Schilling, DR, IBV, Université Côte d'Azur	Directeur de thèse

RÉSUMÉ

La fibronectine (FN) cellulaire, composante majeure de la matrice extracellulaire, est organisée en réseaux fibrillaires de manière différente suivant les deux extra-domaines EDB et EDA. Notre objectif a été le développement de biomarqueurs quantitatifs pour caractériser l'organisation géométrique des quatre variants de FN à partir d'images de microscopie confocale 2D, puis de comparer les tissus sains et cancéreux. Premièrement, nous avons montré à travers deux pipelines de classification fondés sur les curvelets et sur l'apprentissage profond, que les variants peuvent être distingués avec une performance similaire à celle d'un annotateur humain. Nous avons ensuite construit une représentation des fibres (détectées avec des filtres Gabor) fondée sur des graphes. Les variantes ont été classées en utilisant des attributs spécifiques aux graphes, prouvant que ceux-ci intègrent des informations pertinentes dans les images confocales. De plus, nous avons identifié différentes techniques capables de différencier les graphes, afin de comparer les variants de FN quantitativement et qualitativement. Une analyse des performances sur des exemples simples a montré la capacité des méthodes fondées sur l'appariement de graphes et le transport optimal, de comparer les graphes. Nous avons ensuite proposé différentes méthodologies pour définir le graphe représentatif d'une certaine classe. De plus, l'appariement de graphes nous a permis de calculer des cartes de déformation des paramètres entre tissus sains et cancéreux. Ces cartes ont ensuite été analysées dans un cadre statistique montrant si la variation du paramètre peut être expliquée ou non par la variance au sein d'une même classe.

Mots clés: traitement d'images, apprentissage automatique, appariement de graphes, cartes statistiques des paramètres, matrice extracellulaire, fibronectine, cancer.

ABSTRACT

A major constituent of the Extracellular Matrix is a large protein called the Fibronectin (FN). Cellular FN is organized in fibrillar networks and can be assembled differently in the presence of two Extra Domains, EDA and EDB. Our objective was to develop numerical quantitative biomarkers to characterize the geometrical organization of the four FN variants (that differ by the inclusion/exclusion of EDA/EDB) from 2D confocal microscopy images, and to compare sane and cancerous tissues. First, we showed through two classification pipelines, based on curvelet features and deep learning framework, that the FN variants can be distinguished with a similar performance to that of a human annotator. We constructed a graph-based representation of the fibers, which were detected using Gabor filters. Graph-specific attributes were employed to classify the variants, proving that the graph representation embeds relevant information from the confocal images. Furthermore, we identified various techniques capable to differentiate the graphs, allowing us to compare the FN variants quantitatively and qualitatively. Performance analysis using toy graphs showed that the methods, which are based on graph matching and optimal transport, can meaningfully compare graphs. Using the graph-matching framework, we proposed different methodologies for defining the prototype graph, representative of a certain FN class. Additionally, the graph matching served as a tool to compute parameter deformation maps between the variants. These deformation maps were analyzed in a statistical framework showing whether or not the variation of the parameters can be explained by the variance within the same class.

Keywords: image processing, machine learning, graph-matching, statistical parametric maps, extracellular matrix, fibronectin, cancer.

All you need are these: certainty of judgment in the present moment; action for the common good in the present moment; and an attitude of gratitude in the present moment for anything that comes your way. — Marcus Aurelius, Meditations, 9.6

ACKNOWLEDGMENTS

The journey that made possible the accomplishment of this manuscript and of the PhD experience, is far from being individual. Here is my attempt at outlining the impact of those around me throughout these last 3 years and a half, acknowledging the lessons I have learnt during this time.

I would like to start by expressing my gratitude to the people involved in the SIGNALIFE Labex program. This platform has provided the resources for the thesis development within an international framework, and facilitated the interaction with researchers across several life sciences institutes in the Nice Sophia-Antipolis region. I would like to acknowledge the efforts of Konstanze Beck, the Signalife Labex officer, for her meticulous guidance upon navigating the intricate administrative issues, even before arriving in Nice, and until the end of the contract.

My sincere gratitude goes to my thesis advisors, Xavier Descombes, Ellen Van Obberghen-Schilling and Laure-Blanc Féraud. Thank you for giving me the opportunity to start a thesis under your supervision, and for accepting me in your groups after a long period of looking for PhD options. Thank you for guiding me through these sometimes challenging years with patience, dedication, and with all the constructive criticism that I frequently needed, in order to improve during this rewarding learning experience.

Xavier, I have benefited greatly from your creativity and openness to ideas during this thesis. Thank you for always having an open door when I needed your advice, for your critical eye determining me to question everything I learn, and lastly for always seeking ways to improve. With you, there's always a better idea around the corner that one can explore. Laure, I am grateful for our interactions, for your attention to detail and inspiring rigour that have surely shaped this thesis. Thank you for prompting me to dig deeper into the understanding of all the concepts with calm and diligence, and always knowing how to ask the right questions. Ellen, your patience in face of my precarious knowledge of biology and your willingness to introduce me to this new world, have encouraged me to gain confidence and made me feel welcomed in your team. Your flexibility and openness to our proposals has challenged me to learn to better communicate my ideas, which was essential during our interdisciplinary collaboration.

I would like to acknowledge the efforts of the members of the thesis jury, who agreed to take the time to provide valuable insight over the course of our interactions. In particular, I would like to thank both reviewers, Jean-Pierre da Costa and Alin Achim for their insightful comments, for taking the time to go through the manuscript, and suggest improvements and alternative ideas. I am grateful to Isabelle Bloch for agreeing to preside over the jury. I would also like to thank them

for their responsiveness and encouragements during the defense schedule changes generated by the pandemic.

My greatest thank you goes to the team I have been part of during these years, Morpheme, joint team between INRIA, I3S and IBV. I have been fortunate to be part of an international group of people with whom I've shared so many moments. Be it long coffee breaks, picnics at the beach, lunches at the INRIA cafeteria, my all-time favourite apéro-pétanque, scientific conferences, cinema nights, wine/gastronomic and jazz music festivals, these interactions made the routine easier to bear and my life in Nice more interesting and enjoyable. Going back in time, I will mention Lola, my first office mate and patient guide in the beginning, Gaël, thank you for your support, friendship, lessons of French culture/language, pétanque and ahem, car driving. Arne, I am extremely grateful for our friendship consolidated during the coffee/afternoon breaks filled with philosophical discussions, and during the jazz festivals we enjoyed attending. The positive energy you usually radiate made this challenge a bit easier. Sarah, Somia, Cédric, you are gentle and generally very nice people to be around with. Kevin, I will remember your easy-going attitude and genuine ability to make people laugh, which I had the chance to benefit from, while sharing the same office. Friends from Media Coding team on the same floor: Cyprien, thank you for your gentleness, support and interesting discussions. And of course, for letting us prove our mastery at pétanque. Arnaud, thank you for your availability for any kind of question. I will also keep good memories of the occasional but engaging or helpful interactions over coffee breaks/lunch, with both permanent and non-permanent researchers/students: Melpo, Jean-Marie, Ninad, Fernando, Dino, Eric, Henrique, Luca, Agus, Diana, Danny, Vasilina, Simone, Froso, Emma, Flora, Xuchun, Rudan, Djampa, Zhankeng.

My second team, "Adhesion Signaling and Stromal Reprogramming in the Tumor Microenvironment" led by Ellen Van Obberghen-Schilling, has made a powerful impact over this thesis, despite having spent less time at the IBV institute on the Valrose campus. George, I am grateful for our collaborations and fruitful interactions. Your questions were always on point and you helped quite a great deal introducing me to biology, always trying your best to explain in layman terms the intricate terms and biochemical phenomena. I would like to acknowledge the efforts of Sébastien Schaub, who as I always said, was the optimal "interface" between us the researchers/engineers and the biologists. Thank you for proving to be an excellent "biology-maths" translator, for your ideas that helped this thesis in a great way, and for your availability. I am also grateful to other members of the team that were always present during our meetings, with helpful feedback: Delphine Ciais and Dominique Grall.

I keep nice memories of people that I briefly worked with, such as master and undergraduate students I had the fulfilling opportunity to teach and even learn from, both at PolyTech Nice and University of Nice Sophia Antipolis; on this note, I thank Marc Antonini and Frédéric Payan for giving me the chance to be in charge of the lab sessions. I would like to mention Avrajit Ghosh, the first student I had the chance to supervise, thank you for our collaboration. I will also mention here the network of international Signallife students-mostly researchers in biology that I interacted with during our workshops, as well as various doctoral training sessions. Lucie, Nino, Kavya, Cristina, Gaurav, and many others, thank you for the fun and

interesting chats during the trainings. To the secretaries at I3S, Nadia Belfegas-Fagues, and INRIA, Jane Desplanques and Isabelle Strobant, I sincerely thank you for assisting me with the administrative issues.

I was fortunate enough to have, during this journey, friends and acquaintances who brought a lot of value and support over these years. Thank you to my close friends at home, in Romania, and elsewhere in Europe. My life has been enriched with new friends here, whose support and companionship were both invigorating, and soothing. Konstantinos, you are an inspiring person and I am happy to have such a valuable friend in my life. I should perhaps blame the "infamous" 230 bus for providing the chance of really engaging and uplifting conversations: I am thinking of Ariel, Sidhant. Mohit, thanks for showing me that comforting discussions can go well with some (not too spicy) Indian food.

Lastly, I would like to thank all of my family, particularly my parents, and my two younger sisters. Without your help, support and love, none of this would have been possible. Thank you for being there for me, showing encouragement and understanding even in the most stressful and uncertain times. Sergiu, I am blessed to have had you by my side taking care of me even from distance, under any kind of circumstances. Your support and love are invaluable and this thesis is dedicated to you and my family.

CONTENTS

List of Figures	xii
List of Tables	xiv

I INTRODUCTION AND MAIN MOTIVATION

1 INTRODUCTION	3
1.1 Preface	3
1.2 General context	3
1.2.1 Brief introduction of biological aspects	4
1.2.2 Objective and motivations	4
1.3 Thesis organization	5
1.3.1 Main contributions	5
1.3.2 Manuscript organization	6
1.3.3 List of publications	7

II BIOLOGICAL BACKGROUND AND IMAGE DATABASE

2 BIOLOGICAL BACKGROUND	11
2.1 Extracellular Matrix and components	11
2.1.1 Fibronectin - Major Component of the ECM	12
2.1.2 Experimental procedure for FN matrix preparation	14
2.2 Confocal image database	15
2.2.1 Fluorescence principle	16
2.2.2 Confocal microscopy	16
2.3 Conclusions	18

III FIBRONECTIN VARIANTS CLASSIFICATION FROM CONFOCAL IMAGES

3 CLASSIFICATION OF THE FN VARIANTS FROM CONFOCAL IMAGES	25
3.1 Fiber detection using discrete curvelets	25
3.1.1 Fast Discrete Curvelet transform	26
3.1.2 Fiber detection using curvelets for classification	28
3.2 FN fiber classification using curvelets	30
3.2.1 Bag-of-words and image signatures	30
3.2.2 Classification using a DAG-SVM framework	31
3.2.3 Application to FN images classification with DAG-SVM, using a curvelet based representation	33
3.2.4 Classification of the FN variants using Convolutional Neural Nets	36
3.3 Conclusions	37

IV FIBRONECTIN VARIANTS FIBER DETECTION AND CHARACTERIZATION WITH GRAPHS

4 CONSTRUCTION OF THE GRAPH-BASED REPRESENTATION OF FN NETWORKS	41
4.1 Gabor filters	41

4.1.1	Gabor kernels definition	41
4.1.2	Filtering of the FN confocal images using Gabor kernels . . .	43
4.2	Graph representation of the FN fibers- Methodology	44
4.2.1	Computation of the graphs associated to FN morphological skeletons	45
4.2.2	Post-improvement methodology for the FN graph-based representation	46
4.3	Conclusions	48
5	LOCAL CHARACTERIZATION OF FIBER FEATURES USING THE GRAPH-BASED REPRESENTATION	53
5.1	Feature extraction from graph-based FN representation	53
5.2	PCA visualisation of the Gabor and graph-based FN fiber features .	54
5.2.1	Classification of the Gabor and graph-based FN fiber features	55
5.3	Conclusions	55
6	GLOBAL STATISTICAL CHARACTERIZATION OF THE FN PARAMETER MAPS	59
6.1	Gaussian Random fields and decision testing of parametric maps . . .	59
6.2	Application of the statistical analysis to the study of FN parametric maps	62
6.2.1	Statistical analysis based on GRF	64
6.2.2	Statistical analysis of the empirical distributions	67
6.3	Conclusions	68
V	TOWARD MODELLING	
7	GRAPH MATCHING FOR GRAPH COMPARISON	77
7.1	Graph matching general background	77
7.1.1	Exact and inexact matching	79
7.1.2	Graph kernels and graph embeddings	81
7.2	Formulation of the graph-matching problem	82
7.2.1	General matching	82
7.2.2	Different instances of graph matching	84
7.2.3	Many-to-many-assignment problem	85
7.2.4	Matching of weighted labeled graphs	86
7.3	Summary	87
8	OPTIMAL TRANSPORT THEORY - APPLICATION TO HISTOGRAM AND GRAPH MATCHING	89
8.1	Optimal transport general background	89
8.2	Optimal transport for discrete domains	91
8.2.1	Optimal Transport with Linear Programming	92
8.2.2	Metric properties of optimal transport	93
8.2.3	Optimal transport 1D	94
8.3	Entropic Regularization of the OT	94
8.4	Gromov-Wasserstein discrepancy for optimal transport between structured objects	97
8.4.1	Gromov-Wasserstein discrepancy - adaptation to a graph-matching context	97
8.4.2	Gromov-Wasserstein barycenter	99

8.5	Summary	100
9	COMPARISON OF PERFORMANCES OF OPTIMAL TRANSPORT AND MANY-TO-MANY ASSIGNMENT	101
9.1	Toy graphs generation	101
9.2	Methodology for establishing a common framework for comparison	103
9.2.1	Many-to-many assignment framework parameter selection	104
9.2.2	Optimal transport framework parameter selection	104
9.3	Results and interpretation of the comparative performance analysis on toy-graphs	105
9.4	Conclusions	106
10	METHODOLOGIES FOR DERIVING THE REPRESENTATIVE OF A SET OF GRAPHS	109
10.1	Methodology based on a majority voting after matching to a common graph	111
10.2	Methodology based on a heuristic derived from the longest chains of matched nodes	120
10.3	Conclusions	121
11	STATISTICAL ANALYSIS OF PARAMETRIC DEFORMATION MAPS	125
11.1	Deformation maps based on graph matching	125
11.2	Statistical framework to quantify the parameter variation	126
11.2.1	Statistical analysis of the deformation maps based on GRF	126
11.2.2	Statistical analysis of the deformation maps based on the empirical distributions	128
11.3	Conclusions	129
VI CONCLUSIONS AND PERSPECTIVES		
12	CONCLUSIONS AND PERSPECTIVES	135
12.1	Conclusions	135
12.2	Perspectives	136
	BIBLIOGRAPHY	139

LIST OF FIGURES

Figure 2.1	Schematic representation of the extracellular matrix in relation to epithelial cells and vascular endothelial cells	12
Figure 2.2	Linear structure of Fibronectin	13
Figure 2.3	Protocol for the production and 2D confocal image acquisition of FN variants	15
Figure 2.4	Fluorescence principle	16
Figure 2.5	Comparison wide field-confocal	17
Figure 2.6	Basic setup of a confocal microscope	17
Figure 2.7	FN confocal images - Normal ECM and "Tumour-like" ECM FN B-A-	19
Figure 2.8	FN confocal images - Normal ECM and "Tumour-like" ECM FN B+A-	20
Figure 2.9	FN confocal images - Normal ECM and "Tumour-like" ECM FN B-A+	21
Figure 2.10	FN confocal images - Normal ECM and "Tumour-like" ECM FN B+A+	22
Figure 3.1	Frequency tiling of discrete curvelets	26
Figure 3.2	Curvelets in spatial and frequency domain at various scales and orientations	28
Figure 3.3	Curvelet scale decomposition for a sample image of 512×512 pixels	29
Figure 3.4	Reconstruction of a sample image after keeping 85% of total curvelet coefficients energy.	29
Figure 3.5	Bag of features pipeline - curveletes	31
Figure 3.6	Graphical representation of binary SVM	32
Figure 3.7	DAG-SVM Decision Scheme	32
Figure 3.8	Vertical and diagonal line of 1 pixel width	35
Figure 3.9	Curvelet coefficients energy distribution over the wedges	36
Figure 4.1	Gabor filter 3D view	42
Figure 4.2	Elliptic Gaussian rotation	43
Figure 4.3	List of Gaussian kernels of different orientations	43
Figure 4.4	List of Gabor kernels	43
Figure 4.5	Maximum Gabor filtered Image	44
Figure 4.6	Morphological skeleton of a rectangle	45
Figure 4.7	Morphological skeletons of the FN fibers and the corresponding graphs	47
Figure 4.8	Methodology for morphological skeleton and graph improvement	49
Figure 4.9	Pipeline for fiber detection and reconnection	50
Figure 4.10	Reconnected graphs of the four FN-specific variants (normal state)	51
Figure 5.2	PCA Analysis of the Gabor and graph-based FN fiber features	54

Figure 5.1	Graph-based normalized feature distributions	57
Figure 6.1	Connected components (clusters of pixels) taken at a certain intensity threshold	61
Figure 6.2	Computation of parameter (fiber length) maps from graph-based fiber representations	63
Figure 6.3	Gaussianization of the parametric maps using optimal transport	65
Figure 6.4	Result of Gaussianization of parametric map	65
Figure 6.5	Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the maximum cluster intensity)	70
Figure 6.6	Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the cluster surface)	71
Figure 6.7	Detection of the clusters considered foreign elements to the empirical distributions of maximum cluster intensity $p = 0.05$	72
Figure 6.8	Detection of the clusters considered foreign elements to the empirical distributions of cluster surface $p = 0.05$	73
Figure 7.1	Graph matching - Exact matching (Isomorphism)	79
Figure 7.2	Graph matching - Edit Distance	80
Figure 7.3	Graph embedding into a vector domain	82
Figure 7.4	Graph matching principle	83
Figure 7.5	Graph matching- discrepancy between one graph and the optimally permuted version of the second one	84
Figure 7.6	Many-to-one assignment between two graphs	85
Figure 7.7	Many-to-many assignment as two many-to-one matchings	86
Figure 7.8	The effect of modifying the maximum number of nodes matched together	87
Figure 8.1	Relaxed and regularized optimal transport	90
Figure 8.2	Monge Map	92
Figure 8.3	Monge mass splitting: No feasible transport map	93
Figure 8.4	Kantorovich mass splitting	93
Figure 8.5	OT-1D:Permutation scheme and optimal assignment	95
Figure 8.6	OT-1D between 2 images	96
Figure 8.7	OT-1D for histogram conversion	96
Figure 8.8	Optimal transport-Gromov-Wasserstein discrepancy	97
Figure 8.9	Mass dispersion-matching graphs using optimal transport framework	99
Figure 9.1	Methodology for generating graphs with Voronoi cells and Poisson point process	102
Figure 9.2	Generated toy-graphs	102
Figure 9.3	Modified toy-graphs (rotation, removal of nodes)	103
Figure 9.4	Adjacency matrix type for a 4-nodes graph example (binary or the length of the shortest path between the nodes)	104
Figure 10.1	Prototype for isomorphic toy-graphs	113
Figure 10.2	Prototype for random toy-graphs - Configuration I	114

Figure 10.3	Prototype for random toy-graphs - Configuration II	115
Figure 10.4	Prototype for random toy-graphs - Configuration III	116
Figure 10.5	Prototype for random toy-graphs - Configuration IV	117
Figure 10.6	Prototype for FN networks - FN B-A+ (Normal) ECM and "Tumour-like" ECM FN B-A+	118
Figure 10.7	Prototype for FN networks - FN B-A+ (Normal) ECM and "Tumour-like" ECM FN B-A+	119
Figure 10.8	Methodology for deriving the representative graph based on chains of nodes connected among them across graphs	120
Figure 10.9	Prototype for random toy-graphs- Method based on the longest cycles of matched nodes	123
Figure 10.10	Prototype for random toy-graphs - Method based on the longest cycles of matched nodes	124
Figure 11.1	Computation of the fiber length difference after graph matching	126
Figure 11.2	Computation of the fiber length difference after graph matching between graph representations of Normal-Normal FN B-A+	127
Figure 11.3	Computation of the fiber length difference after graph matching between graph representations of Normal-Tumoral FN B-A+	128
Figure 11.4	Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the maximum cluster intensity) within the Fiber Length Deformation Map Normal-Tumoral	131
Figure 11.5	Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the surface of the cluster) within the Fiber Length Deformation Map Normal-Tumoral	132

LIST OF TABLES

Table 3.1	Confusion matrix automatic classification with curvelets	33
Table 3.2	Confusion matrix in percentage form - Trained specialist	34
Table 3.3	Confusion matrix in percentage form of the CNN classification of FN variant confocal images	37
Table 5.1	Confusion matrix in percentage form of the DAG-SVM classification of FN variants, using Gabor and graph-based features	55
Table 6.1	Average number and area of clusters per tumoral parametric map identified as foreign to a GRF, at various thresholds	66
Table 6.2	Average number and area of clusters per normal parametric map identified as foreign to a GRF, at various thresholds	66

Table 6.3	Average number and area of clusters per tumoral parametric map identified as foreign elements to the empirical distributions, at various thresholds	68
Table 6.4	Average number and area of clusters per normal parametric map identified as foreign elements to the empirical distributions, at various thresholds	68
Table 9.1	Maching cost for the perfect matching (PM), many to-many matching (MM) and optimal transport (OT) for graphs of 16 nodes	106
Table 9.2	Maching cost for the perfect matching (PM), many-to-many matching (MM) and optimal transport (OT). Graphs have 181 nodes and the transformations are: rotation, removal of one-degree node (Remove A), removal of multiple-degree node (Remove B). G and H have the following representations: binary adjacency matrices (Int1), shortest path integer values and subunitary values at order 2,3, total (Int2, Int3, IntT, Sub2, Sub3, SubT).	107
Table 11.1	Average number and area of clusters per normal-tumoral deformation map identified as foreign, at various thresholds	127
Table 11.2	Average number and area of clusters per normal-normal deformation map identified as foreign, at various thresholds	128
Table 11.3	Average number and area of clusters per normal-tumoral deformation map identified as foreign, ($p = 0.05$) to the empirical distributions of clusters (size and intensity) at various thresholds	129
Table 11.4	Average number and area of clusters per normal-normal deformation map identified as foreign, ($p = 0.05$) to the empirical distributions of clusters (size and intensity), at various thresholds	129

ACRONYMS

cFN	Cellular Fibronectin
CNN	Convolutional Neural Networks
DAG-SVM	Directed Acyclic Graph Support Vector Machines
ECM	Extracellular Matrix
EDA/EDB	Extra Domain A/B
FN	Fibronectin
GM	Graph Matching
GRF	Gaussian Random Field
MDS	Multi-dimensional Scaling
NP	Nondeterministic Polynomial Time
OT	Optimal Transport
PCA	Principal Component Analysis
pFN	Plasma Fibronectin
SVM	Support Vector Machines
TACS	Tumor Associated Collagen Signatures
TGF	Transformation Growth Factor

Part I

INTRODUCTION AND MAIN MOTIVATION

INTRODUCTION

1.1 PREFACE

Across biological systems, deciphering the molecular mechanisms of life remains one of the main challenges to be faced by researchers. To unravel the function of certain microscopic components, a thorough knowledge of their underlying composition is essential, along with the interactions within their micro-environment. Nowadays, getting access to the molecular structure is facilitated by the development of various imaging techniques. These are capable to provide an insight at a molecular level, thus enabling a better understanding of key processes that occur both in normal and disease states. Despite the ability of the dedicated imaging instruments (microscopes, digital scanners, etc.) to generate vast amount of data at high resolution, a full and comprehensive analysis is a laborious and difficult task, even when performed by a trained specialist.

The advances in digital image processing and modelling tools over the last years, have proven their effectiveness in supplementing the human observer analysis, through object detection, various structure segmentation and delineation, classification, statistical analysis, etc. The challenges, however, are numerous: from identifying the most effective tools that can extract meaningful information for the particular biological/biomedical context, designing specific approaches adapted to a particular set of images, to dealing with a shortage of data due to experimental constraints, etc.

This thesis, the result of a collaboration-based effort, attempts to bring forward a set of methodologies inspired from the signal/image processing and computer science fields. These approaches were adapted or developed specifically for the analysis of a biological problem, in the hope to provide a numerical ground that facilitates its analysis for potential clinical/therapeutical purposes.

Deciphering the role of certain structures within their environment comes down to understanding their structure.

1.2 GENERAL CONTEXT

This thesis was funded through the SIGNALIFE Labex program, part of the French governmental initiative "ANR - Investments for the Future", focused on the study of biological systems, within an interactive network of regional Life Science Institutes, in Nice, France.

Within this framework, the current project represents the product of a cooperation between the Adhesion signaling and stromal reprogramming in the tumor microenvironment team, from the Biology Institute (IBV) in Nice, and Morpheme

(Computational morphometry and morphodynamics of cellular and supracellular structures), a joint team between INRIA, CNRS and Université Côte d'Azur.

1.2.1 *Brief introduction of biological aspects*

The Extracellular Matrix (ECM) is a cell-produced micro-environment, responsible for providing structural support for cells to undergo several functions related to their survival, proliferation, motility, etc. Fibronectin (FN), a major constituent, is a large protein which serves as a template upon which other components are attached to form a functional ECM. It exists in two forms: plasma and cellular FN. Cellular FN (cFN) is organized in fibrillar networks and can be assembled differently in the presence or absence of two extra protein domains in the molecule, named EDB and EDA.

Cellular FN is a major protein of the extracellular matrix that can take multiple forms in the presence of the extrodomains. The role of these extra domains has yet to be fully deciphered.

So far, a comparison among the FN variants (obtained through the inclusion/exclusion of EDA/EDB) has been difficult to achieve. However, a better understanding of the impact of EDA/EDB on the FN structure and functions, is essential to fully describe the role of the FN in several processes linked to tissue repair, fibrosis and tumor progression.

To generate the different FN forms, an experimental procedure was set up, leading to the production of four recombinant FN variants (B-A, B+A-, B-A+, B+A+). The resulting fibrillar structures were subsequently imaged with a confocal microscope in a 2D acquisition.

Additionally, the variants were presented to cells in presence of a growth factor that simulates an activated "tumor-like" state, resulting in the production of a disease-like, fibrotic ECM. Understanding how the FN is assembled in a pathological condition is critical for diagnostic and clinical applications.

1.2.2 *Objective and motivations*

The motivation behind this thesis, as introduced before, was provided by a biological problem focused on the study of the different types of ECM assembly within the tumor micro-environment (head and neck cancer). More specifically, the object of interest is FN, a major constituent of the tumor ECM, and of its four different forms that can be analyzed from a database of 2D confocal microscopy images. Assessing how FN is assembled is crucial for understanding the structure-function relationships operating in the tumor ECM and exploiting them for diagnostic and therapeutic purposes.

Therefore, the focus of the thesis is the numerical characterization of the different architectures of the FN networks, based on geometrical features of the fibrillar structures (e.g. fiber length, thickness, orientation, etc.). The proposed set of computational tools need to play a discriminating role to distinguish and significantly compare the FN variants.

Throughout this thesis, we are mainly focused on the characterization of the four "normal-state" forms, to determine whether the matrices are differently organized, and, if so, identify the relevant attributes that help make the distinction. Additionally, we were interested in the comparison of FN architecture between the normal and pathological state conditions.

The first step within this characterization has been the fiber detection with multi-resolution techniques and subsequent classification of the extracted features (e.g. curvelet features). After showing that the four variants can be discriminated by a classification pipeline with a similar performance to that of a trained specialist, we subsequently constructed a graph-based representation of the fibers previously detected with a set of Gabor filters. Graph-specific attributes (geometrical and topological) were computed, submitted to a PCA analysis and then employed for the classification of the variants, proving that the graph representation embeds the most relevant information provided by the confocal images. Additionally, we performed a statistical analysis (based on Gaussian random fields as well as empirical distributions) of the fiber parametric maps, in order to illustrate quantitative and qualitative differences between the normal and tumoral state FN networks.

The next part of the thesis was devoted to taking the first steps towards building a numerical model of the FN variants, starting from the graph representations. Therefore, we were interested in identifying the appropriate methods that can provide a measure of similarity between the graphs, in order to obtain a quantitative and qualitative comparison of the FN variants as well as a differentiation between normal and tumour-like FN fibers.

Methods based on graph matching that provide a metric between graphs whilst comparing their global structure, and alternatively, on optimal transport, were selected for this purpose. In the hope of acquiring a better understanding of how these two approaches can be adapted to compare the FN graph representations, we proposed a preliminary analysis on randomly generated graphs, that showed the capacity of two of those approaches to provide a meaningful distance between the graphs.

Based on the metric provided by the chosen graph-matching framework, we developed two different approaches for defining a prototype graph, representative of a certain FN class. Additionally, since the graph matching serves as a registration tool between the graphs, this enabled the computation of various fiber parameter deformation maps between the variants (after matching). These deformation maps were subsequently analyzed within the same statistical framework previously employed for the analysis of parametric maps, showing here whether the variation of the parameter (e.g. fiber length) can be explained by the variance within the same class or not.

1.3 THESIS ORGANIZATION

1.3.1 *Main contributions*

The main contributions of this thesis can be articulated as follows:

- A classification pipeline based on curvelet features invariant to rotation and DAG-SVM for which results were compared to a trained specialist.
- Graph-based representation of the FN fibers constructed on top of Gabor filters, using morphological-based operations that provide the fiber morphological skeleton and its associated graph. Subsequently, an approach to

reconnect the missing fibers in the skeleton to improve the representation of the fiber graphs, was proposed.

- Local characterization of the fiber features based on their graph-based representation (fiber length, proportion of node degree, median pore size, etc.) and Gabor parameters (fiber width) with PCA analysis.
- Global characterization of the fiber features, e.g. fiber length through a statistical analysis framework based on Gaussian random fields and on the computation of empirical distributions. Application for the comparison of parametric maps for normal and tumoral-like FN.
- Preliminary analysis of graph matching performance between many-to-many assignment framework and optimal transport approach (adapted to a graph matching setting) on randomly generated graphs.
- Methodologies for the computation of the prototype graph using the many-to-many assignment framework: the first approach is based on (edge) majority voting after matching to a common reference graph and the second approach uses a heuristic based on the longest chains of matched nodes.
- Statistical analysis (revisited) of the fiber deformation maps (after graph matching) to study the variation of certain fiber parameters between normal and tumoral-like conditions.

1.3.2 *Manuscript organization*

Part I introduces the main context and objectives of this thesis focused on the study and numerical characterization of the FN variants. Main contributions as well as publications are illustrated here. Part II ([Chapter 2](#)) focuses on the biological context of the project, introducing the protocol for the confocal microscopy 2D image acquisition. In Part III ([Chapter 3](#)) we present the proposed classification pipeline of the FN confocal images, based on curvelet features and deep-learning framework. Part IV describes the methodologies proposed to obtain a local characterization ([Chapter 5](#)) of the fibers, starting from a graph-based representation built on top of Gabor filters ([Chapter 4](#)). Additionally, it presents the statistical framework that provides a global characterization of the fiber features ([Chapter 6](#)).

Part V introduces the set of approaches that were considered in order to take the first steps toward the development of a numerical model of the FN variants, based on graphs. [Chapter 7](#) introduces the state of the art of the graph-matching methods as well as the many-to-many assignment approach for graph comparison. [Chapter 8](#) presents the state of the art of the discrete optimal transport framework, with a focus on the Gromov-Wasserstein discrepancy for similarity matrices comparison. [Chapter 9](#) illustrates the proposed framework for a performance analysis between the many-to-many assignment approach and the optimal transport for randomly generated graphs. [Chapter 10](#) describes two proposed methodologies for defining the prototype for a set of graphs. [Chapter 11](#) revisits the statistical framework presented in ([Chapter 6](#)) for an analysis of the fiber deformation maps obtained after matching of the corresponding graphs.

1.3.3 *List of publications*

Conferences

- Anca-Ioana Grapa, Raphael Meunier, Laure Blanc-Feraud, Georgios Efthymiou, Sébastien Schaub, Agata Radwanska, Ellen Van Obberghen-Schilling, and Xavier Descombes. **Classification of the fibronectin variants with curvelets**. Proceedings - International Symposium on Biomedical Imaging (ISBI). Vol. 2018-April, 2018.
- Anca-Ioana Grapa, Laure Blanc-Feraud, Ellen Van Obberghen-Schilling, and Xavier Descombes. **Optimal Transport vs Many-to-many assignment for Graph Matching**. GRETSI- XXVIIème Colloque francophone de traitement du signal et des images, 2019.

Journals

- Georgios Efthymiou*, Agata Radwanska*, Anca-Ioana Grapa, Stéphanie Beghelli-de la Forest Divonne, Dominique Grall, Xavier Descombes, Laure Blanc-Feraud, Sébastien Schaub, Mallorie Poet, Didier Pisani, Laurent Counillon, Maurice Hattab, and Ellen Van Obberghen-Schilling, **The presence of alternatively spliced extra domains EDB and or EDA domains of cellular fibronectin confers topographically and functionally distinct features** (in submission to Journal of Cell Science).
- Anca-Ioana Grapa, Laure Blanc-Feraud, Georgios Efthymiou, Sébastien Schaub, Agata Radwanska, Ellen Van Obberghen-Schilling, Xavier Descombes. Results of chapters 6, 11 (in preparation for submission to PLOS Computational Biology).

Dissemination (workshops)

- Poster presentation, 3rd Labex SIGNALIFE Meeting - Cell Signaling, May 9-10th, 2017, Le Saint Paul Hôtel, Nice, France.
- Oral presentation, SIGNALIFE Joint PhD Projects Workshop March 19th, 2018, Théâtre Valrose, Campus Valrose, Nice, France.
- Poster presentation, Modelife days (UCA JEDI: Modélisation, Physique et Mathématique du Vivant), June 18-19th, 2018, Le Saint Paul Hôtel, Nice, France.

Part II

BIOLOGICAL BACKGROUND AND IMAGE DATABASE

BIOLOGICAL BACKGROUND

A cell is the basic structural, functional, and biological unit that contains all the necessary information to build an organism, such as the human body. To correctly build the organism, cells generate and sculpt a scaffold-like structure known as the extracellular matrix (ECM), on which they grow, interact, and exert their various functions. A major component of this ECM is fibronectin (FN), a protein indispensable for normal development, that regulates cell adhesion, motility, proliferation, and differentiation, by forming fibrillar networks that provide lattices for the assembly of a complex ECM.

This chapter begins with brief introduction to the ECM, before focusing on different forms of FN networks, in health and disease. To study the different networks of FN, tissue-like ECMs were generated in a cell biology laboratory, and a series of confocal microscopy images was acquired, as described below. This image dataset was used for characterization of the FN architecture with a longterm goal of generating a prediction model of clinical significance. A representative set of images showing different FN networks is shown at the end of this chapter.

2.1 EXTRACELLULAR MATRIX AND COMPONENTS

The ECM is a complex and dynamic network of macromolecules that surrounds cells in tissues. It represents a scaffold that provides structural and mechanical support, and mediates diverse biological processes that are crucial for tissue formation and function [MK13]. Over the years, it has been shown that the ECM plays vital roles in the behaviour of the residing cells in terms of signaling, motility, proliferation, survival, and function [Hyn09]. Furthermore, it has been shown that ECM architecture undergoes significant remodelling in pathological conditions (cancer, fibrosis, etc) that in turn influences the behaviour of surrounding cells. Hence, there is an increased interest in the study of interactions between cells and the ECM in development, tissue homeostasis and disease.

Structurally, the ECM is composed of various types of molecules (Figure 2.1), namely proteins, carbohydrates, and collagens, as well as molecules with functional roles rather than structural ones. The focus of this manuscript is the numerical characterization of the networks of FN, a major glycoprotein (protein decorated with carbohydrates) of the ECM, which serves as a template upon which collagen and other components are attached and polymerized in order to form a mature and functional ECM.

Extracellular matrix is the scaffold that provides the mechanical support and facilitates cellular signalling, motion and proliferation.

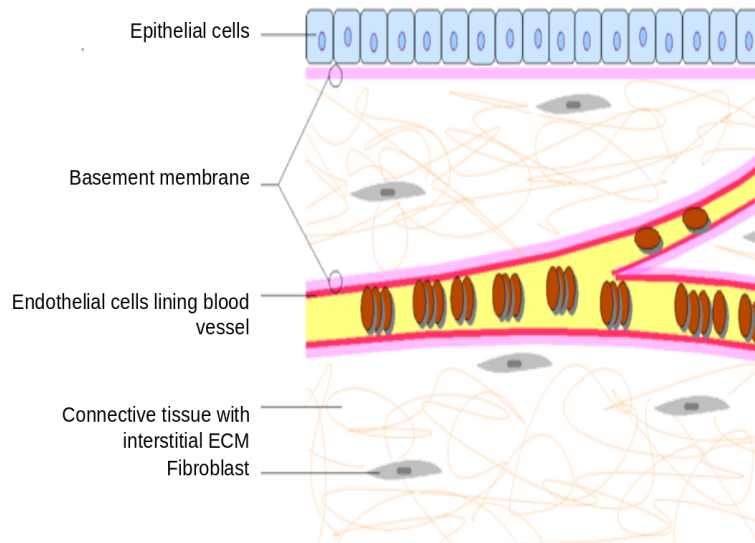


Figure 2.1: Schematic representation of the extracellular matrix in relation to epithelial cells and vascular endothelial cells. The basement membrane, a specialized ECM rich in laminin, non-fibrillary collagen and proteoglycans, separates epithelial cells and endothelial cells from underlying connective tissue. The connective tissue ECM is rich in fibrillary collagen, fibronectin, other glycoproteins and proteoglycans. Fibroblasts are the major producers of the interstitial ECM.

2.1.1 Fibronectin - Major Component of the ECM

Fibronectin is a glycoprotein found in the extracellular space, surrounding fibroblasts, major ECM-producing cells of connective tissue (Figure 2.1). Two types of FN exist:

1. Plasma FN (pFN) is secreted by the liver, in a soluble form circulating through the blood stream.
2. Cellular FN (cFN) is produced by fibroblasts. cFN is mainly found in the form of an insoluble mesh around the cells secreting it, where it is assembled into fibrils and fibers to form the ECM and its production is strictly regulated.

FN is assembled by fibroblasts¹ into a fibrillar network with a complex structure and functions. It serves as a template for the deposition of other matrix components, and possesses binding sites necessary for cell function regulation for cells and growth factors that are, both in physiological and pathological conditions. Moreover, the importance of FN in ECM generation has been underlined in several studies. In an *in vivo* setting (e.g. mouse models), FN plays a vital role in embryonic development [Dar+90], and it participates in early stages of wound healing [CM06]. In *in vitro* experimental approaches ("test-tube experiments"), removal of the FN gene from fibroblasts, results in complete abolishment of ECM formation [Cse+10].

FN is a macromolecule involved in many cellular processes, including tissue repair, embryogenesis, blood clotting, cell migration, adhesion.

¹ A fibroblast is a type of mammalian cell that is able to produce and assemble FN, as well as remodel a pre-existing FN network. Fibroblasts in connective tissue support a wide range of functions: tissue homeostasis, wound healing, pathological conditions.

FN is a high-molecular-weight dimeric protein composed of two similar subunits derived from a single gene. The linear structure of a FN subunit (Figure 2.2) is characterized by a highly repetitive, modular structure composed of three distinct domains termed FN type I, FN type II, and FN type III repeats. Both types of FN (pFN and cFN) have a similar but not identical linear structure. The difference lies in the presence in cFN of two FN type III repeats with similar but not identical amino acid sequences, termed Extra Domain B (EDB) and Extra Domain A (EDA).

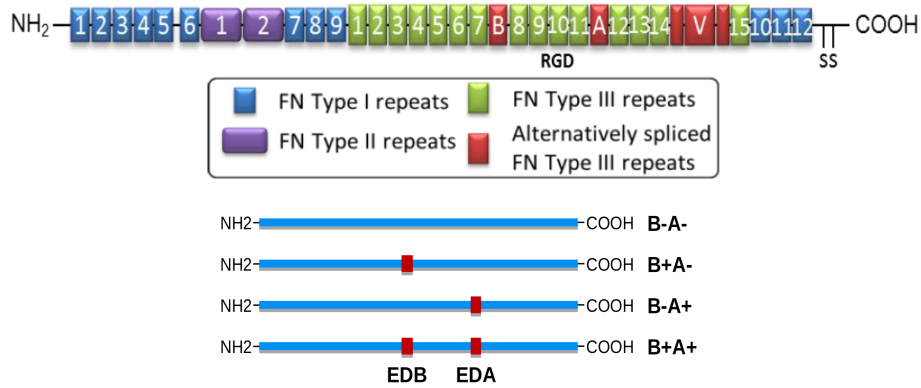


Figure 2.2: Linear structure of Fibronectin. Adapted from [OS+11].

The presence of either (or both) of these Extra Domains gives rise to different forms of FN, collectively termed cFN. Four distinct Extra Domain-specific FN variants exist, namely FN B-A-, FN B+A-, FN B-A+, and FN B+A+. Since the Extra Domains are only present in cFN, and not pFN, the FN B-A- variant is the equivalent of pFN, while FN B+A-, FN B-A+, and FN B+A+ are by definition cFN variants.

A role of these variants in development was suggested when it was shown that the Extra Domains were present in embryonic tissues but absent in adult organisms. Their importance was highlighted when experimental deletion of both Extra Domains in a mouse model, resulted in early embryonic death [ELGH93]. Later it was demonstrated that FN containing EDB and EDA reappeared in the adult organism in some pathophysiological situations such as wound healing and cancer [Ven+10; al99]. Thus, the term *oncofetal FN* was used to describe these isoforms reflecting the biological context in which they are found.

So far, the precise functional properties of EDB and EDA have yet to be fully understood. Several lines of evidence have suggested roles for cFN in cell adhesion, migration and differentiation [Cse+10]. Presence of the Extra Domains have recently been found to enhance FN assembly and to differentially affect fiber organization by cultured cells [Eft+]. cFN variants have also been reported to regulate inflammatory responses, angiogenesis, and tumor progression [Ven+10]. Indeed, FN containing EDB and/or EDA is highly upregulated in several tumor types, including head and neck cancer, which is of particular interest in the context of my thesis project [Gop+17].

ECM architecture is severely altered in the stroma of tumor tissue. The assembly of tumor ECM is largely carried out by carcinoma-activated fibroblasts. Patterns of fibrillar collagen organization called *Tumor Associated Collagen Signatures* (TACS

The inclusion of the two alternatively spliced extra regions EDB/EDA gives rise to different FN variants bound to have different properties.

1-3), have been defined to describe the structural changes in fiber arrangement that accompany carcinoma progression. TACS 1 corresponds to curly/anisotropic fibrils in normal tissue, whereas TACS 3 (advanced stage cancer) is characterized by linearized and aligned collagen fibers oriented perpendicular to the tumor boundary. In breast cancer, TACS 3 was found to be an independent prognostic indicator of poor survival in breast cancer [Con+11]. As mentioned above, collagen type I binds to FN and collagen type I deposition is primarily dependent on previously assembled FN. Therefore, assessing how FN is assembled, which factors regulate its assembly is crucial for understanding the structure-function relationships operating in the tumor ECM and exploiting them for diagnostic and therapeutic purposes.

Assessing how FN is assembled is crucial for understanding the structure-function relationships operating in the tumor ECM and exploiting them for diagnostic and therapeutic purposes.

2.1.2 Experimental procedure for FN matrix preparation

Since cFN is a major component of the ECM, an important goal is to understand whether and how the ECM in development and disease, is differently organized when the extra domains EDA/EDB are present in the molecule (together or separately), compared to when they are absent. To achieve this, a set of biological tools comprised of full-length human FN proteins was generated, containing either one, both or none of the Extra Domains, resulting in four different FN variants.

With this toolset, we set out to elucidate how the presence of alternatively spliced (EDA and EDB) domains affects the fibrillar assembly of FN at the surface of assembly-competent fibroblasts. The steps (Figure 2.3) that were followed experimentally, in the laboratory, to produce the variants are illustrated below:

1. FN-null fibroblasts were placed at a high density in tissue culture plates.
2. The next day, FN variants (FN B-A-, FN B+A+, FN B+A-, FN B-A+) were added to the culture medium.
3. Cells were placed in an incubator for eight days in order for the matrix to be generated.
4. In the end of the experiment, cells were removed from the tissue culture plates. Matrices were fixed with chemicals and stained to allow the visualization of FN variants with a confocal fluorescent microscope. Staining was performed using an immune-based approach in which a fluorescent secondary antibody is bound to a primary antibody that binds directly to FN fibrils.
5. Fluorescently-labeled FN in matrices were visualized using a confocal microscope.

This experimental approach is designed for the study of FN networks in normal tissue. However, the addition of tumor-/fibrosis-promoting growth factors to the cell cultures during matrix generation leads to the assembly of an ECM that resembles that observed in cancer or fibrotic tissue. To that end, we used a soluble growth factor, namely Transforming Growth Factor β 1 (TGF- β 1) to induce a tumor-like cellular state. In this way, we could examine not only the differences among FN variant-specific matrices (normal-state), but also the differences between normal and tumor-like matrices.

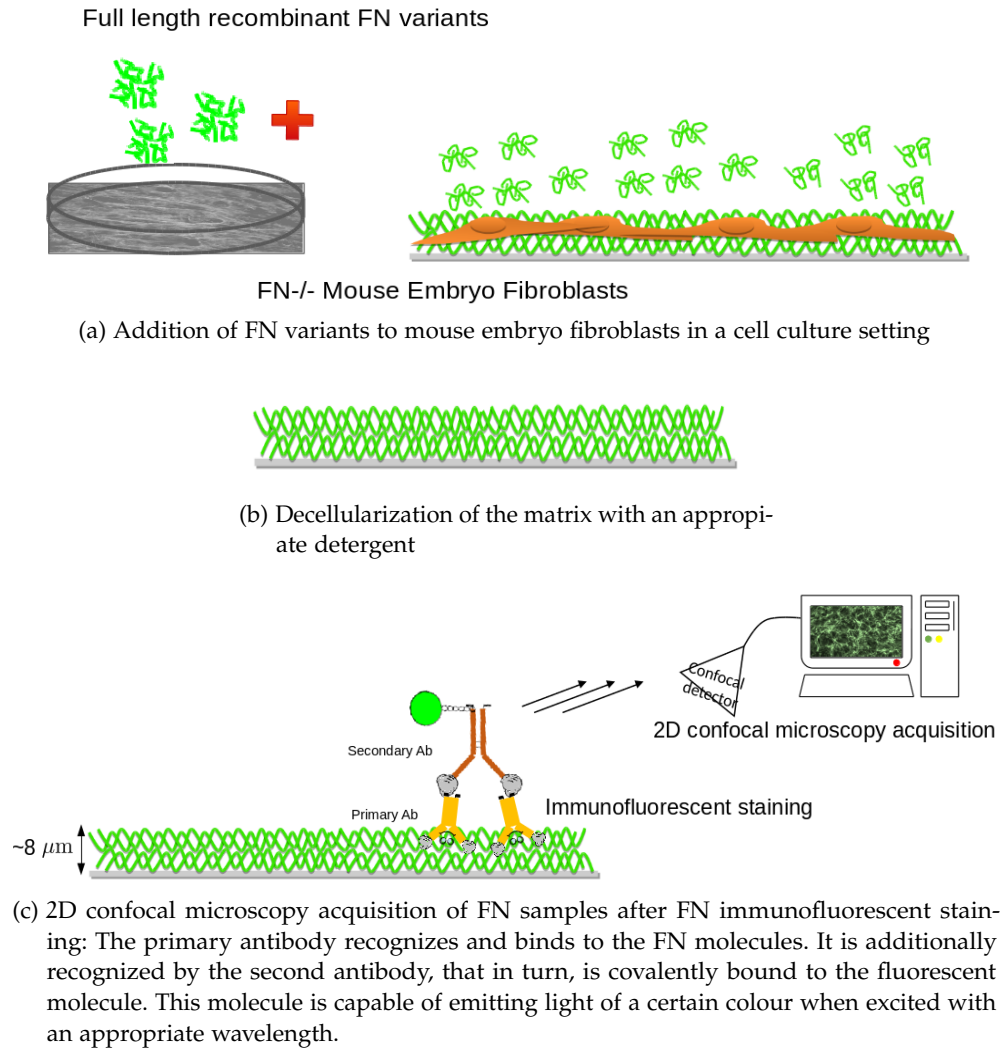


Figure 2.3: Protocol for the production and 2D confocal image acquisition of FN variants

2.2 CONFOCAL IMAGE DATABASE

A microscope generates a magnified image of an object above the human eye resolution ability [Mic]. There are several types of microscope technology based on the information carrier (electron microscopy, photon or optical microscopy), employed in various fields of science and medicine to study objects in great detail. In cell biology, a frequently used microscopy technique is fluorescence microscopy, the most popular optical technique in biology, that allows the visualization of a specific molecule in cells by staining them with fluorescent dye. The fluorescence combined with confocal microscopy, yields precise 3D imaging of the molecule of interest in the sample.

In the following sections, we introduce the basic principles of microscopy and subsequently define the confocal microscope which was used for the acquisition of FN image samples.

2.2.1 Fluorescence principle

Fluorescent staining combined with an appropriate imaging instrument, is widely used in cell biology for a variety of experimental applications.

Traditional fluorescence microscopy relies on a physical process in which special types of molecules called fluorophores, fluorochromes, or fluorescent dyes, when illuminated with light (photons) of an appropriate wavelength, will absorb it and emit light of a different wavelength (fluorescence). The basic principle is illustrated in [Figure 2.4](#). When fluorescent molecules are excited with photons, their energy increases to a higher (less stable) level. Typically the molecule dissipates partially the absorbed energy in thermal energy, and may subsequently lose the remaining energy difference by emitting light of a longer wavelength. Finally, the emitted light is captured by a detector and an image is created.

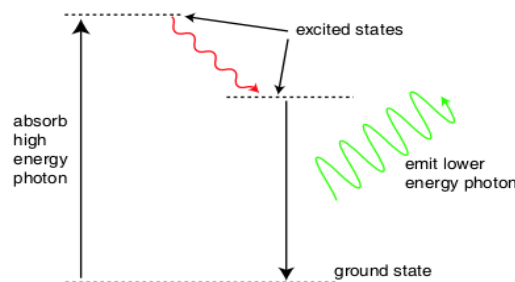


Figure 2.4: Fluorescence principle. Figure reproduced from [SW].

Except for cell expressing fluorescent proteins, the fluorophores are commonly introduced artificially in the biological structures of interest. There are several methods to accomplish this. Immunofluorescence [Imm] is such an approach that relies on the characteristic of antibodies to bind specifically a protein of interest, in order to add fluorescent dyes to the desired molecular target within the stained specimen tissue.

2.2.2 Confocal microscopy

Confocal microscopy [Min] is nowadays, the most favourite technique in biology, providing sharper images. Due to its ability of optical sectioning, it removes the out of focus light, compared to widefield microscopy. In conventional widefield microscopes, the entire specimen is flooded evenly with light from a light source ([Figure 2.5](#)) acquiring simultaneously all pixels, while the confocal microscope scans the sample pixel per pixel.

The modern confocal microscope consists of an optical microscope and an integrated electronic system (similarly to any modern microscope), composed of one or more electronic detectors, a computer (for image display, processing, output, and storage), and a system of laser lines coupled to wavelength selection devices and a beam scanning assembly.

Focusing the excitation light (using a laser to get enough power) on the sample on the smallest volume (limited by optical resolution), the emitted light is then collected through the pinhole which blocks the out of focus light ([Figure 2.6](#)). The

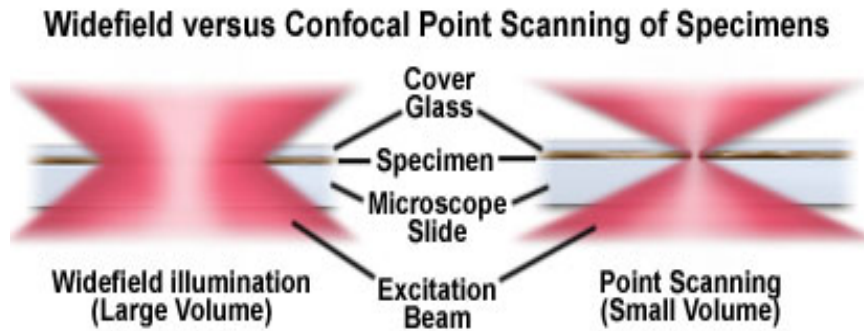


Figure 2.5: For traditional widefield microscopes, all planes contribute to the image formation, while in a confocal system, only the focal plane contributes to image formation. Adapted from [Con].

setup then will scan all the sample using the excitation light, moved across the specimen, by the oscillating mirrors.

As already mentioned, one of the most common techniques used by a confocal microscope to produce images is by excitation of fluorescent dyes (fluorophores) in the specimen. The secondary fluorescence emitted from the specimen passing through the dichromatic mirror, is focused as a confocal point at the detector pinhole aperture. There is a small part of the out-of-focus (above/below the focal plane) fluorescence emission that reaches the pinhole aperture, hence it does not contribute to the resulting image.

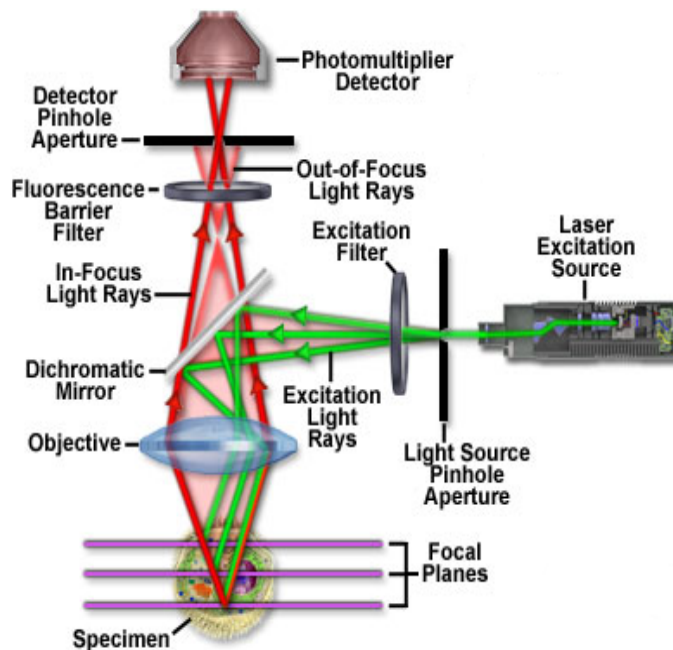


Figure 2.6: Basic setup of a confocal microscope. Light from the laser is scanned across the specimen by the scanning mirrors. Optical sectioning occurs as the light passes through a pinhole on its way to the detector. Adapted from [Con].

For the present analysis, two dimensional (2D) confocal images (3128×3128 pixels) of the four FN variant-specific matrices were acquired both for the normal and tumor-like state. Acquisition was performed with a confocal microscope LSM 710 System by Zeiss, with a 10x/0.45 M27 objective lens. Fluorophore excitation was done with an Argon laser at 488 nm. Pixel size was set to $0.27\mu\text{m}$.

Due to computational constraints, the approaches developed within this manuscript were applied to regions of 512×512 pixels, as the information contained within this area was shown to be representative for fiber characterization in all of the four variants. Naturally, the pipelines can be applied to larger sized images to incorporate more information for future analysis. We recall that a central question of our work is to determine whether the confocal images can provide discriminant features for the FN variants.

2.3 CONCLUSIONS

Fibronectin is a major matrix protein that provides a template upon which other matrix components attach to form the ECM. Expression of cFN harboring EDB and/or EDA is upregulated in development and disease, yet the specific roles of these domains are yet to be completely understood. cFN assembly by fibroblasts was found to be differently assembled depending upon the presence of the extra domains. Thus, an experimental procedure was designed to analyze these differences. To this end, a database of 2D confocal microscopy images was subsequently generated for both normal and tumor-like conditions (Figure 2.7, 2.8, 2.9, 2.10).

Our overall objective is the numerical characterization of topological/geometrical features of ECM landscapes of healthy and disease (fibrotic/tumoral) tissue. The goal of my work was thus the development of algorithms that distinguish variant-specific networks while extracting biologically relevant information to provide a quantitative/qualitative analysis of the fibrillar FN parameters.

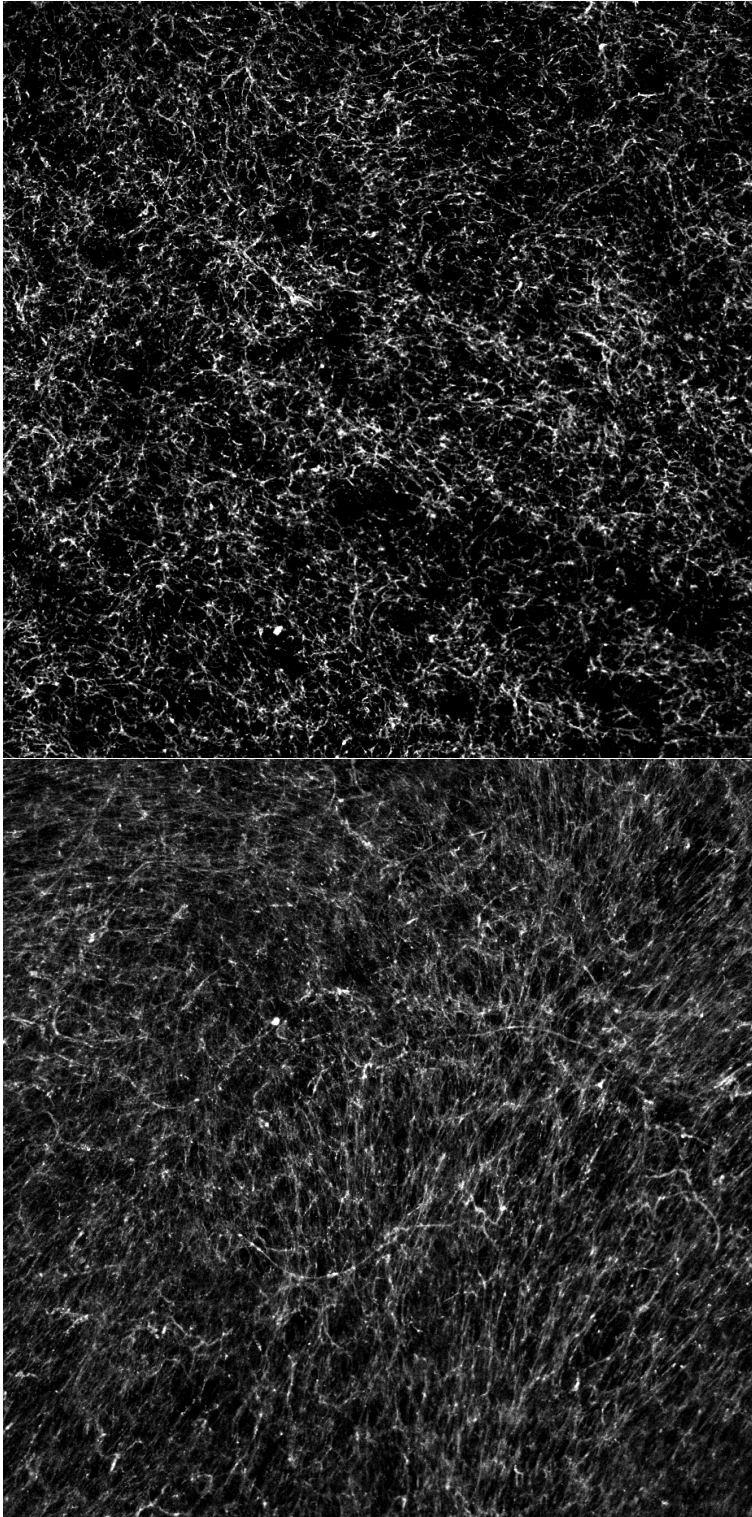


Figure 2.7: FN confocal images - Normal ECM (top) and "Tumour-like" ECM (bottom) FN B-A-. Image size : 3128×3128 pixels. Pixel size is $0.27\mu\text{m}$.

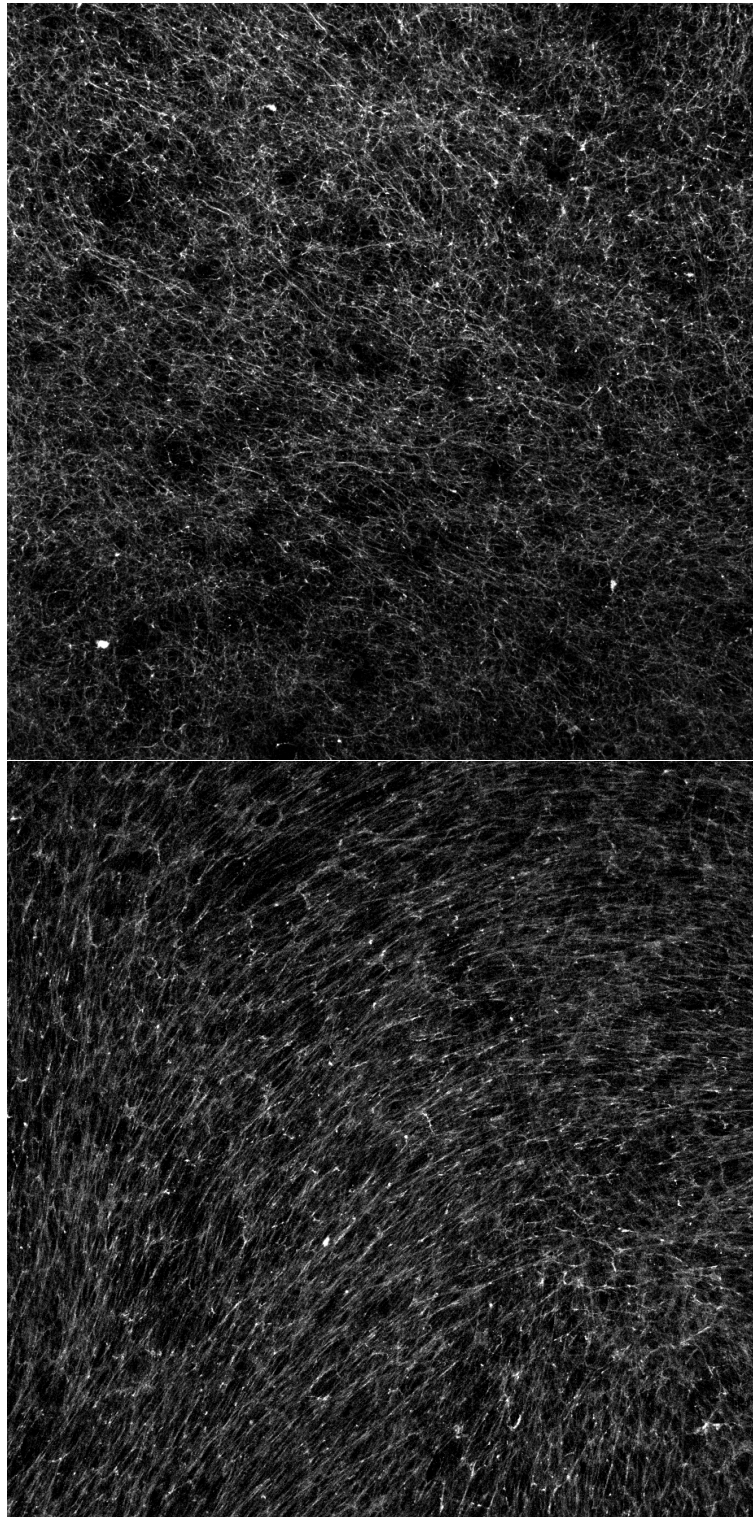


Figure 2.8: FN confocal images - Normal ECM (top) and "Tumour-like" ECM (bottom) FN B+A-. Image size : 3128×3128 pixels. Pixel size is $0.27\mu\text{m}$.

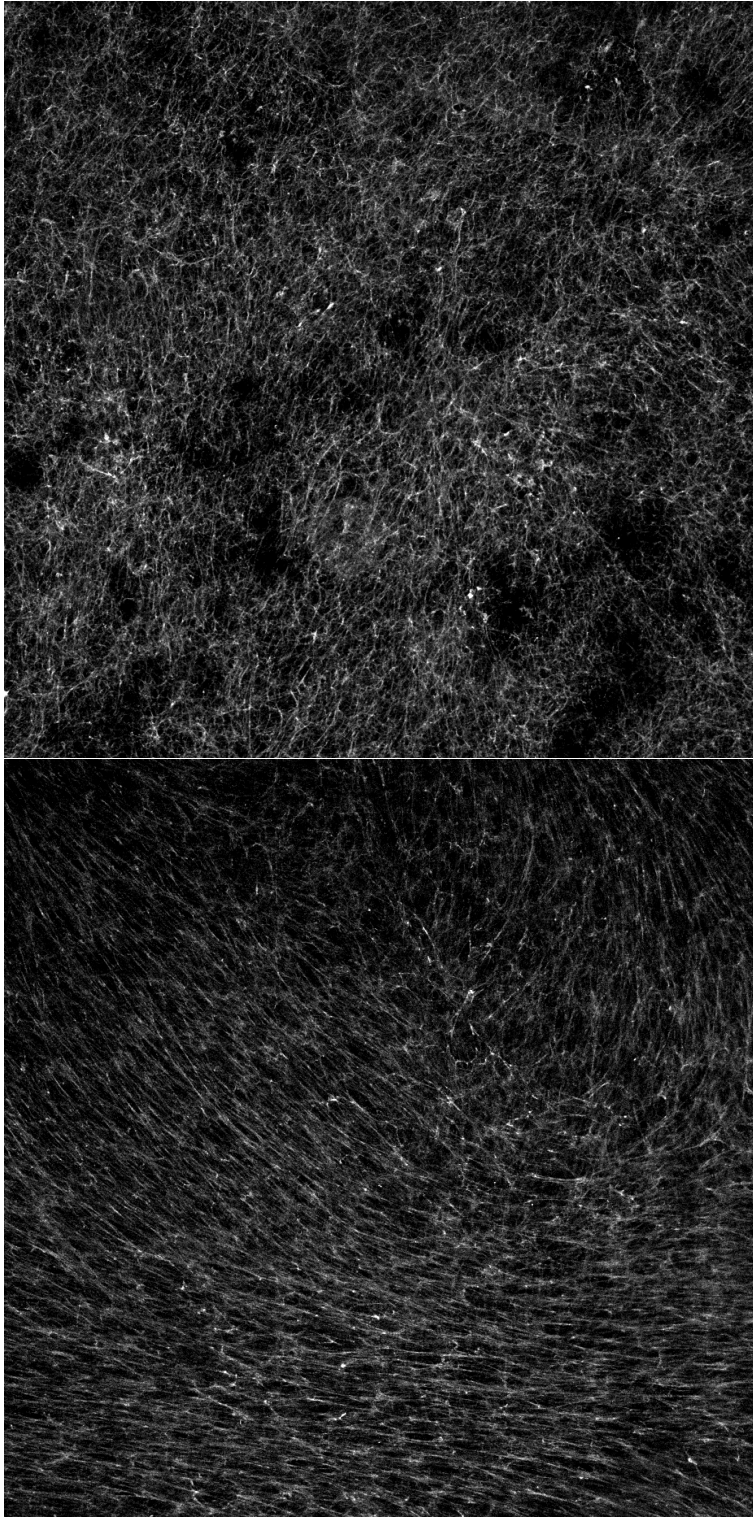


Figure 2.9: FN confocal images - Normal ECM (top) and "Tumour-like" ECM (bottom) FN B-A+. Image size : 3128×3128 pixels. Pixel size is $0.27\mu\text{m}$.

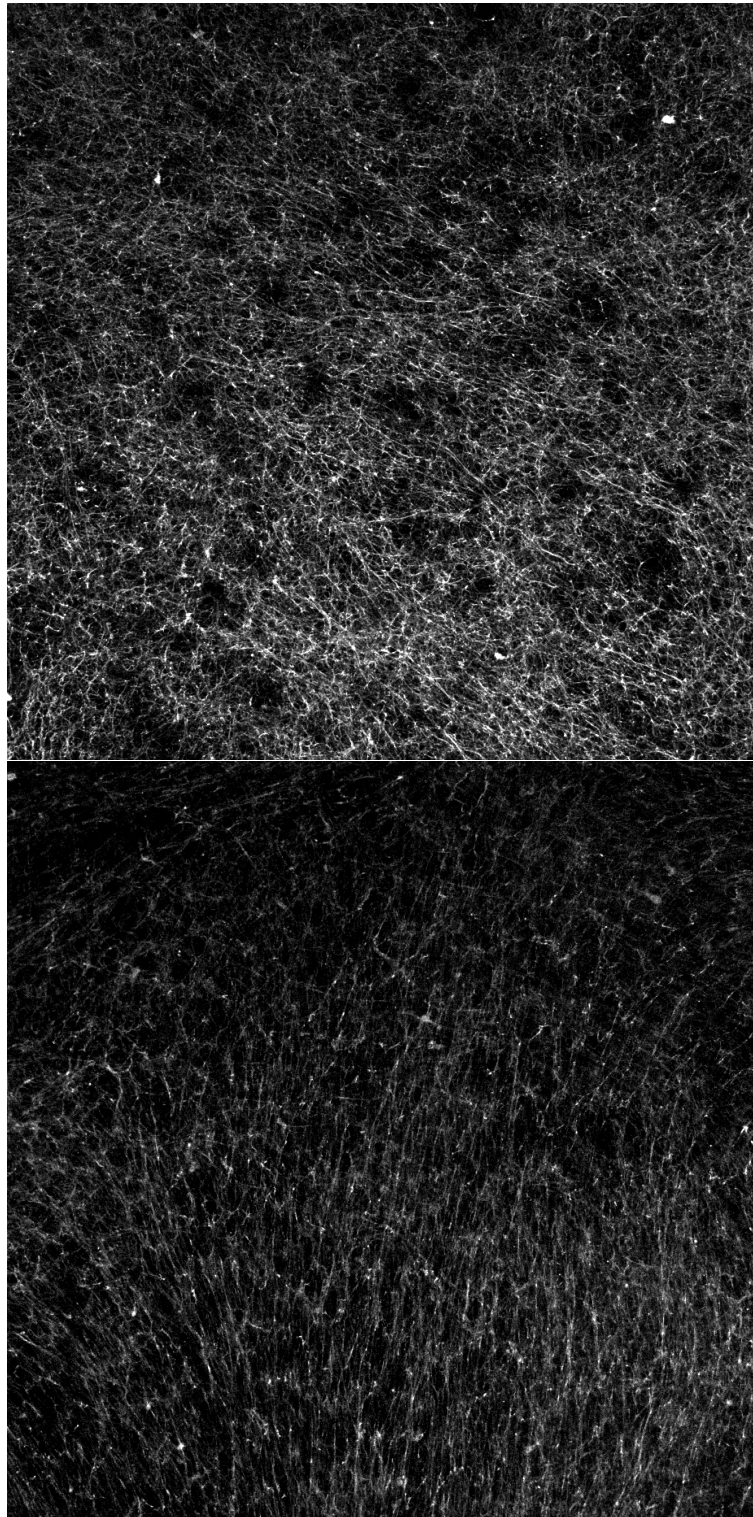


Figure 2.10: FN confocal images - Normal ECM (top) and "Tumour-like" ECM (bottom) FN B+A+. Image size : 3128×3128 pixels. Pixel size is $0.27\mu\text{m}$.

Part III

FIBRONECTIN VARIANTS CLASSIFICATION FROM
CONFOCAL IMAGES

CLASSIFICATION OF THE FN VARIANTS FROM CONFOCAL IMAGES

The mechanics and behaviour of the FN fiber networks is thought to have an impact on the surrounding cell structure and dynamics. Through confocal microscopy, the FN-specific fibers corresponding to the four different conformations, were acquired at $0.27\mu\text{m}/\text{pixel}$. A qualitative analysis of the confocal images allows a trained specialist to detect differences in the fiber architecture for each variant.

An essential step in the methodology that provides the fiber numerical representation is the fiber detection. There are several image processing multi-resolution analysis methods which are able to identify directional, anisotropic structures, occurring at different resolutions. The extracted features, provided by the aforementioned techniques, can subsequently be employed in discriminative models, capable of measuring the extent of variability of certain parameters among and within variants.

Across this chapter, we present a standard techniques that enables the detection of anisotropic, oriented structures, namely *discrete curvelets*. We show the advantages and inconvenients of using curvelets for feature extraction in a classification and modelling context. Moreover, we present a classification pipeline based on curvelets and SVM-type classifier to prove that curvelet features contain relevant information to differentiate the FN variants (in normal state). Finally, we classify the confocal images using a deep-learning approach (with a pretrained model network), to verify whether the variants can be distinguished based on the information contained in the confocal images.

3.1 FIBER DETECTION USING DISCRETE CURVELETS

Generally, the feature extraction is performed as a transformation of the images into a characteristic space defined, for instance, (but not limited to) by "dictionary" elements (atoms) with certain attributes. Within this space, the detected structures will thus be characterized according to the specific attributes. Among the various techniques, multi-resolution analysis provides the framework for various geometrical feature detection in spatial and frequency domain. For the work presented in this manuscript, we selected two of these methods that we have applied for FN fiber detection and enhancement, the discrete curvelets [Can+05] and Gabor filters [PK97]. In this chapter, we focus mainly on the first method, namely, discrete curvelets.

3.1.1 Fast Discrete Curvelet transform

In the field of image processing, multiscale/multiresolution tools have been extensively used for anisotropic feature extraction (points, lines, edges) and detection, compression etc.

Among existing methods, the wavelets form a basis (dictionary) of isotropic elements occurring at all scales and locations but represented by a fixed number of directional elements [FS09]. Wavelets can mostly detect features such as point singularities, but do not constitute a well-suited techniques to represent curvilinear features (e.g. such as those that describe the FN fibers).

The curvelet transform is a family of frames that is constructed to better identify the anisotropy and curvilinear characteristic of certain structures of interest that occur at various orientations. Among the different curvelet theoretical frameworks, the second generation of curvelets is a multiscale pyramid allowing the representation of a certain number of possible directions at multiple scales [FS09]. For the analysis of the confocal images of FN fibers, we have chosen to work with the implementation of the fast discrete curvelet transform found in the Curvelet Toolbox [Can+05].

The output of the linear transform is a collection of coefficients $c_{j,l,k}$ evaluated in Fourier domain (real-valued), indexed by discrete-valued scale j , orientation l and location k . These coefficients can, in turn, be used to perform an analysis of anisotropic objects (e.g. different parameter statistics, classification, etc.). An important aspect resulting from the construction of the curvelets, is that a finer scale is associated with a higher number of possible orientations. This property allows for highly anisotropic elements to be represented at a fine scale. Another specificity is given by the scaling property: at scale 2^{-j} the length and the width of the envelope of a specific curvelet obeys a consistent scaling (the length is approximately $2^{-j/2}$ and the width 2^{-j}).

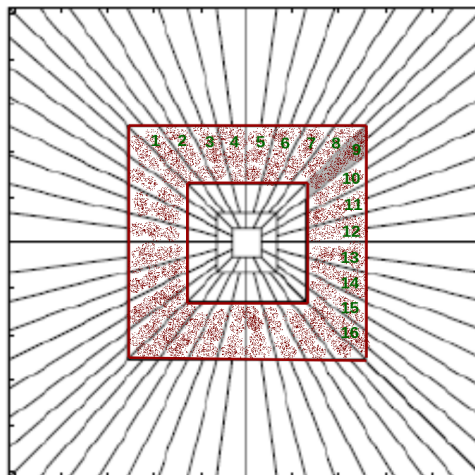


Figure 3.1: Frequency tiling of discrete curvelets of a 256×256 image (grid): frequency representation of the curvelet wedges (sectors) corresponding to different orientations organized at different scales starting from the lowest scales (in the center), up to $(N - 3)$ -th scale, where $N = \log_2(256)$. In this example, the wedges drawn in red colour belong to the 4th scale, with 32 possible orientations.

Formally, a curvelet can be represented as the product of a radial dyadic frequen- tial window (bandpass filter) and an angular window in the frequency domain, in a polar coordinate system. This representation provides the means to obtain a directional analysis at different scales.

In practice, the window functions are built on trapezoids (as an adaptation of the polar wedges to Cartesian arrays), so the frequency tiling is based on shears (Figure 3.1). There are multiple options for the construction of Cartesian arrays, instead of polar tiling, in the frequency plane. We chose the wrapping method implemented in Curvelet Toolbox [Can+05] for its simplicity in the handling of the discretization grid and for the fast computation algorithm of an otherwise redundant transformation.

If we denote by $\phi_{j,0,0}$ the basic "mother curvelet" at scale 2^{-j} , $j > 0$ the family of curvelets constructed via wrapping method ([Can+05]) at arbitrary angles (slopes: $\tan \theta_{j,l} = \lfloor 2^{-l/2} \rfloor + 1, \dots, \lfloor 2^{l/2} \rfloor - 1$), is obtained by shearing and translation of this basic element on a discrete Cartesian grid, $b \approx (k_1 2^{-j}, k_2 2^{\lfloor -j/2 \rfloor})$, where $k = (k_1, k_2) \in \mathbb{Z}^2$ [Can+05]:

$$\phi_{j,l,k}(x) = 2^{3j/4} \phi_{j,0,0}(S_{\theta_l}^T(x - b)) \tag{3.1}$$

where $S_{\theta} = \begin{pmatrix} 1 & 0 \\ -\tan(\theta) & 1 \end{pmatrix}$ is the shear matrix. The family of curvelets is completed by symmetry and rotation with $\pm\pi/2$; the coarse curvelet elements for low frequen- cies are non-directional. For more details regarding the discrete implementation methodology, the reader is referred to [MP11; Can+05].

For the modelling perspective, an advantage of the discrete curvelet transform is given by the fact that it implements a tight frame ¹, meaning that every function $f \in L^2(\mathbb{R}^2)$ can be represented [Can+05] as shown in Equation 3.2. An important property of being a tight frame is that it can recover f from the coefficients $c_{j,l,k}$ the L^2 sense, (perfect reconstruction property of a frame).

$$f = \sum_{j,l,k} c_{j,l,k} \phi_{j,l,k} \tag{3.2}$$

where $\phi_{j,l,k}$ is the discrete curvelet waveform and the curvelet coefficients are $c_{j,l,k} = \langle f, \phi_{j,l,k} \rangle$. The Parseval identity then holds:

$$\sum_{j,l,k} c_{j,l,k}^2 = \|f\|_{L^2(\mathbb{R}^2)}^2, \quad \forall f \in L^2(\mathbb{R}^2) \tag{3.3}$$

In Figure 3.1, the frequency representation of the decomposition of a 256×256 image, is illustrated as a tiling of wedges (trapezoids) - curvelet support in the frequency plane. The geometric pyramid structure is divided in dyadic scales, from coarsest centre of the representation (with non-directional elements) to the finest scale (higher frequencies), each with a different number of orientations. Usually for this type of representation, the angular resolution commonly doubles when

¹ A tight frame is different to an orthogonal basis in that it does not need to be linearly independent, (such as the orthogonal basis). This means that a frame has some redundancy, the positive impact being an increased selectivity in orientation, among others [Fuj19].

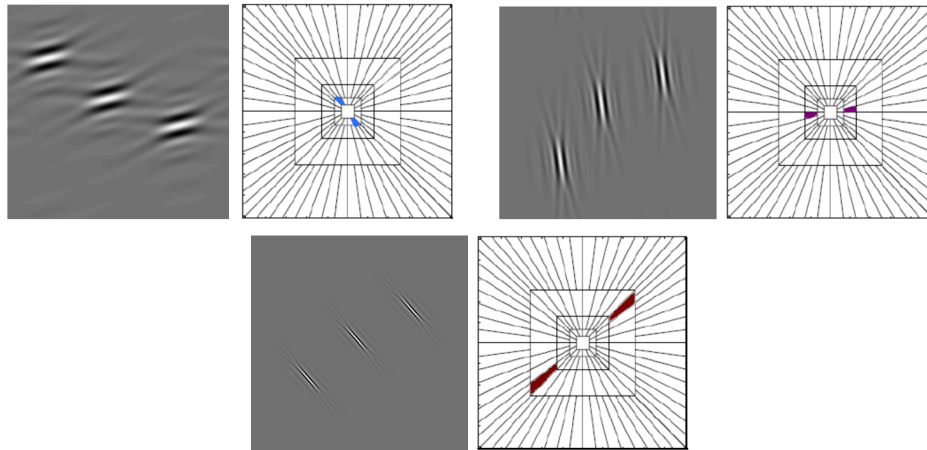


Figure 3.2: Curvelets in spatial (left side) and frequency domain (corresponding right side) at various scales and orientations: top-left, 3 curvelets at scale 2, top-right, 3 curvelets at scale 3, bottom-centre, 3 curvelets at scale 4.

passing at a finer scale. **Figure 3.2** displays several examples of curvelets in spatial-frequency domain at different scales with various orientations, corresponding to the decomposition of a 256×256 image.

3.1.2 Fiber detection using curvelets for classification

Each image of size $N \times N$ is decomposed into $(\log_2(N) - 3)$ dyadic scales and the number of angular sectors for each scale differs according to the following example: for $N = 512$, the curvelet transform returns 6 scales with 1, 16, 32, 32, 64, 64 possible orientations from coarse to fine scales respectively.

Figure 3.3 illustrates the curvelet coefficients amplitude matrices for 3 levels of decomposition corresponding to coarsest scale 1, 2, and finest scale 6. The multiple matrices at each level belong to different orientations. A certain fiber will be reconstituted by a linear combination of curvelet coefficients at different scales and orientations.

Related studies consider different statistical properties of the curvelet coefficients for texture characterization, such as energy [SDL08], entropy or curvelet subband distribution [GR11; IZ09]. As our main interest is to perform geometrical modeling of the fibers rather than a pure assessment of their discriminative power, instead of computing average statistical features to facilitate the classification, we have worked with the coefficients themselves.

However, we chose to reduce the vast number of coefficients taken into account for the classification, and thus keep the most significant ones. To do so, we selected the largest curvelet coefficients that contain a suitable percentage of the total energy. A percentage of 85% seems to be a good compromise between the speed of the training and classification algorithm and the fidelity of image reconstruction, as illustrated in **Figure 3.4**. Finally, the coefficients that belong to the finest scale, susceptible of capturing the eventual acquisition noise present in the images, were not taken into account.

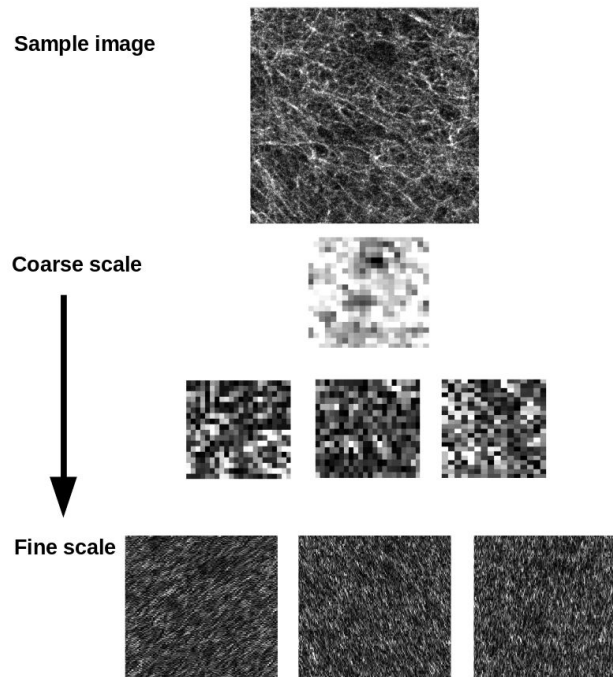


Figure 3.3: Curvelet scale decomposition of a sample image of 512×512 pixels, at 3 scales, from coarse to fine: 1, 2, 6.

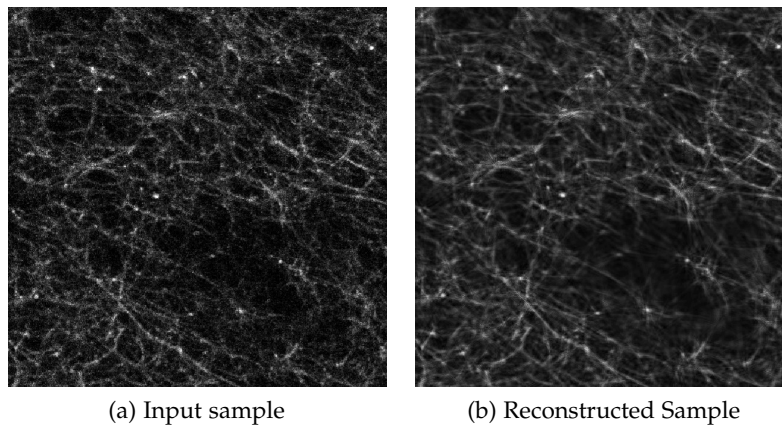


Figure 3.4: Reconstruction of a sample image after keeping 85% of total curvelet coefficients energy.

3.1.2.1 Invariance to rotation of curvelet coefficients

The curvelets that are described above are not invariant to rotation. In a discriminating context, this aspect can be quite problematic, as it can impact the accuracy of the classification. What is important in the FN images is the presence of multiple dominant orientations, relative to each other. During the acquisition process, the

samples of the ECM may also be differently oriented. Hence we needed to ensure that the images follow the same main privileged direction.

To do so, we estimated the dominant orientation of the fibers and rotated every image according to its own dominant orientation. Since this information is hidden in the energy distribution over the subbands, we opted for an estimation of the dominant orientation using the gradient vector of the images.

For a function $f \in \mathbb{R}^2$, we consider its gradient vector $\nabla f = (f_x, f_y)$ with magnitude defined by $|\nabla f| = \sqrt{f_x^2 + f_y^2}$ and orientation $\theta = \arctan(\frac{f_y}{f_x})$. We can now estimate the dominant orientation Θ as:

$$\Theta = \frac{\sum_i |\nabla f_i|^2 \theta_i}{\sum_i |\nabla f_i|^2} \quad (3.4)$$

where $|\nabla f_i|$ is the magnitude and θ_i is the orientation of the image gradient at pixel i .

Subsequently, the images were aligned to the same direction, after performing a rotation by interpolation with the corresponding Θ . In order to validate the assumption that curvelets can provide a suitable model for the characterization of FN fibers, we first needed to show their ability to describe the fiber geometry in terms of physical characteristics (e.g. scale, orientation, location). In addition to that, we were interested in determining the discriminating capacity of the curvelet features (i.e. ability to discriminate among the different FN variants). Therefore a bag of features model [Csu+], adapted to our data, was developed in order to analyze the classification results of the four FN variants, as detailed below.

3.2 FN FIBER CLASSIFICATION USING CURVELETS

3.2.1 Bag-of-words and image signatures

The curvelet features that describe the fibers are the collection of coefficients $c_{j,l,a}$ with scale j , orientation l and magnitude a . We performed a K-means clustering of the curvelet coefficients after the curvelet decomposition of the image database, referred to as the training dataset. In order to determine an appropriate number of clusters, we used a heuristic elbow method [KM13] and found $K = 400$ number of total clusters.

The normalized feature histogram was computed as the rate of the number of the curvelet coefficients of an image, assigned to each cluster, as shown in Figure 3.5. Also referred as image signature, it is stored as a K-dimension vector of real-positive values. The image signature constitutes the input data for the chosen classifier.

The classification of the feature histograms is performed using a DAG-SVM classifier, using LibSVM [CL01], as described below.

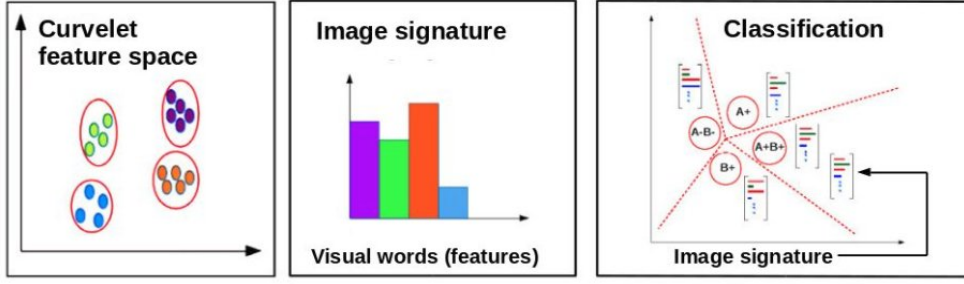


Figure 3.5: Bag of features pipeline (from left to right): K-means clustering in curvelet feature space, image signature (feature histogram) and classification of the image signatures.

3.2.2 Classification using a DAG-SVM framework

Generally, a classification problem starts by separating the data into training and test sets with class labels and several features (observations). The objective is to predict the test class labels given the test data features.

Support vector machine (SVM) [Vap95], originally proposed for binary classification, is a machine learning methodology, that seeks the optimal hyper-plane to best separate the two classes from each other with the widest margin (Figure 3.6). It is formulated as an optimization problem as illustrated below.

Given a training set of instance-label pairs (x_i, y_i) , $i = 1, \dots, l$ where $x_i \in \mathbb{R}^n$, (a data vector), and $y \in \{1, -1\}$, (class label of x_i), the support vector machines (SVM) require the solution of the following optimization problem [Vap95; HL02] :

$$\min\left(\frac{1}{2}w^T w + C \sum_{i=1}^l \xi_i\right) \quad \text{s.t.} \quad (3.5)$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, i = 1, \dots, l.$$

- x_i , the training vectors, are mapped into a higher dimensional space by the function ϕ .
- w (weights vector) is a normal vector to the hyper-plane $w^T x + b = 0$.
- $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel function. We consider here the radial basis function kernel $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, $\gamma > 0$.
- $C > 0$ is the penalty parameter of the error term that determines the trade-off between the maximization of the margin and the minimization of the error cost.
- ξ_i is called a slack variable, representing the distance from x_i to the margin plane $w^T x + b = y_i$.

The parameters for the SVM model described above, are C and γ , which have to be carefully selected for every given problem. In [HL02], the authors suggest a grid-search of both parameters as follows: various pairs of (C, γ) from a given range of values, are considered at a time. For each classifier defined by the pair

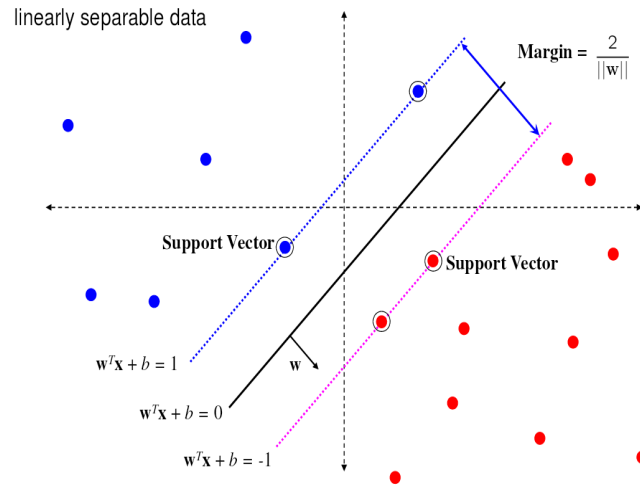


Figure 3.6: Graphical representation of binary SVM: separation of two classes (features (observations) represented by blue (one class) and red (second class)). Figure reproduced from [Bin].

of parameters, we perform a cross-validation during the training (learning) phase. The pair of values for which the cross-validation accuracy is the highest, is kept as the chosen classifier that will subsequently be used for the prediction phase (i.e. prediction of the class labels for the test samples).

The problem of classifying the FN networks involves comparing more than two classes, therefore we were interested in the possible ways of adapting the SVM framework to a multi-classification context. Directed Acyclic Graph Support Vector Machines (DAG-SVM) proposed in [PCST00] is one of the solutions that suggests using several binary SVMs to classify multiple classes. It is known and proven a superior algorithm comparing to other multiclass SVM techniques with respect to the training, evaluation time, generalization capacity [PCST00].

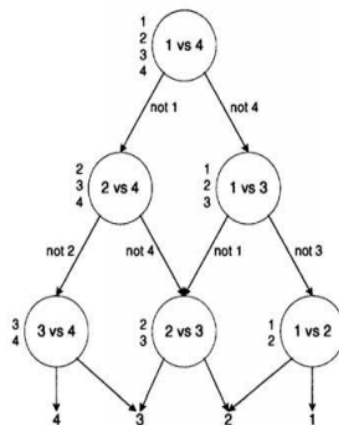


Figure 3.7: DAG-SVM Decision scheme for finding the best class out of four classes. The equivalent list state for each node is shown next to that node. Figure reproduced from [PCST00].

Typically, during the test phase, for a N -class problem, the DAG-SVM uses a directed acyclic graph (DAG)² with $N(N - 1)/2$ nodes which represent binary classifiers, one for each pair of classes (Figure 3.7). The DAG is obtained by training each i (vs) j -node only on the subset of training samples labeled by i or j [PCST00]. As for the decision step, starting at the root node, the binary decision function evaluates whether the next node to visit is at left or right. Finally, after $N - 1$ decision nodes, the leaf node indicates the output class. Thus, an important advantage with respect to other approaches (e.g. one-against-one method) is given by a smaller testing time.

3.2.3 Application to FN images classification with DAG-SVM, using a curvelet based representation

For the classification of the four FN variants, we deployed a database of 280 images of 3128×3128 pixels at $0.27\mu\text{m}/\text{pixel}$, acquired with a Zeiss 710 confocal system. Each class contains 70 images corresponding to the four FN variants. For speed convenience, we selected a representative region of 512×512 pixels from each image and used those regions for feature extraction and classification. We decided to use a non-exhaustive k -fold cross validation technique with $k = 4$ to evaluate the classification performance and its generalization capabilities.

The classification results were compared to those of a trained specialist, in terms of general classification accuracy, as well as confusion matrices.

Table 3.1 indicates the values of the confusion matrix for the automatic classification, while Table 3.2 shows the results of the specialist. The confusion matrix indicates that the classifier is highly capable of distinguishing the FN images belonging to variant B-A- from the rest of the others. Additionally, the classifier is presented with a greater challenge when it comes to distinguishing among classes B+A- and B+A+. A similar pattern was noted in the confusion matrix that corresponds to the classification performed by the specialist.

Actual \ Predicted	FN B-A-	FN B+A-	FN B-A+	FN B+A+
FN B-A-	90	0	0	10
FN B+A-	4.3	45.7	25.7	24.3
FN B-A+	2.9	25.7	64.3	7.1
FN B+A+	15.7	8.6	0	75.7

Table 3.1: Confusion matrix in percentage form of the DAG-SVM classification of FN variants, using curvelets

The fiber geometry associated to the B-A- FN variant, characterized by short filaments without a specific pattern, seems to be represented by a more discriminative geometric model. On the other hand, the topological properties of the fibers corresponding to FN A+, and FN B+ (i.e. fiber length and the presence of an apparent directionality) are quite similar, thus increasing the difficulty in differentiating between them. FN variant that incorporate the B+ domain is the least distinguishable,

² A DAG is a graph whose edges have an orientation and no cycles.

Actual \ Predicted	FN B-A-	FN B+A-	FN B-A+	FN B+A+
	FN B-A-	65.7	5.7	0
FN B+A-	0	48.6	34.3	17.1
FN B-A+	0	18.5	77.2	4.3
FN B+A+	5.7	37.2	2.9	54.2

Table 3.2: Confusion matrix in percentage form - Trained specialist

both in automatic and manual classification. Regarding the general accuracy of classification, the classification scheme that is proposed in this chapter (68.92%) outperforms the results obtained by a trained specialist (61.42%).

DISCUSSION: From a modelling perspective, the framework defined by the discrete curvelets has the benefit of providing a mathematical setting which is favourable to feature extraction and reconstruction from the coefficients. Anisotropic features at different resolutions that exhibit different orientations can be successfully detected through the curvelet transform. Moreover, the resulting curvelet coefficients can be used directly or through statistical measures for classification, compression, etc.

However, due to the specificity of the discrete frequency tiling methodology, the wedges (sectors) in neighborhood locations have often different sizes, i.e. different orientations at the same scale are captured by different number of curvelets. On one hand, this fact impedes a proper direct manipulation of the coefficients in the frequency domain (i.e. shifting the curvelet coefficients from one wedge to another so that rotation is achieved in spatial domain is difficult here because the wedges have different dimension). On the other hand, curvelet representation in the frequency domain of similarly-sized structures with different orientations may be represented by different amount of energy per scale, although the difference should be only illustrated by the wedge number.

To illustrate this difficulty, let us analyse the following example: [Figure 3.8](#) shows two line segments of 1 pixel width and different orientations. A vertical and diagonal orientation can be represented by the curvelets occurring at two neighboring sectors for each scale. The curvelet coefficients wedge plot at each scale in [Figure 3.9](#) show that for larger scales, i.e. 2,3 and 4, the amount of energy corresponding to the vertical line is higher than the diagonal line's one. We expected the total energies at different scales to only differ by their location (corresponding to a different wedge, i.e. different orientation).

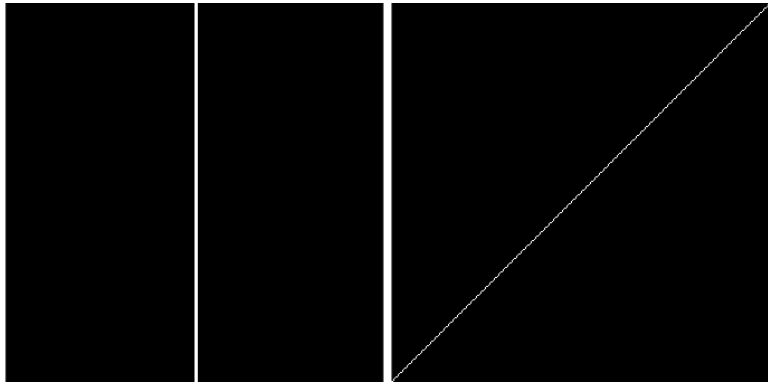
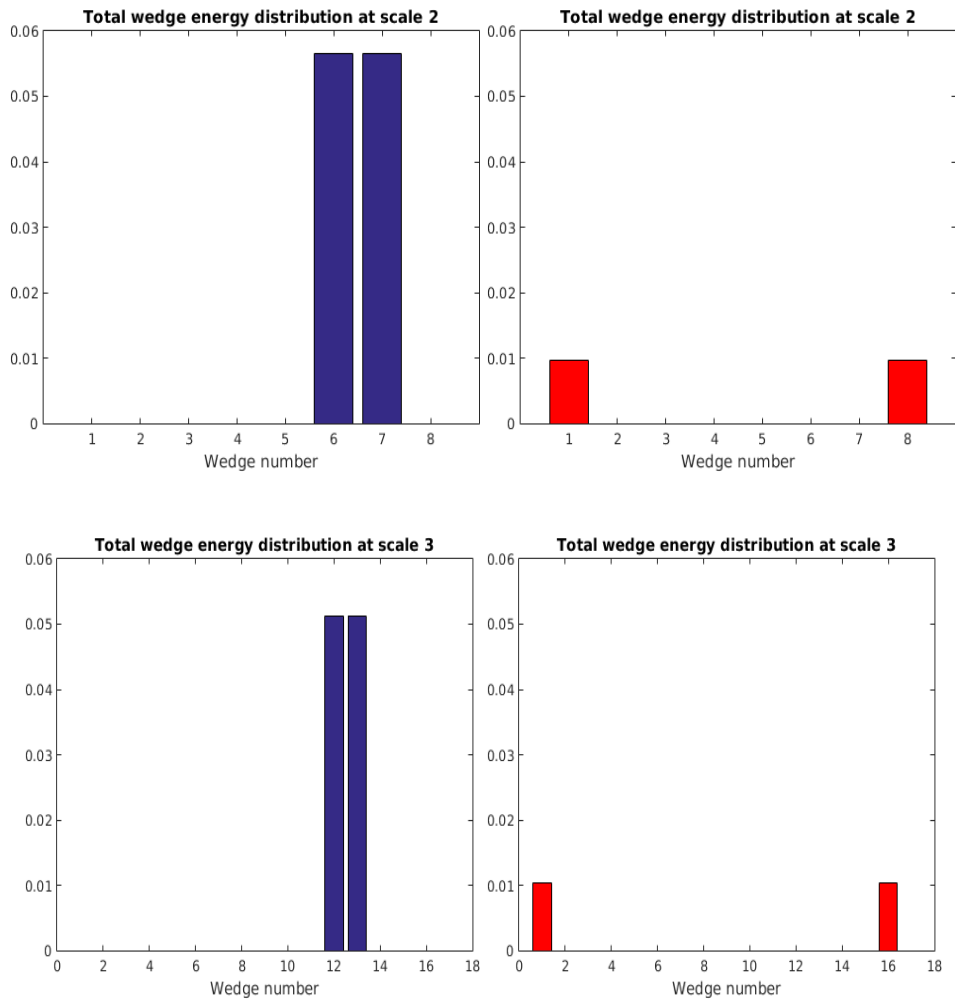


Figure 3.8: Vertical and diagonal line of 1 pixel width



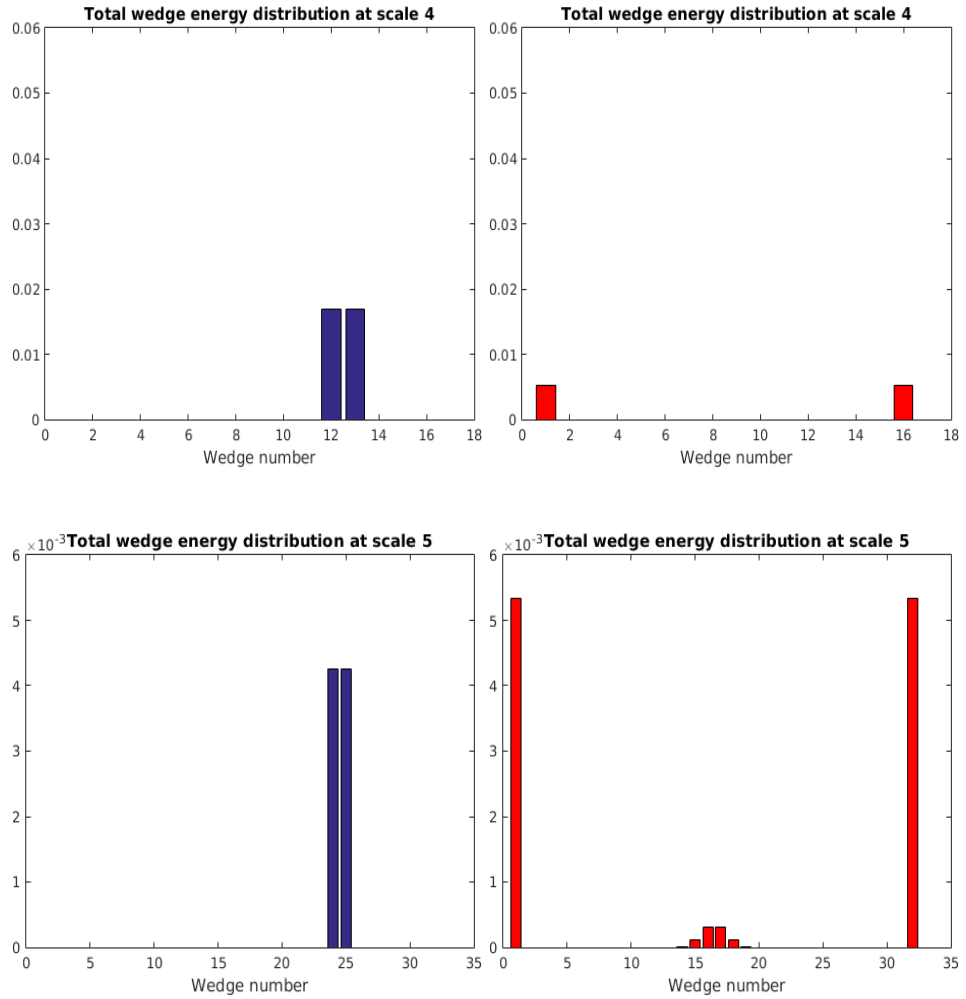


Figure 3.9: Curvelet coefficients energy distribution over the wedges (sectors) for each scale numbered in an ascending order starting from the center (2,3,4,5). The numbering of the wedges start from the upper left corner. Left side illustrations correspond to the vertical line, right-side illustrations correspond to the diagonal line. Only half of the sectors for each scale are illustrated here, since the energy levels are symmetric for the other second half.

These aspects have encouraged us to consider an alternative technique of the multiresolution approaches family, that has the potential of detecting "fiber atoms" of different dimensions and orientations, providing a simpler characterization. Gabor filters and their use for the FN fiber modelling is presented in the next section [Section 4.1](#).

3.2.4 Classification of the FN variants using Convolutional Neural Nets

The set of 280 gray-scale images (representative regions of 512×512 pixels), were additionally classified using a Convolutional Neural Net (CNN) [Cnn] architecture, typically developed to perform classification, segmentation, object recognition tasks by learning different features and pattern from images, text, sound, etc. In order to classify the FN images database using the MATLAB Deep Learning Toolbox, we have downloaded a pretrained model (i.e. previously trained on more than 1

million images to classify them into 1000 object categories), provided by GoogLeNet [Sze+14], a 22 layers deep network developed by Google.

A set of 196 images was used for the training of the algorithm, and the remaining 84 for testing it. The training image set was presented to the algorithm 25 times (epochs), in order to improve classification accuracy.

The results (Table 3.3) show that the information in the FN images is relevant enough in a CNN-based classification to distinguish FN variants better than curvelet-based features.

Actual \ Predicted	FN B-A-	FN B+A-	FB B-A+	FN B+A+
	FN B-A-	85.7	0	0
FN B+A-	0	80.9	14.3	4.8
FN B-A+	0	9.5	90.5	0
FN B+A+	9.5	14.3	0	76.2

Table 3.3: Confusion matrix in percentage form of the CNN classification of FN variant confocal images. General mean accuracy of classification is 83.3%.

3.3 CONCLUSIONS

FN network fibers exhibit local geometric properties that can be captured by curvelet features. We can reconstruct the fibers as a linear combination of curvelet coefficients at multiple scales and orientations. In addition, we are able to classify among the four variants of interest (in normal state), with a similar performance to that of a trained specialist.

FN-specific variant architecture is distinguishable in the confocal images, as highlighted by the results of the CNN-based classification. As one of the central questions throughout this work, was to determine whether the information contained in the confocal images is sufficient to distinguish among the four classes, the results of this type of classification enforced the idea that the FN variants are organized differently upon inclusion/exclusion of EDA/EDB. However, since the CNN-based architecture infers the features directly from the images, it is rather difficult to determine which fiber characteristics are truly discriminant. This is the reason why we adopted multiscale resolution techniques for feature extraction, such as discrete curvelets and Gabor filters, as we are about to see in the next chapter.

Part IV

FIBRONECTIN VARIANTS FIBER DETECTION AND
CHARACTERIZATION WITH GRAPHS

4

CONSTRUCTION OF THE GRAPH-BASED REPRESENTATION OF FN NETWORKS

In this chapter, we illustrate an alternative (standard) technique for detection of anisotropic, oriented structures, namely *Gabor filters*. We show that this choice can be more appropriate for a future modelling context of fiber features, than the discrete curvelets. We subsequently proceed to describe the pipeline that we have designed to extract a graph-based description of the fibers from the confocal images. This representation allows us to compute different statistics of FN fibrillar features and thus, compare and distinguish the FN-specific variants architecture. Properties that describe the geometry of the fibers, such as the general fiber orientation, thickness, anisotropy, fiber/pore density are bound to provide a meaningful characterization of the tissues.

4.1 GABOR FILTERS

Within the multi resolution methods, Gabor filters [Gab47] represent an alternative technique to capture various structures, at different frequencies and orientations. It belongs to the linear local filters category and has been extensively used for edge detection [PK97], texture discrimination [Tur86], facial expression recognition [SGP09], optical character recognition, etc.

In the spatial domain, a 2D Gabor filter is represented by an elliptic Gaussian kernel function modulated by a sinusoidal function, see Figure 4.1. In order to detect objects in an image that appear at various frequencies characterized by preferred directionalities, one can typically proceed by constructing a set (bank) of filters with the appropriate characteristics (given by the shape of the Gaussian kernel or by the frequency of the sinusoidal wave). Subsequently, the image is filtered with this set of Gabor kernels, the result of which can provide a feature database that can be further used for analysis, classification, or segmentation [KPG02]. Fibrillar structures were detected and enhanced with Gabor filters, such as defined in [Pet95; Dau85], commonly employed in image processing for the detection of structures with different frequencies, and certain directionalities.

Gabor functions can be used to numerically simulate the 'simple cells' of the primary visual cortex (mammalian brain), as frequency and orientation representations of Gabor filters are similar to those of the human visual system [Pet95]. Besides, filter banks are among the biologically inspired recognition systems [Ham13].

4.1.1 Gabor kernels definition

The exponential term provides the shape of a bivariate Gaussian kernel (Figure 4.3), and the cosine function (carrier) describes its oscillations in space, while $\mathbf{v} = (x, y)^T$ is the 2D coordinate vector, indicating pixel localization in a bi-dimensional

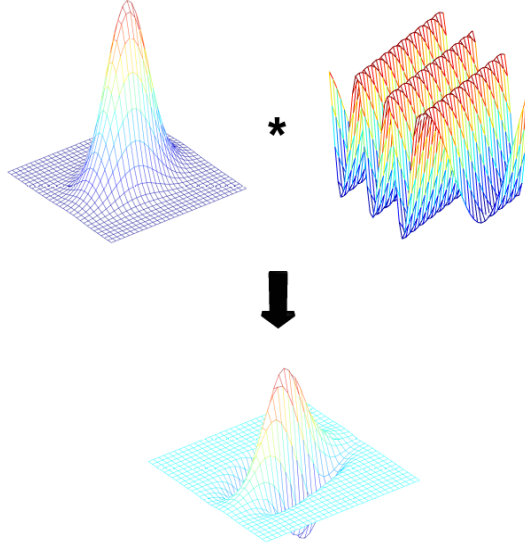


Figure 4.1: Gabor filter 3D view: Top-left: Gaussian kernel, top-right: Cosine wave function, bottom-middle: Gabor kernel function obtained by the modulation of the Gaussian kernel on the cosine carrier.

Cartesian coordinate system. One Gabor kernel g_k is characterized by the following formulation ¹:

$$g_k = \exp\left(-\frac{1}{2}\mathbf{v}^t \Sigma_{\theta_i}^{-1} \mathbf{v}\right) \cos\left(2\pi \frac{x_{\theta_i}}{\lambda_j} + \phi\right) \quad (4.1)$$

where $x_{\theta_i} = x \cos \theta_i + y \sin \theta_i$ and θ_i is the rotation angle (with respect to the horizontal axis $-x$) of the Gaussian envelope.

Let Σ be the covariance matrix of the bivariate (anisotropic) Gaussian kernel. σ_x and σ_y represent the standard deviations (in pixels) of the Gaussian function along the two axes (x, y) . Then Σ is a symmetric and positive definite matrix (that consequently admits an inverse Σ^{-1}). A counterclockwise rotation (Figure 4.2) in the 2D Cartesian system with R_θ is applied to Σ^{-1} . It follows that $\Sigma_{\theta_i}^{-1}$, the inverse of the covariance matrix of the bivariate Gaussian function, rotated with θ_i , has the form in Equation 4.2.

$$\Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}, \quad R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}, \quad \Sigma_{\theta_i}^{-1} = R_{\theta_i} \Sigma^{-1} R_{\theta_i}^t \quad (4.2)$$

Other specific parameters of the Gabor kernel are λ_j , the wavelength (in pixels) of the cosine term and ϕ , which represents the phase of the (cosine) carrier.

¹ Some works depict an alternative formulation of the Gabor kernel, which is based on a complex sinusoidal wave that modulates the Gaussian function. The formulation in Equation 4.1 relies only on the real part of this complex sinusoidal wave (shown to be an appropriate edge detector) to capture the fibers.

The family of 2D Gabor functions is shown to achieve the theoretical limit of joint uncertainty of spatial location and frequency [Dau85].

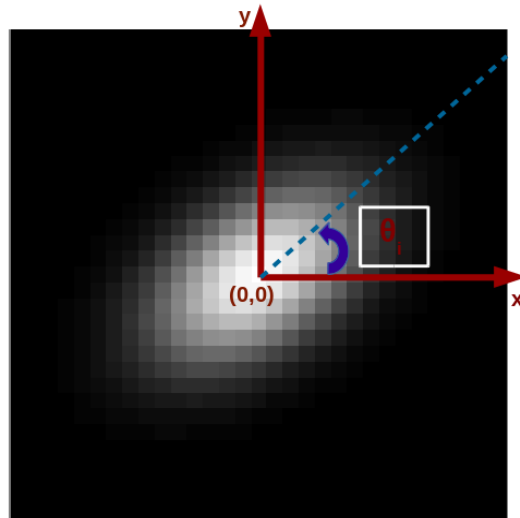


Figure 4.2: Counterclockwise rotation with an angle θ_i around the origin of a 2D Cartesian plane of an elliptic Gaussian kernel

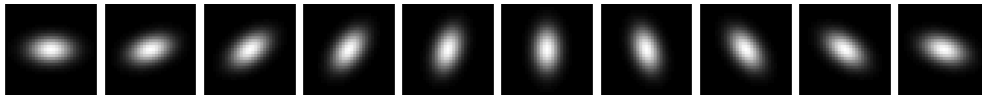


Figure 4.3: Gaussian kernels: variation of θ_i from 0 to $9\pi/10$. $\sigma_x = 5$ pixels and $\sigma_y = 3$ pixels

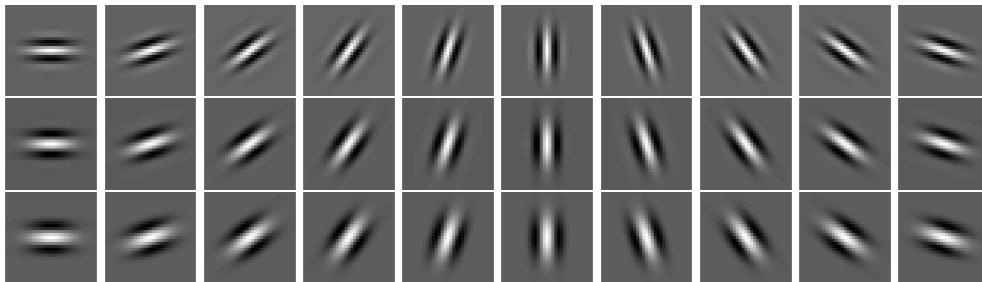


Figure 4.4: Gabor kernels: column-wise: variation of λ_j from 6 to 8 pixels, row-wise: variation of θ_i from 0 to $9\pi/10$, with a step-size of $\pi/10$. $\sigma_x = 5$ pixels and $\sigma_y = 3$ pixels.

4.1.2 Filtering of the FN confocal images using Gabor kernels

Using Gabor filters for feature detection is commonly performed by applying a set of predefined filters to the input image: the filtered output provides the spatial localization of the detected structures corresponding to the filter response. For a given kernel having a certain shape in the spatial domain (see [Figure 4.4](#)), the pixel intensity of the output is subsequently higher in the regions where there is a positive correlation between the input image and the given kernel, in other words where detection of the specific structure occurs. This practical aspect is formalized as follows.

If we consider $I(x, y)$ the pixel intensity of 2D grayscale image defined on a discrete grid, then the convolution of I with a Gabor filter whose kernel, denoted g_k , has the form of the [Equation 4.1](#):

$$I_k^c(x, y) = (I * g_k)(x, y) \quad (4.3)$$

In order to capture the various geometrical properties of the FN fibers using Gabor filters, we have constructed a set of Gabor kernels g_k , $k \in \{1, 2, \dots, 60\}$ defined by the following parameters:

- Fiber orientation is represented by θ_i , computed as $\theta_i = \frac{i\pi}{20}$, where $i = \{0, 1, 2, \dots, 19\}$.
- Fiber thickness is represented by λ_j , $j = \{1, 2, 3\}$ that corresponds to the wavelength (in pixels) of the cosine term, the values of which are equal to $\lambda_j/2$ and vary between 3 and 5 pixels. The thinnest fibers are detected when $\lambda_j = 6$ pixels, medium thickness fibers correspond to $\lambda_j = 8$ pixels, while the thickest are characterized by $\lambda_j = 10$ pixels.
- For accurate localization of fibers the phase of the cosine function, ϕ , is set to 0.
- The spatial support of the kernels is given by $\sigma_x = 5$ pixels and $\sigma_y = 3$ pixels, indicating an anisotropic filter that is appropriate for fiber detection.

At one specific location indicated by (x, y) within the maximum Gabor filtered image M_G , we retain the Gabor kernel (and its parameters) that returns the highest coefficient among the predefined kernels set ([Figure 4.5](#)):

$$M_G(x, y) = \max_k(I_k^c(x, y)) \quad (4.4)$$

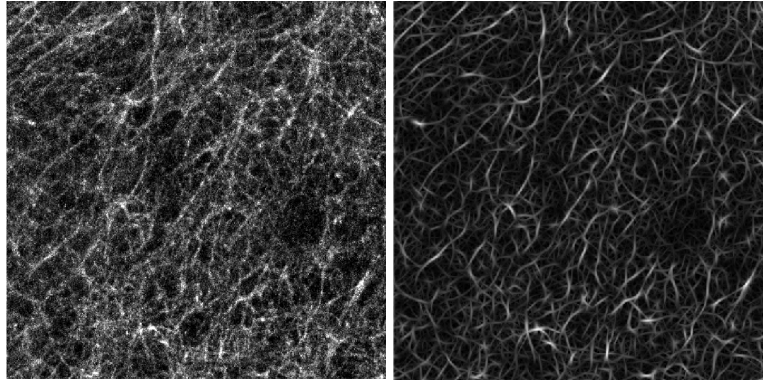


Figure 4.5: Maximum Gabor filtered Image: FN B-A+ Sample of 512×512 pixels (left) and the representation of the detected fibers using Gabor kernels (right).

4.2 GRAPH REPRESENTATION OF THE FN FIBERS- METHODOLOGY

Since the pixel intensity of a detected fiber of M_G corresponds to the best-responding Gabor filter, we are thus able to retain its specific parameters which can directly

be linked to physical attributes, such as fiber thickness and local fiber orientation. Gabor filters can provide a good descriptive characterization of the fiber geometry, however, we are interested to extend this characterization to allow the inclusion of supplementary parameter statistics e.g. fiber length, degree of fiber cross-links, etc, into the general model.

The following subsections illustrate the methodology for deriving the FN graphs starting from the images of detected fibers using Gabor filters.

4.2.1 Computation of the graphs associated to FN morphological skeletons

Fiber detection and enhancement with Gabor filters constitutes the first step of the graph-based model that we intent to construct for fiber characterization. The images of previous extracted fibers represent the "canvas" for the subsequent series of image morphological transformations meant to simplify the fiber delineation and allow the conversion to graphs. We managed to compute the morphological skeleton of the FN detected fibers, i.e the medial axis, a complete shape descriptor, using different morphological operations. Subsequently, the graph-based description of fiber skeletons was obtained using a toolbox originally dedicated for generating the network graph of a 3D skeleton voxel that we have adapted to the 2D setting [Kol+17a].

First, we start by binarizing (using hysteresis thresholding) the gray-scale images containing the FN detected fibers, previously obtained, to prepare the skeletonization. Converting the fiber structure into a skeleton is useful to reduce the amount of data to a representation that encapsulates the shape and preserves the connectivity of the original region. The morphological skeleton is also defined as the loci of centers of bi-tangent circles that fit entirely within the foreground region being considered (see an example in Figure 4.6).

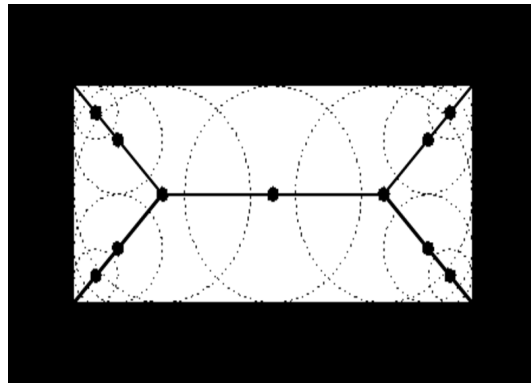


Figure 4.6: Skeleton (black) of a (white) rectangle defined in terms of bi-tangent circles. Figure reproduced from [Ske].

We subsequently chose to compute the morphological skeleton of the fibers using the toolbox [Kol+17a] based on a homotopic (i.e. one object can be continuously deformed into the other) thinning algorithm [Ta-94]. The underlying idea is to thin the objects into skeletons whose thickness is one pixel in at least one of the two dimensions. The thinning algorithm iteratively removes the border points of an object, while preserving the end points of the line segments until no more thinning

Intuitive explanation of the skeletonization algorithm inspired from [Ske]: consider that the foreground regions in the input binary image are made of some uniform slow-burning material. Light fires simultaneously at all points along the boundary of this region towards the interior. At the points where the fire traveling from two different boundaries meets itself, the fire will extinguish itself forming the skeleton.

is possible, in other words, satisfying the following topological and geometrical constraints:

- topological: preservation of the number of connected objects in the object (fibers) (by respecting Euler's constraint from digital topology that connects the number of objects, cavities and holes in a consistent manner [Ta-94].)
- geometrical: ensure the desired width and location of the skeleton.

Finally, the network graph associated to the skeleton representation of the fibers, (example illustrated in [Figure 4.7](#)) is the collection of nodes (fiber ends or fiber crosslinks (such as they are perceived in the 2D representation ² of various degrees indicated by the number of adjacent fibers) that unite the fiber edges. This structure has the advantage of providing the means to compute different parameters characteristic of the (fiber) graph topology and geometry, such as the number of fibers, the fiber lengths (in terms of number of (non-zero) pixels (from the corresponding skeleton) between two given nodes), the degree of network connectivity based on the different proportions of node degree, etc. Additionally, Gabor-based features can be combined with the graph data to describe fiber density (in terms of local fiber thickness coupled with fiber length), predominant fiber orientation, etc. The analysis of the pore shape/size can be made starting from the fiber skeleton.

4.2.2 Post-improvement methodology for the FN graph-based representation

The skeletonization tool does not always manage to capture the fibers as desired and thus, a post-improvement step was necessary to capture the fibers with a higher degree of fidelity. The graph-based representation facilitates the set up of the proposed method to reconstruct the missing fibers due to previous morphological operations and noise in the data. We intend to reconnect the fibers within a predefined area around the fiber extremities, i.e., degree 1 nodes, denoted n_k , assuming that there are more chances for a fiber to be reconnected in a surrounding area, along the direction of the local orientation of n_k .

In short, for an extreme node n_k , predefined radius R and cone sector $\delta_c \in (0, \pi/2)$, we search for the candidate pixels that can be reconnected with n_k . Below we describe the methodology that selects the set of eligible pixels as fiber ends (the following steps were applied to every n_k):

- Selection of the set of pixels intersecting a disk of radius R centered on the pixel of index n_k and within a cone sector δ_c around the local Gabor orientation θ_i , where $\theta_i \in [0, \pi)$:

$$S_p = \{(r_x, r_y) : r_x^2 + r_y^2 \leq R^2, \theta_i - \delta_c/2 \leq \arctan(r_y/r_x) \leq \theta_i + \delta_c/2\} \quad (4.5)$$

where (r_x, r_y) are the relative coordinates of pixels in the 2D Cartesian system considering that $(r_x(n_k), r_y(n_k)) = (0, 0)$. ([Figure 4.8, a](#))

² We note here, that the thickness of the matrix tissue is not taken into account when constructing the graph representation, as we are mainly interested in describing the two dimensional aspect of the FN fibers, corresponding to the available data. The microscope integrates the signal across the 3rd dimension (tissue thickness is negligible here with respect to the microscope's optical resolution in z-axis) to obtain the 2D figure.

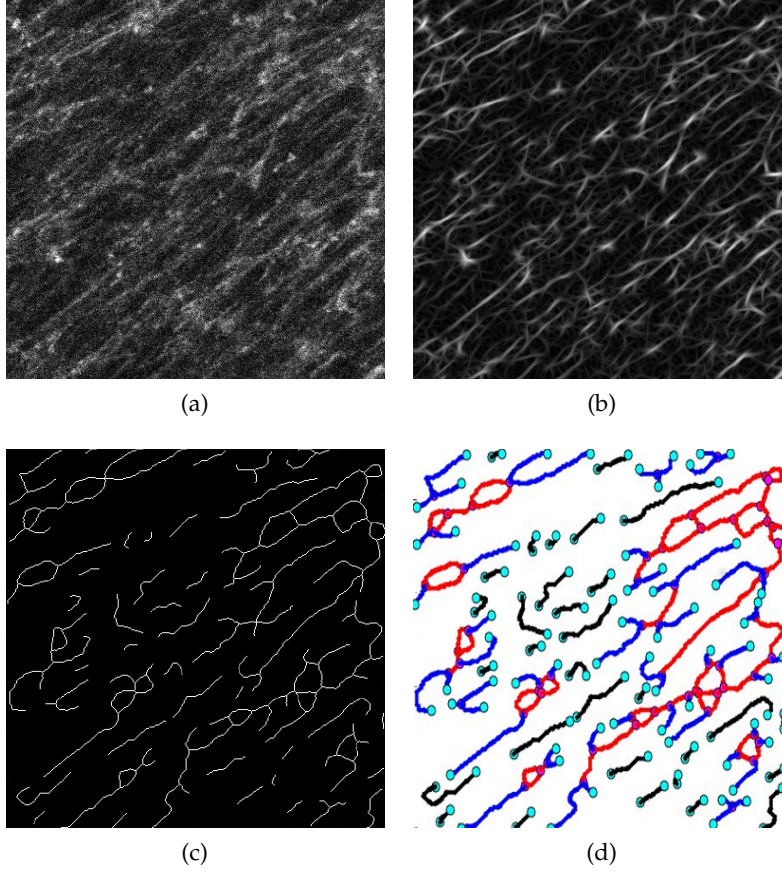


Figure 4.7: Morphological skeletons of the FN fibers and the corresponding graphs: (a) FN B-A+ confocal image sample mimicking a tumoral state of 784×784 pixels, (b) Fibers detection using Gabor filters, (c) Morphological skeleton of the extracted fibers at (b), (d) Skeleton-associated graph illustration

- To narrow down the number of pixels in S_p , we proceed to select them as follows:
 1. for the pixels that correspond to degree 1 nodes, they will be considered eligible candidates only if $n_k \in S_{\text{sym}}$, where we define by $S_{\text{sym}} = \{(r_x^s, r_y^s) : \theta_p - \delta_c/2 - \pi \leq \arctan(r_y^s/r_x^s) \leq \theta_p + \delta_c/2 - \pi\}$ as the set of pixels coordinates relative to $n_p(0,0)$ in the cone sector around the local orientation θ_p of the candidate pixels in S_p . In other words, we keep the candidate pixels for which n_k belongs in the symmetric cone around the local θ_p . (Figure 4.8, b,c)
 2. for the pixels that correspond to the edges, the decision rule implies that they are accepted only if $n_k \in S_{\text{ort}}$, where we define by $S_{\text{ort}} = \{(r_x^o, r_y^o) : \theta_p - \delta_c/2 \pm \pi/2 \leq \arctan(r_y^o/r_x^o) \leq \theta_p + \delta_c/2 \pm \pi/2\}$ as the set of pixels coordinates relative to $n_p(0,0)$ in the cone sector perpendicular on the local orientation θ_p . (Figure 4.8, d).
- At this step, the feasible candidates were chosen. In order to reconstruct the missing fibers we start by computing the minimal weighted paths from n_k to

all candidates, using Dijkstra's shortest path algorithm.³ The weights account for the image intensity, i.e. the reconnection is considered only on the paths where fibers were "defined" in the original image (local intensity is above a certain threshold), but not necessarily captured during skeletonization.

- Choose the path of minimal length among all possible paths and recompute the weighted path from n_k to the chosen pixel for reconnection, using as a guideline the maximum Gabor filtered image M_G (Figure 4.5).

4.3 CONCLUSIONS

In this chapter, we have illustrated one of the main contributions of this thesis, essential for FN characterization, which is the construction of a graph-based representation of the fibers built on top of Gabor features.

Within the literature dedicated to the multiresolution analysis, the discrete curvelets have earned a better reputation than Gabor filters when it comes to applications concerning anisotropic feature detection, classification, etc. The discrete curvelet transformation is a tight frame with low redundancy, while Gabor filters, (not built as orthogonal structures), fail to cover the entire frequency spectrum (unlike curvelets) and present a higher redundancy.

However, the general context of the work presented in this manuscript is focused on the design of meaningful modelling approaches to characterize the geometry of the FN fibers. Thus, upon applying a multi-scale transformation based on dictionary atoms with certain features, it is essential to be able to easily manipulate the coefficients/parameters subsequently employed under different forms. In this regard, we found that the curvelet coefficients are more difficult to operate with (Chapter 3), while the Gabor-based detection provides a simple fiber delineation, (i.e. for each pixel, we can access the corresponding Gabor parameters). Additionally, the Gabor-detected fibers served as starting point for building a faithful graph description of the fibrillar structures. This representation is built on top of Gabor detected fibers, and employs morphological operators to extract the fiber skeleton and subsequently associate the graph, as a collection of nodes (fiber crosslinks, or fiber ends) that connect the edges.

A post-processing tool for the missing fibers (due to noise, skeletonization defects, etc.) reconnection was proposed (see example in Figure 4.9), in order to capture the information from the confocal images with a higher degree of fidelity. This is a 2D graph-based representation of the available 2D data (Figure 4.10), which is useful for extracting additional fiber features (e.g. fiber length, node degree, etc). Furthermore, as we are about to see in the next chapters, it will also provide the appropriate setting to extend the FN characterization model to compare FN networks, (e.g. through matching their associated graphs), to classify the FN graphs, to derive the graph barycenter for as the mean "individual" of a class, etc.

³ Dijkstra's shortest path algorithm finds the shortest paths from the source node to all the other nodes in the graph. In this setting, the source node is n_k and the "destination nodes" represent the candidate pixels. The "graph" in the algorithm's context is constituted by pixels and the cost to move from the current pixel C_p to the neighbouring pixels V_k is a function of image gray-level intensity: $\forall k \in \mathbb{Z} : 1 \leq k \leq 8, \text{Cost}(C_p, V_k) = \frac{1}{\text{Intensity}(V_k) + \epsilon}$.

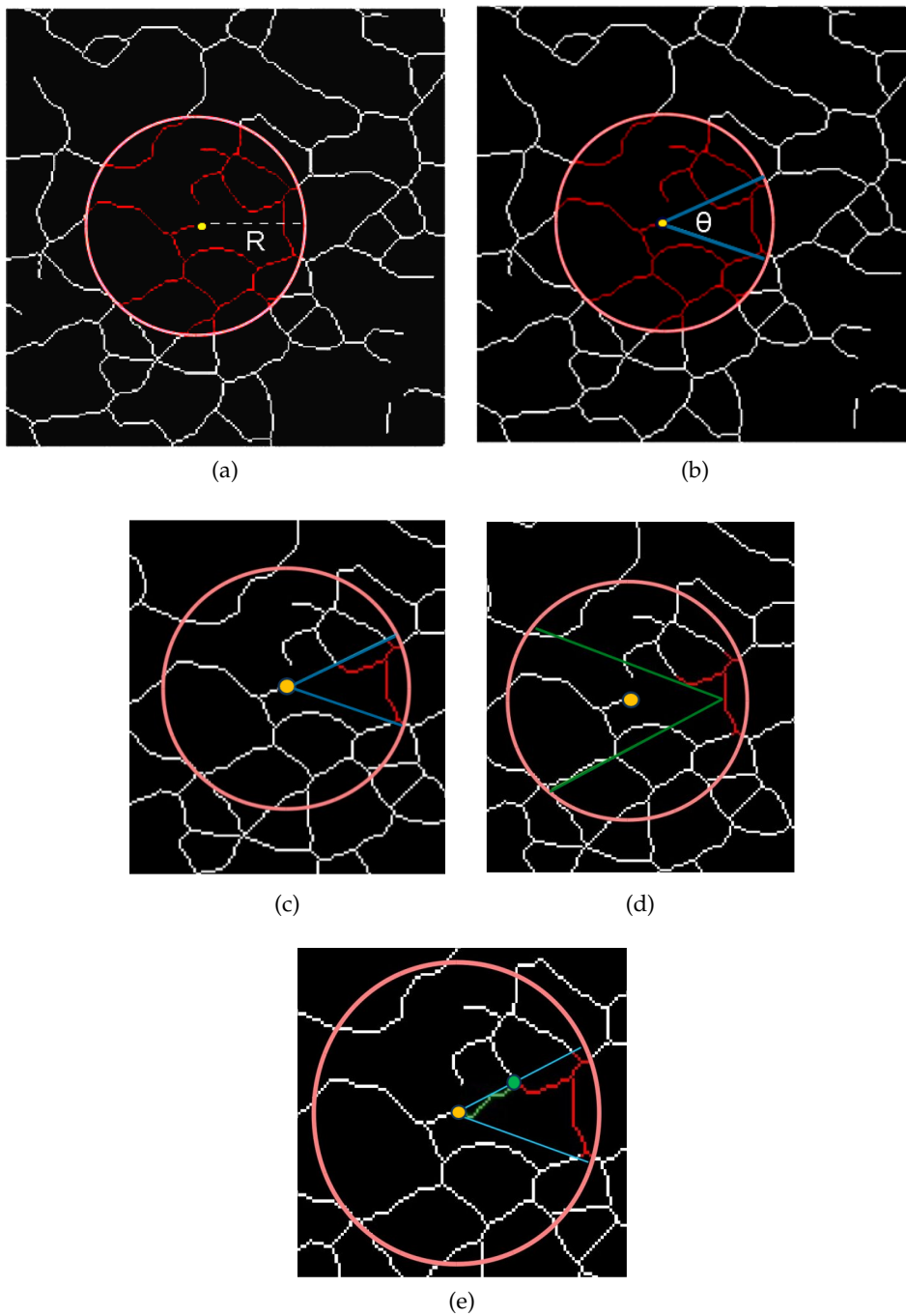


Figure 4.8: Methodology for morphological skeleton and graph improvement: (a) Selection of pixels intersecting a disk of radius R centered on the pixel n_k (yellow disk). (b,c) Selection of pixels inside a cone sector around θ_i . (d) Eligible pixels on the fibers define a cone sector perpendicular on the corresponding local orientation θ_p . (e) The accepted pixel (green disk) has the minimal path length towards the n_k ; the reconstructed fiber corresponds to the green path, which was computed with Dijkstra's weighted algorithm when the weights are proportional to the pixel intensity in the image of detected fibers with Gabor kernels.

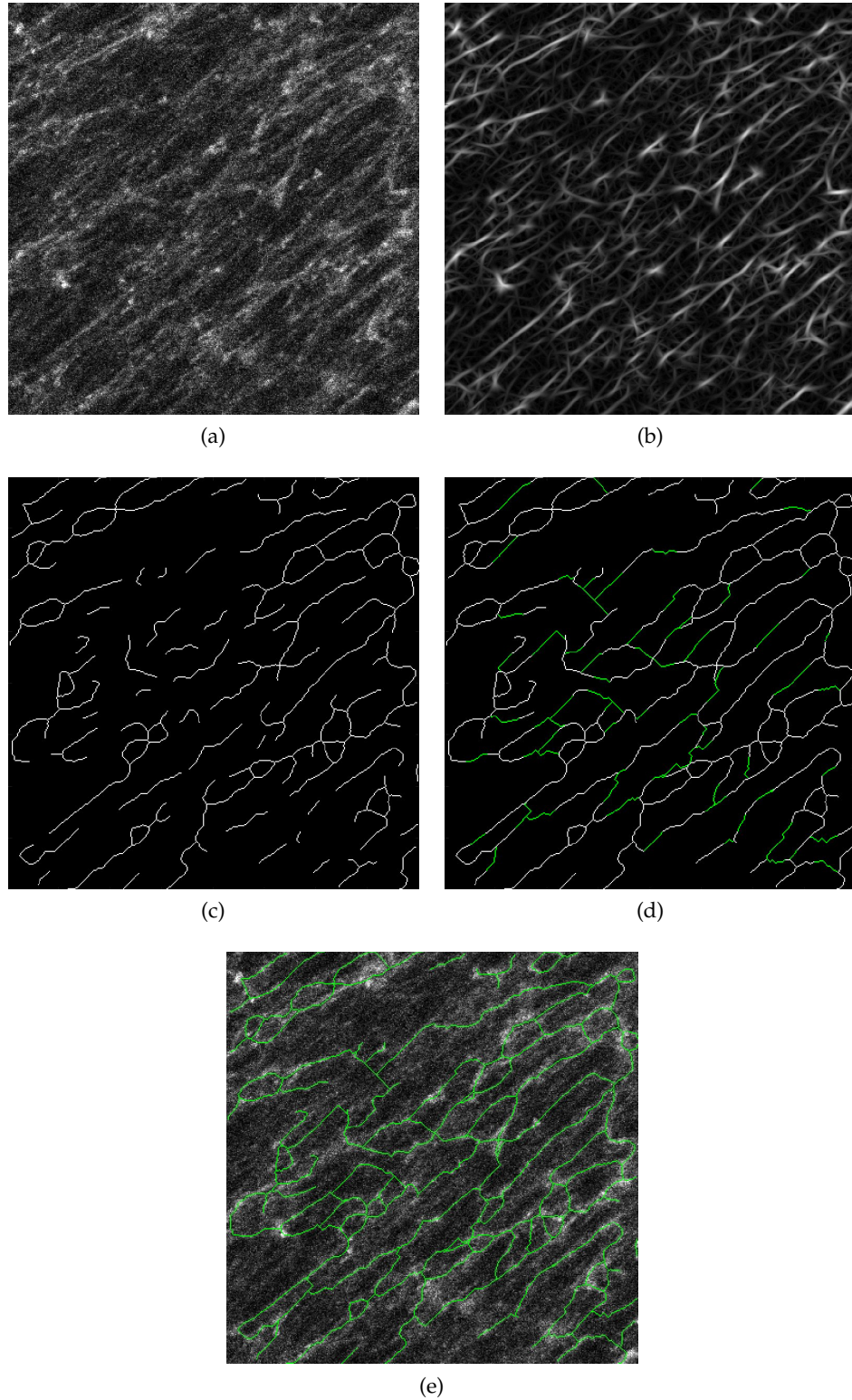


Figure 4.9: Pipeline fiber detection and reconnection: (a) FN B-A+ confocal image sample mimicking a tumoral state of 784×784 pixels, (b) Fibers detection using Gabor filters, (c) Morphological skeleton of the extracted fibers at (b), (d) Reconnected skeleton of the extracted fibers at (b), (e) Final reconnected skeleton displayed on top of the original confocal image (a).

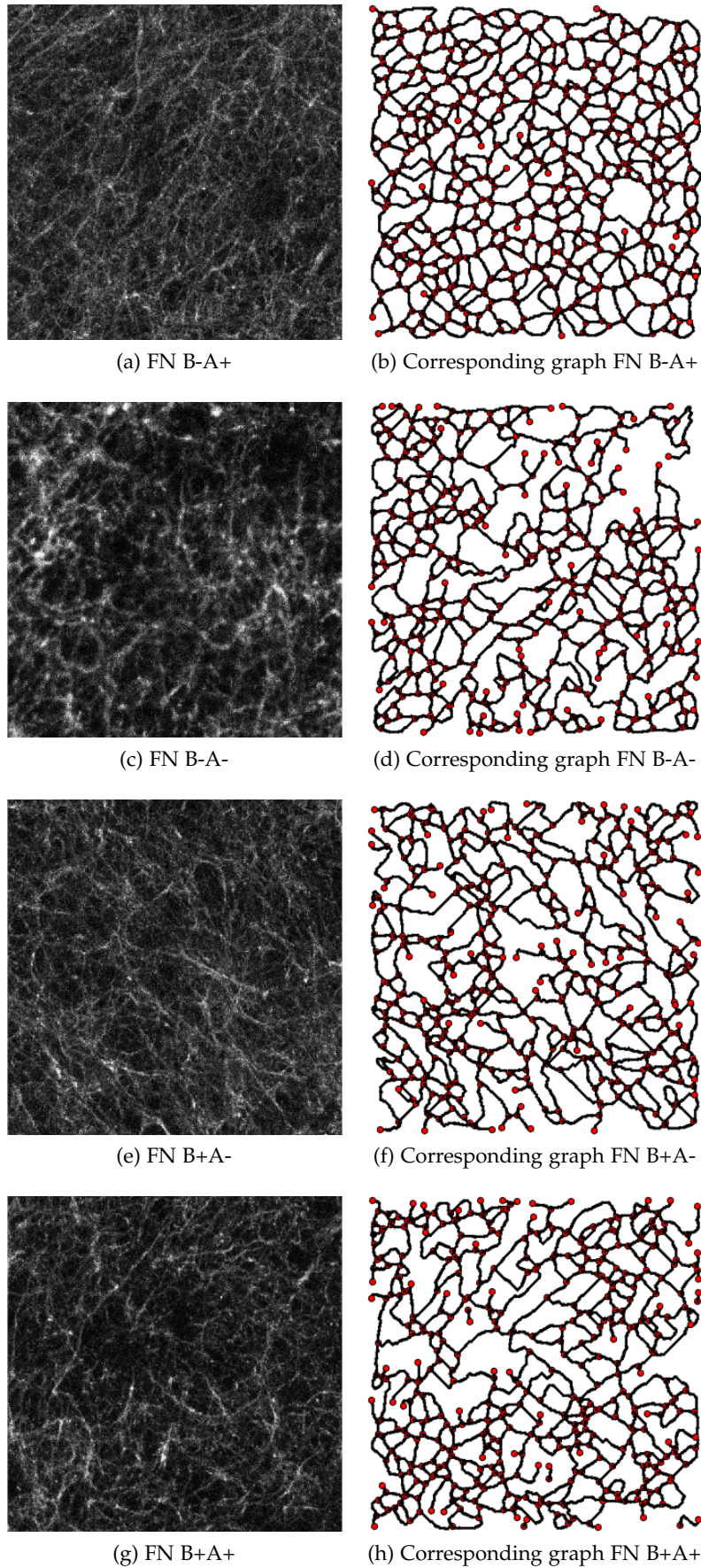


Figure 4.10: Reconnected graphs of the four FN-specific variants (normal state) corresponding to images of 512×512 pixels: (a,b) FN B-A+, (c,d) FN B-A-, (e,f) FN B+A-, (g,h) FN B+A+.

5

LOCAL CHARACTERIZATION OF FIBER FEATURES USING THE GRAPH-BASED REPRESENTATION

In this chapter, we establish a local characterization of the FN networks based on local geometrical features, extracted from the graph representation (see [Chapter 4](#) for details) (e.g. fiber network connectivity, fiber length, median pore dimension, etc.), as well as Gabor features (e.g. fiber local thickness). We subsequently show that the FN variants (normal state) can be compared and distinguished among them, first after performing PCA analysis and computing local parameter distributions across classes. Secondly, we classify the FN variants using the above features classified by a DAG-SVM classifier. These results are bound to prove that the graph-based representation contains relevant and meaningful information about the fibers.

5.1 FEATURE EXTRACTION FROM GRAPH-BASED FN REPRESENTATION

Features related to fiber thickness and connectivity were directly computed using Gabor kernels and graph-specific parameters:

- connectivity was defined as the proportion of degree 1 nodes (those corresponding to fiber ends) relative to the nodes with a degree higher than 2 (corresponding to branching and intersecting points). The variant-specific connectivity distributions are shown in [Figure 5.1 \(a\)](#). Interestingly, B-A- fibers are characterized by a higher abundance of fiber ends, delineating a low level of connectivity, compared to the other variants, especially to B-A+. These results reveal that the absence of Extra Domains leads to a less branched FN fiber arrangement.
- Next, we considered fiber thickness by computing the proportion of thin to thick fibers. As shown in [Figure 5.1 \(b\)](#), B-A- fibers display low proportion of thin fibers, hence characterized by the presence of medium and thick fibers, while the opposite is observed for B-A+.
- In order to analyze fiber thickness heterogeneity, or fiber diversity, we computed the fiber thickness kurtosis, a parameter that indicates how outlier-prone the fiber thickness distribution is relative to a normal distribution with identical variance. In terms of fiber thickness, B-A+ values are distributed around the mean, suggesting a high homogeneity in fiber thickness, compared to B-A- fibers which are highly heterogeneous ([Figure 5.1 c](#)).
- Pore shape was measured through a circularity parameter and the average pore size. Circularity measures pore anisotropy allowing us to distinguish

circular and oval-like pores. Pore circularity was determined by the formula $4\pi\text{Area}/\text{Perimeter}^2$, the values of which vary between 0 (line) and 1 (perfect circle). **Figure 5.1 (d)** shows that B-A- FN arrangements are characterized by a high number of oval pores, while pores in B-A+ FN networks are predominantly circular.

- The same pattern is observed in terms of pore size, a parameter whose values were considered starting from the 90th percentile. Large pore sizes are found within B-A- FN networks, while smaller pore size is observed in B-A+ FN networks (**Figure 5.1 (e)**).

5.2 PCA VISUALISATION OF THE GABOR AND GRAPH-BASED FN FIBER FEATURES

After establishing a faithful representation of the fibers by describing them in terms of graphs, a set of features was selected to characterize fiber geometry and to perform principal component analysis (PCA). The PCA method was used to explore the relatedness between the different FN variants with respect to various physical attributes, or features, including i) connectivity, ii) fiber thickness, iii) fiber heterogeneity, iv) pore shape v) and pore size distribution. The PCA in **Figure 5.2** was performed with the aforementioned features, by adopting the representation provided by the first two principal components. The plot illustrates both the samples (images) projected in a bi-dimensional space, and the five features represented by a vector, the direction and length of which indicate the contribution of each feature to the two principal components.

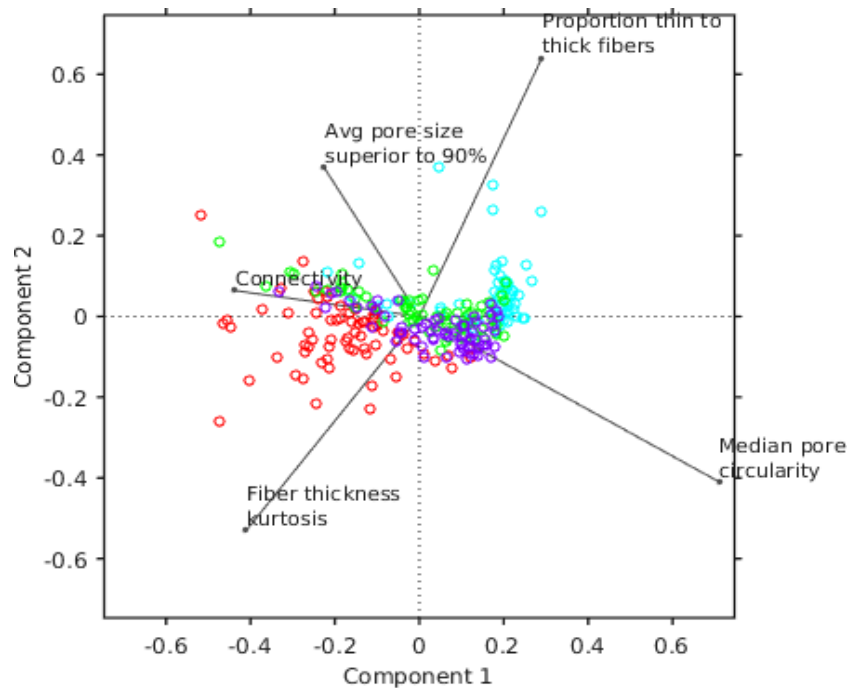


Figure 5.2: PCA Analysis of the Gabor and graph-based FN-specific fiber features: **FN B-A-**, **FN B+A-**, **FN B-A+**, **FN B+A+**.

Generally, the samples belonging to FN B-A- and FN B-A+ are concentrated in non-overlapping areas, displaying the distinguishability of these two variant-specific FN networks through the chosen features.

Altogether, these analyses demonstrate that B-A+ FN matrices feature highly branched, homogeneous, thin fibers that form small pores. In contrast, B-A- FN forms thicker, unbranched networks with larger more elongated pores. Interestingly, the presence of EDB results in matrices (either B+A-, or B+A+) characterized by a mixture of the attributes seen in B-A- and B-A+.

5.2.1 Classification of the Gabor and graph-based FN fiber features

The 280 set of 512×512 pixels images represented by the Gabor and graph-based features has been classified with a DAG-SVM multi classifier with 5-fold cross-validation. The results in Table 5.1 illustrate that the five features help distinguish the two variants, FN B-A- and FN B-A+, the best out of the four variants, confirming the observations made both in the PCA analysis and also during the classification using discrete curvelets.

Concerning the general classification accuracy, the value of 66% is higher than that obtained by the trained biologist and comparable to the curvelets performance, indicating that the information captured by the graph description is relevant to distinguish the FN fibers and meaningful for obtaining an appropriate geometrical characterization.

Actual \ Predicted	FN B-A-	FN B+A-	FN B-A+	FN B+A+
FN B-A-	92.9	7.1	0	0
FN B+A-	7.1	43	21.4	28.5
FN B-A+	0	28.5	71.5	0
FN B+A+	7.1	35.7	0	57.2

Table 5.1: Confusion matrix in percentage form of the DAG-SVM classification of FN variants, using Gabor and graph-based features

5.3 CONCLUSIONS

This chapter illustrated an application of the graph representation of FN variants (normal state) for local fiber characterization, based on meaningful physical attributes describing the fiber thickness, heterogeneity, pore shape and network connectivity. These features are significant not just in a classification context, but also for understanding which are the meaningful properties that determine the structural differences within the FN fiber organization. We have thus shown that the constructed graphs embed meaningful information of the FN networks.

In the following chapters, we extend this analysis to include differences on a topological level between the graph-based fiber representations, using techniques dedicated to graph comparison. Furthermore, the methodologies that we have

identified as being relevant in this context, prove to be favourable for studying the variation of certain fiber geometrical properties in a statistical analysis framework.

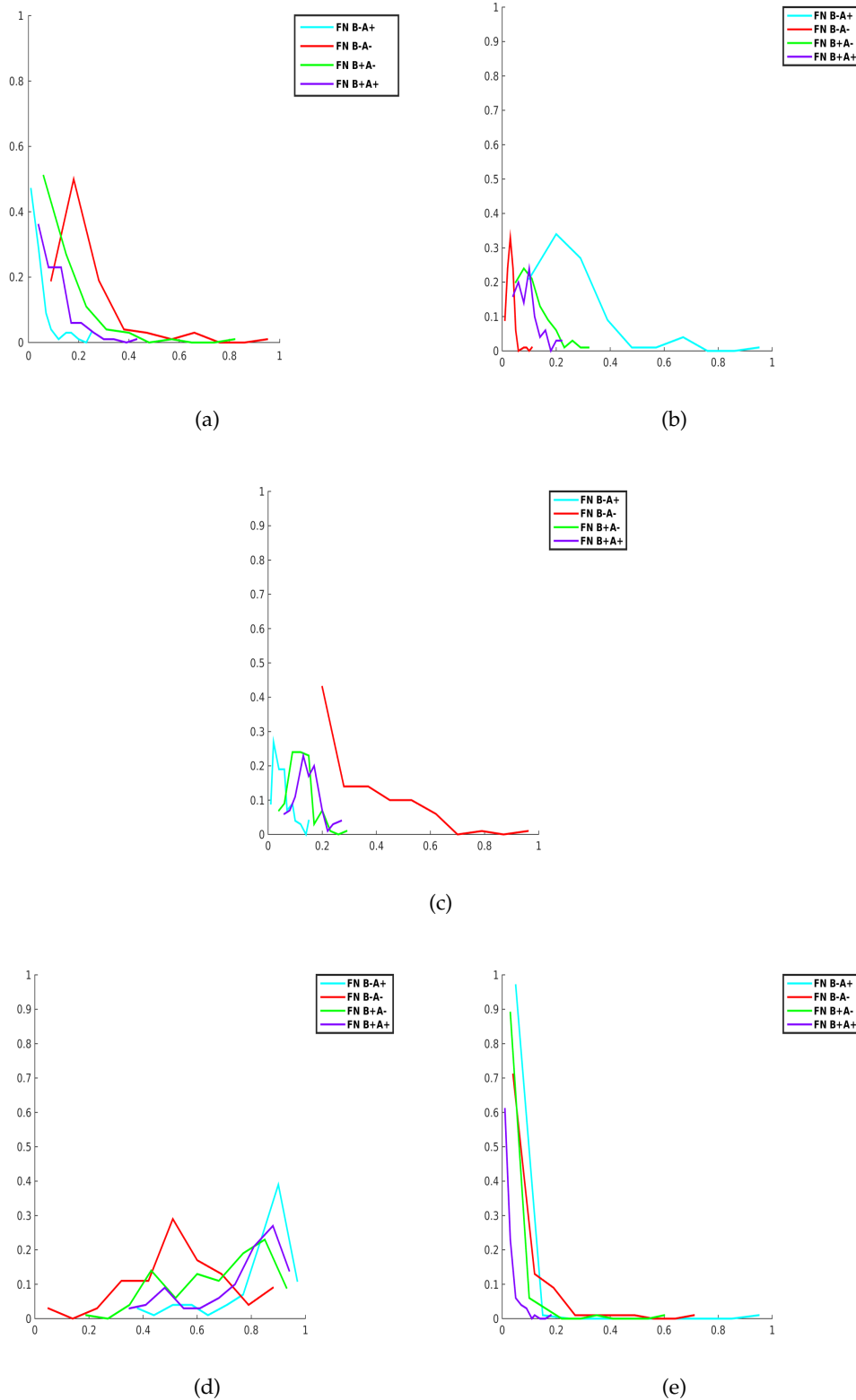


Figure 5.1: Graph-based normalized feature distributions: (a) Proportion of degree 1 nodes relative to the nodes of superior degree (connectivity), (b) Proportion of thin fibers to thick fibers (fiber thickness), (c) Fiber thickness kurtosis (distribution of fiber thickness values with respect to the mean), (d) Pore circularity average value, (e) Pore size (mean of values superior to the 90th percentile)

GLOBAL STATISTICAL CHARACTERIZATION OF THE FN
PARAMETER MAPS

The graph-based FN fibers representation achieved following the pipeline presented in [Chapter 4](#), provides an appropriate setting for computing local characteristic features (e.g. connectivity, fiber thickness, median pore shape, etc.), such as illustrated in [Chapter 5](#). These features were proven to be meaningful for differentiating the four FN variants corresponding to "normal" state ECM.

Within this chapter, we are interested in the study of the parameter maps (e.g. fiber length) in a statistical framework based on the random field theory, that can provide a comparison between the "normal" and "tumour-like" state FN variants. Inspired from the methodologies employed in the statistical mapping of the functional images [[Fri+94](#); [Pol+97](#)], we intend to model the "normal" images through Gaussian Random Fields (GRF) and thus determine a set of probabilities that characterize a degree of belonging to the Gaussian field of certain regions of the tumoral images. More precisely, the purpose of the statistical analysis is to identify the foreign regions with respect to the GRF within both normal and tumoral parameter maps under the null hypothesis, and subsequently compare their properties (e.g. number, size).

Here, we implement two methodologies, the first one based on the theory of GRF, and the second relying on the computation of empirical distributions, and show in both cases the differences between the 2 classes quantitatively and qualitatively with respect to the fiber length, exemplified for one variant, FN B-A+.

6.1 GAUSSIAN RANDOM FIELDS AND DECISION TESTING OF PARAMETRIC
MAPS

DEFINITION RANDOM FIELDS: Given a complete probability space (Ω, \mathcal{F}, P) , and T a topological space, then a random field of real values is a measurable application $X : \Omega \mapsto \mathbb{R}^T$. The finite-dimensional (F_d) distributions of X are defined as the set of functions F_{t_1, \dots, t_n} , where $F_{t_1, \dots, t_n}(B) = P((X(t_1), \dots, X(t_n)) \in B)$, $\forall n > 0, \forall B \in \mathcal{B}^n$, \mathcal{B} being the Borel set on \mathbb{R} .

A particular class of random fields is represented by the Gaussian random fields, whose marginal (F_d) distributions are Gaussian vectors $X = (X_{(1)}, \dots, X_{(n)})$, characterized by the probability density function:

$$f_X(x) = (2\pi)^{-n/2} |V|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x - \mu)V^{-1}(x - \mu)^T\right] \quad (6.1)$$

where $\mu = (E(X_{(i)}))_{i \in [1, n]}$ is the expectation and V , the covariance matrix, $V = (E[(X_{(i)} - \mu_i)(X_{(j)} - \mu_j)])_{i, j \in [1, n]^2}$ is the covariance matrix.

Within the statistical analysis framework applied to parameter maps, we intend to model the images by approximating them as realisations of a GRF, thus we need to "Gaussianize" its marginal distributions to be sure the hypotheses under which we make this assumption are respected. In practice, we only Gaussianize the one dimensional marginal distribution and consider that the parameter maps under study are smooth enough.

STATISTICAL PARAMETRIC MAPS (SPM) are used to evaluate the probability of change in every pixel [Pol+97] by using decision tests based on the magnitude of the SPM values (i.e. the peak intensity of a cluster in SPM) and on the spatial extent of these clusters formed at a certain threshold. For our application, we consider that the parameter maps are described by the union of 2 classes of pixels: those that appear as the realization of a GRF modeling the "normal-state" case, and those that constitute foreign elements to the GRF. We expect these foreign elements to occur in regions with very high intensity and/or in larger clusters taken at a specific threshold. In the following, we briefly introduce the approach that allows us to estimate whether the clusters (i.e. contiguous regions of pixels - connected components) taken at a specific threshold of intensity, have a low probability of belonging to a GRF, based on the maximum intensity of the cluster, or its spatial extent. (Figure 6.1).

In order to estimate the likelihood of a certain cluster belonging to a GRF, depending on the maximal intensity of this cluster, we will use the following formulations taken from the theory of the random fields [Adl]. The expected (mean) value of the number of clusters that appear at a threshold t , of an image modelled by a zero-mean, homogeneous Gaussian field of dimension 2, is the following [Laf+07]:

$$E[m_t] = S(2\pi)^{-3/2} |\Lambda|^{1/2} t \sigma^{-3} \exp -\frac{t^2}{2\sigma^2} \quad (6.2)$$

where:

- m_t represents the number of clusters at a certain threshold t
- S is the number of pixels of the image
- Λ is the covariance matrix of partial derivatives of the GRF
- σ is the standard deviation of the GRF

Similarly, the mean value of the number of clusters at a threshold $t + H_0$ can be written as such:

$$E[m_{t+H_0}] = S(2\pi)^{-3/2} |\Lambda|^{1/2} (t + H_0) \sigma^{-3} \exp -\frac{(t + H_0)^2}{2\sigma^2} \quad (6.3)$$

Considering $x_0 = t + H_0$ as the intensity peak of a cluster (at threshold t), one can estimate the probability that a cluster (at a threshold t , having an intensity peak equal to x_0 , denoted $C_t^{H_0}$) belongs to G_r , a realization of GRF. This probability can

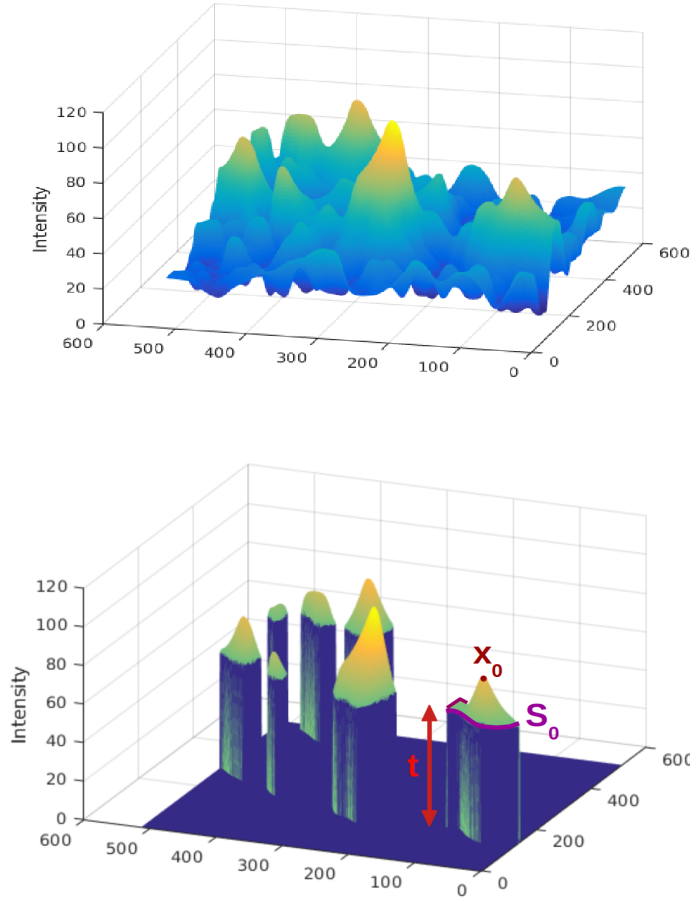


Figure 6.1: Connected components (clusters of pixels) of a 512×512 image (displayed in a 3D plot - top) taken at an intensity threshold t equal to 60 (bottom), where x_0 represents the maximum intensity of the cluster, and S_0 represents its surface (spatial extent).

be seen as the likelihood of a cluster (taken at threshold t) of having an intensity peak higher or equal to $t + H_0$:

$$P(C_t^{H_0} \in G_r) = \frac{E[m_{t+H_0}]}{E[m_t]} = \frac{x_0}{t} \sigma^{-3} \exp \frac{t^2 - x_0^2}{2\sigma^2} \tag{6.4}$$

Next, we are interested in the estimation of the probability that a cluster (at a threshold t) belongs to a realization of GRF, depending on its surface (spatial extent - number of pixels). To estimate the number of pixels (n_t) of a cluster at a threshold t , we use the following equation [Fri+94]:

$$E[n_t] = \frac{E[N_t]}{E[m_t]} \tag{6.5}$$

where N_t is the number of pixels at of higher intensity than t , and m_t is the number of clusters at the threshold t . Since the intensity values follow a normal (zero mean value) distribution, the expectation of N_t is the following:

$$E[N_t] = S \int_t^\infty (2\pi\sigma^2)^{-1/2} \exp -\frac{x^2}{2\sigma^2} dx = S\Phi(-t) \tag{6.6}$$

where $\Phi(-t)$ is the complementary cumulative distribution function (ccdf- tail distribution). It follows, then, that one can approximate the mean value of n_t , accordingly:

$$E[n_t] = \frac{E[N_t]}{E[m_t]} = \frac{\Phi(-t)}{(2\pi)^{-3/2}|\Lambda|^{1/2}t\sigma^{-3} \exp -\frac{t^2}{2\sigma^2}} \quad (6.7)$$

Furthermore, it has been experimentally proven [V.P69] that n_t follows an exponential distribution law, which is commonly defined by a parameter λ_t , (the inverse of the mean expected value of the random variable). Consequently, $P(n_t = x) = \lambda_t \exp(-\lambda_t x)$, where $\lambda_t = \frac{(2\pi)^{-3/2}|\Lambda|^{1/2}t\sigma^{-3} \exp -\frac{t^2}{2\sigma^2}}{\Phi(-t)}$.

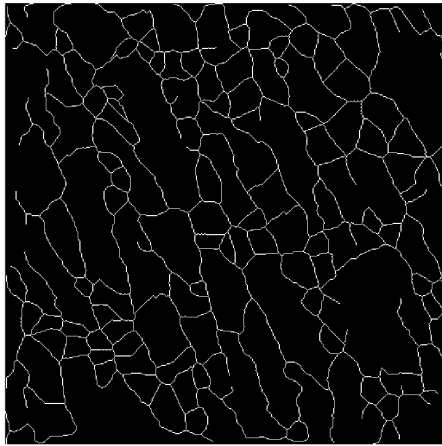
It follows then that the approximation for the probability of a given cluster having a spatial extent S greater than s_0 is given by the following formulation:

$$P(C_t^{S_0} \in G_r) = P(n_t \geq S_0) = \exp(-\lambda_t S_0) = \exp -\frac{(2\pi)^{-3/2}|\Lambda|^{1/2}t\sigma^{-3} \exp -\frac{t^2}{2\sigma^2}}{\Phi(-t)} \quad (6.8)$$

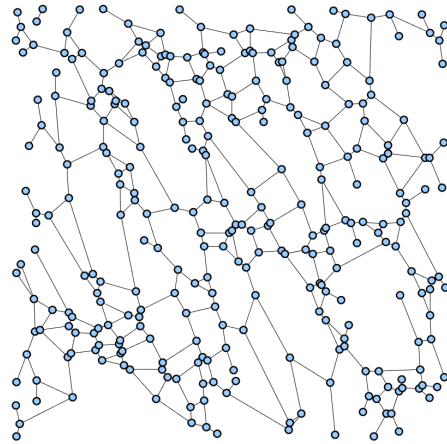
6.2 APPLICATION OF THE STATISTICAL ANALYSIS TO THE STUDY OF FN PARAMETRIC MAPS

Having defined the two probabilities of a cluster belonging to a GRF in (6.4), that will be denoted as P_H , and (6.8), denoted as P_S , we will illustrate an application to the statistical study of parametric maps (e.g. fiber length) derived from the graph-based representations of FN fibers, to further compare between normal state and tumoral state. Figure 6.2 illustrates the proposed workflow that can generate parametric maps of the FN fiber lengths:

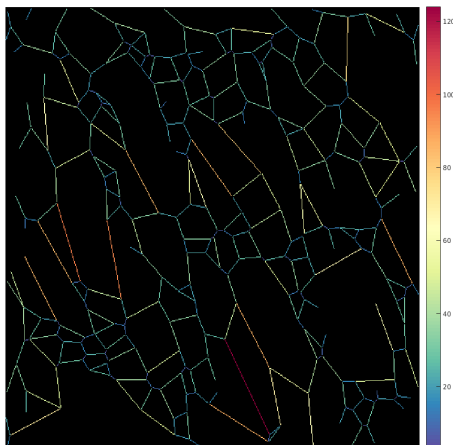
- For any FN confocal image: computation of the morphological skeleton using the pipeline defined within Chapter 4 .
- Simplification of the graph representation obtained by keeping the location of the nodes (corresponding to the fiber ends or to the fiber crosslinks) and replacing the "body" of the fiber from the skeleton (if present) with a connecting line. This representation can be subsequently encoded in adjacencies matrices (indicating the presence/absence) of a certain edge between the nodes and furthermore, will be employed in future chapters for FN fiber modelling and comparison of the graphs.
- Identification of the 2D pixels coordinates that approximate the straight line between the nodes (using Bresenham's line algorithm [Bre]); replacement of the pixels at the concerned locations with the length of the respective connecting line.
- Extrapolation of the values [Ext] of the fiber length map and smoothing with a Gaussian kernel.



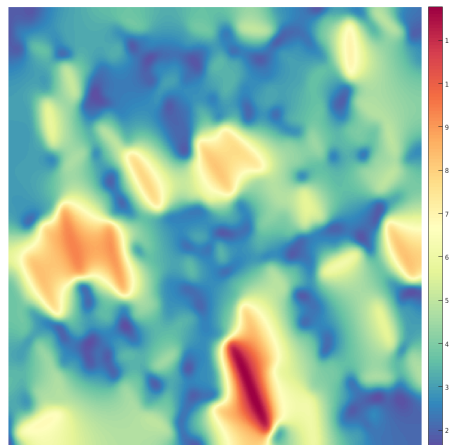
(a) Morphological skeleton of a "Tumour-like" FN B-A+ sample of 512×512 pixels obtained using the pipeline described in Section 4.2.1.



(b) FN fibers corresponding graph (defined as the set of nodes corresponding to fiber crosslinks or fiber ends. The edges between the nodes are represented as connecting lines.)



(c) Fiber length map associated to the graph representation: The lengths of the connecting lines are shown in different colours.



(d) Smoothed fiber length map

Figure 6.2: Computation of parameter (fiber length) maps from graph-based fiber representation: Top row illustrates the morphological skeleton (left) and associated graph (right) of a FN B-A+ sample. Bottom row shows the characteristic fiber length maps (the pixels intensity on the connecting lines is given by the specific line length- left image). Extrapolated values of the lengths are then smoothed out with a Gaussian filter to obtain the map on the right.

6.2.1 Statistical analysis based on GRF

In the first setup, the fiber lengths maps, characterized by the presence of clusters of different dimensions and intensities, corresponding to normal and tumoral-like FN, are approximated by Gaussian fields. This approximation requires that the marginal distributions of the GRF are Gaussian. Therefore, to render the images more appropriate to this setting, for sake of simplicity, we decided that the image intensity histogram is the only distribution variable to be Gaussianized. Thereby, the resulting histogram of intensity follows a normal distribution (centered on 0) with the same variance as the empirical native histogram.

The approach that was considered for map Gaussianization, is based on the optimal transport framework, described in detail in a following chapter (Chapter 8). More specifically, the problem of "converting" the empirical intensity distribution of the fiber length map into a normal distribution with the same variance, can be approached by performing the 1D optimal transport between the two distributions. The reader is referred to Section 8.2.3 for details on how the 1D optimal transport is accomplished.

Briefly, for a map I with an empirical mean and variance of intensity, we consider a second image J , whose intensity pixels follow a normal distribution (Figure 6.3). The 1D optimal transport problem will determine how to optimally permute the pixels indices in J , to "recreate" the image I . Since the intensity of the pixels in the permuted version of J does not change, the result will resemble I , but will have the histogram of J (Figure 6.4).

LEARNING PHASE: For all the images I_j (previously Gaussianized) in the learning set:

- Compute Λ_j (empirical estimator of the covariance of partial derivatives of I_j). If for an image function $f \in \mathbb{R}^2$, we consider its gradient vector $\nabla f = (f_x, f_y) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$, then $\Lambda_j = \text{cov}[f_x, f_y] = E[(f_x - E[f_x])(f_y - E[f_y])]$.
- Compute σ_j , as the I_j sample standard deviation.

Store the average Λ_m, σ_m of the learning dataset.

TEST PHASE: - decision on whether the clusters (blobs) taken at a certain threshold t_i are considered foreign to the GRF:

For all the images I_j in the test set:

- Gaussianize the image I_j , in order to make it more feasible to be approximated by a realization of a GRF. The result is a new image I_g , whose histogram is Gaussian with identical variance to that of I_j .
- For a given list of thresholds $T = (t_1, t_2, \dots, t_n)$:
 - Binarize the image I_g according to the threshold t_i
 - Find the list of connected components in the binary image resulted from thresholding and for every (labeled) connected-component (l_1, l_2, \dots, l_p) :

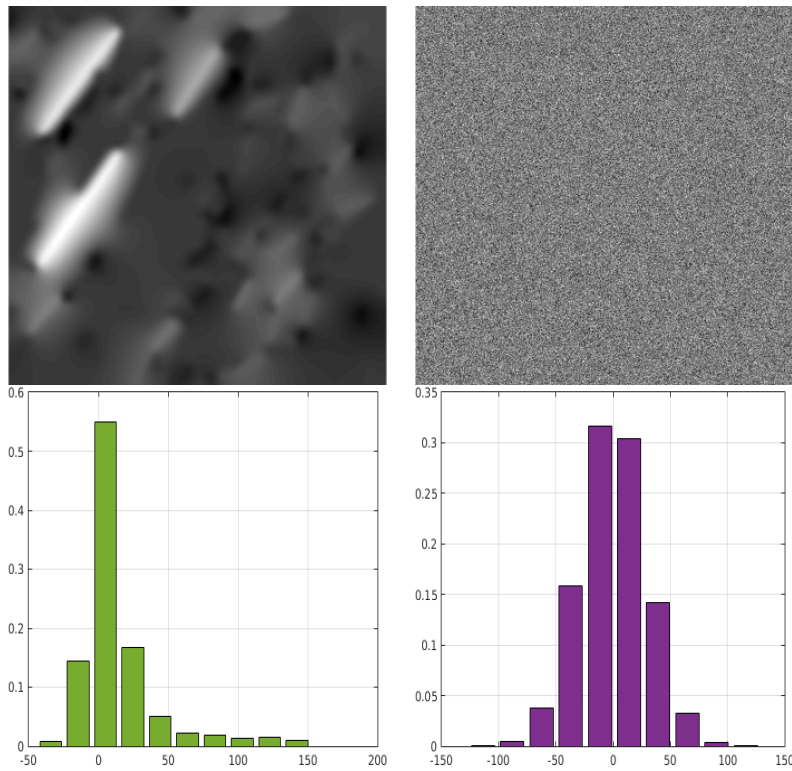


Figure 6.3: Gaussianization of the parametric maps using optimal transport between the empirical intensity distribution of the map (top row, left) and a Gaussian noise image (top row right), whose intensity distribution centered on 0, has the same variance as the first image.

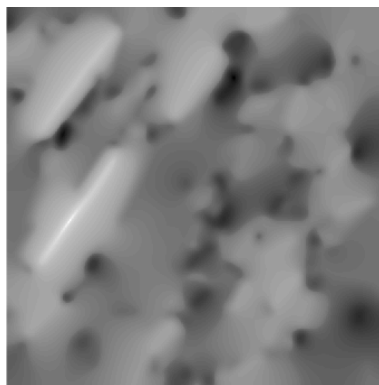


Figure 6.4: Result of Gaussianization of the parametric map in [Figure 6.3](#). The native histogram of the map has been converted to a normal distribution defined by the same variance as the original one, (and mean equal to 0).

- * Compute P_H using the learnt model parameter σ_m . If its value is lower than a chosen p-value (p), the cluster is considered a foreign element to a GRF.
- * Compute P_S using the learnt model parameters Λ_m, σ_m . If its value is lower than a chosen p-value (p), the cluster is considered a foreign element to a GRF.

The methodology described above was applied for the quantitative and qualitative comparison of differences with respect to the fiber length, for one variant FN B-A+, in both normal and tumoral-like state, under the following conditions:

- Learning dataset : 50 parametric maps (Normal-like FN networks)
- Test set: 70 parametric maps (Tumoral) and 20 parametric maps (Normal)
- Thresholds of intensity $T = [45 \ 60 \ 70 \ 80 \ 100]$
- $p = 0.05$. Clusters are considered as foreign to GRF if either P_S or P_H are less or equal than p .

The results in [Table 6.1](#), [Table 6.2](#) illustrate for every threshold, the average number of identified foreign clusters, as well as the average cluster area per image. Within the tumoral parametric maps of the test dataset, the method identifies a higher average number of clusters per image having a higher average spatial extent, in comparison to the normal parametric maps. The results in [Figure 6.5](#), [Figure 6.6](#) illustrate several example of detected clusters for a given $p = 0.05$.

Table 6.1: Average number and area of clusters per tumoral parametric map FN B-A+ 512×512 identified as foreign to a GRF ($p=0.05$) based on either the maximum intensity or cluster surface, taken at various thresholds

Thresholds	45	60	70	80	100
Max Intensity - Avg nb/im	1.00	0.80	0.46	0.33	0.16
Max Intensity - Avg area/im	2595.46	1063.85	728.68	460.55	196.91
Surface- Avg nb/im	1.30	0.99	0.69	0.39	0.20
Surface - Avg area/im	3939.29	3225.60	2119.56	2306.27	1562.82

Table 6.2: Average number and area of clusters per normal parametric map FN B-A+ 512×512 identified as foreign to a GRF ($p=0.05$) based on either the maximum intensity or cluster surface, taken at various thresholds

Thresholds	45	60	70	80	100
Max Intensity - Avg nb/im	0.2	0.1	0.1	0.05	0
Max Intensity - Avg area/im	159.65	10.4	1.15	0.05	0
Surface- Avg nb/im	0.4	0.15	0.1	0.05	0
Surface - Avg area/im	161.6	11.55	0.95	0.1	0

6.2.2 Statistical analysis of the empirical distributions

The second setup for the statistical analysis and comparison of the FN variants in normal vs tumoral state was motivated by the fact that upon Gaussianization of the parametric maps (during the previous methodology), the higher intensity clusters are smoothed out. Since Gaussianization is necessary when approximating the parametric maps as realizations of GRF, we decided to take a different approach to estimate the probabilities of clusters taken at different thresholds to be foreign elements with respect to the maximum cluster intensity and surface.

The approach described below will compute, for a given threshold t , the empirical cumulative histogram of maximum cluster intensities/surfaces for all the images in the learning set. This, in turn, will provide a certain threshold regarding either the cluster area (S_p) or the cluster intensity (C_p) that depends on the chosen p -value, above which the clusters from the test set taken at threshold t , if they exist, will be considered as foreign elements.

LEARNING PHASE: For a given list of thresholds $T = (t_1, t_2, \dots, t_n)$:

- For all the images Im_j in the learning set:
 - Binarize the image Im_j according to the threshold t_i
 - Find the list of connected components in the binary image resulted from thresholding and for every (labeled) connected-component (l_1, l_2, \dots, l_m) :
 - * compute and store the area (total number of pixels) of l_k in a vector S
 - * compute and store the maximum intensity of l_k in a vector C
- compute the cumulative histogram of S as $Q(s)$ and set the area threshold S_p according to a predefined p -value (p): $S_p = \operatorname{argmin}_s\{Q(s) \geq p\}$. Store the resulted S_p for the current t_i .
- compute the cumulative histogram of C as $Q(c)$ and set the intensity threshold C_p according to a predefined p -value (p): $C_p = \operatorname{argmin}_c\{Q(c) \geq p\}$. Store the resulted C_p for the current t_i .

TEST PHASE: decision on whether the clusters (blobs) taken at a certain threshold t_i are considered foreign to the empirical distributions of cluster surfaces and maximum intensities. For a given list of thresholds $T = (t_1, t_2, \dots, t_n)$:

- For all the images I_j in the test set:
 - Binarize the image I_j according to the threshold t_i
 - Find the list of connected components in binary image resulted from thresholding and for every (labeled) connected-component (l_1, l_2, \dots, l_p) :
 - * compute the area (total number of pixels) of l_k and compare it to the already stored S_p (at t_i , during the learning phase). If its value is at least as high as the stored one, the cluster is considered a foreign element.

- * compute the maximum intensity of l_k and compare it to the already stored C_p (at t_i , during the learning phase). If its value is at least as high as the stored one, the cluster is considered a foreign element.

The methodology described above was applied for the quantitative and qualitative comparison of differences with respect to the fiber length for one variant FN B-A+ in both normal and tumoral-like state, under the following conditions:

- Learning dataset : 50 parametric maps (Normal)
- Test set: 70 parametric maps (Tumoral) and 20 parametric maps (Normal)
- Thresholds of intensity $T = [95 \ 105 \ 115 \ 120 \ 127]$
- $p = 0.05$ and 0.1 .

The results in [Table 6.3](#) illustrate for every threshold, the average number of identified foreign clusters per image as well as the average cluster area. By comparison, within the normal parametric maps of the test dataset, the method does not identify any foreign cluster when $p = 0.05$, however it manages to identify foreign clusters (relatively low average number and area per image) for one given threshold upon increasing the p-value $p = 0.1$ ([Table 6.4](#)).

Table 6.3: Average number and area of clusters per tumoral parametric map FN B-A+ 512×512 identified as foreign ($p = 0.05$) to the empirical distributions of clusters (size and intensity) taken at various thresholds

Thresholds	95	105	115	120	127
Max Intensity - Avg nb/im	0.31	0.31	0.36	0.37	0.40
Max Intensity - Avg area/im	6913.33	5761.59	2777.69	2330.17	1723.45
Surface- Avg nb/im	0.04	0.09	0.13	0.23	0.37
Surface - Avg area/im	4922.07	4935.79	3019.36	2529.45	2075.28

Table 6.4: Average number and area of clusters per normal parametric map FN B-A+ 512×512 identified as foreign ($p = 0.1$) to the empirical distributions of clusters (size and intensity) taken at various thresholds

Thresholds	95	105	115	120	127
Max Intensity - Avg nb/im	0.1	0	0	0	0
Max Intensity - Avg area/im	458.8	0	0	0	0
Surface- Avg nb/im	0	0	0	0	0
Surface - Avg area/im	0	0	0	0	0

6.3 CONCLUSIONS

In this chapter we illustrated an analytic framework for the study of FN parametric maps (namely the fiber length), in order to compare normal and tumor-like FN

states. We managed to show, using both approaches (based on the GRF theory and on the computation of empirical distributions), that the tumoral aspect can be differentiated with respect to the normal-like state for one specific FN variant, and statistically characterized, based on the fiber length. The quantitative analysis shows that "normal" variants are well characterized by the both the GRF and empirical model, and that differences with respect to the fiber length, between "normal" and "tumoral" variants are statistically significant. These are promising results that encourage us to extend this analysis for normal vs tumour FN-variant comparison.

Future perspectives can include a more comprehensive study that analyses larger image samples and considers different FN fiber parameters.

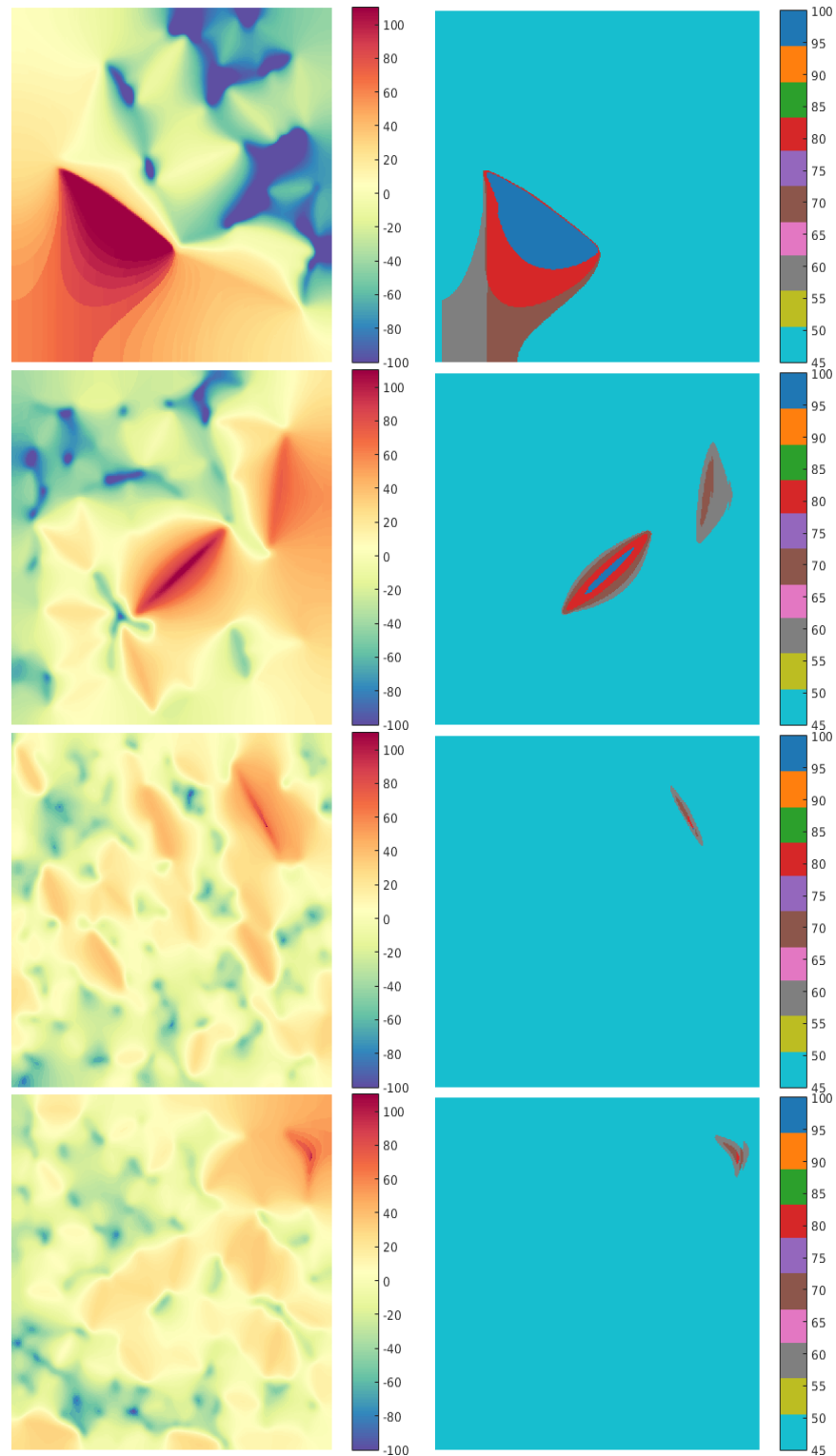


Figure 6.5: Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the maximum cluster intensity), within the Fiber Length Map (left column) corresponding to Tumour-like FN B-A+ of 512×512 . The right columns depicts the clusters at different thresholds (indicated in the colorbar).

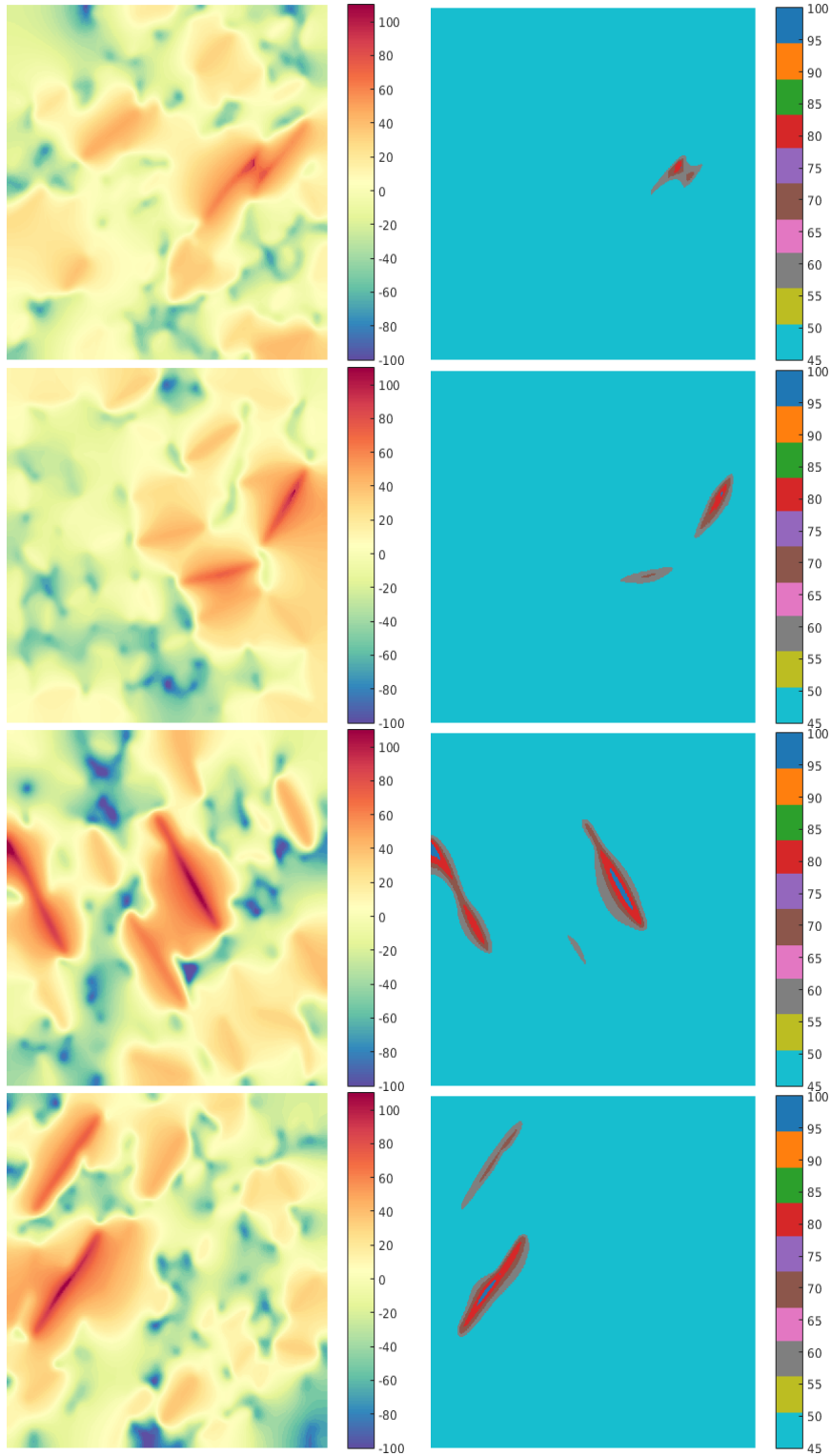


Figure 6.6: Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the cluster surface), within the Fiber Length Map (left column) corresponding to Tumour-like FN B-A+ of 512×512 . The right columns depicts the clusters at different thresholds (indicated in the colorbar).

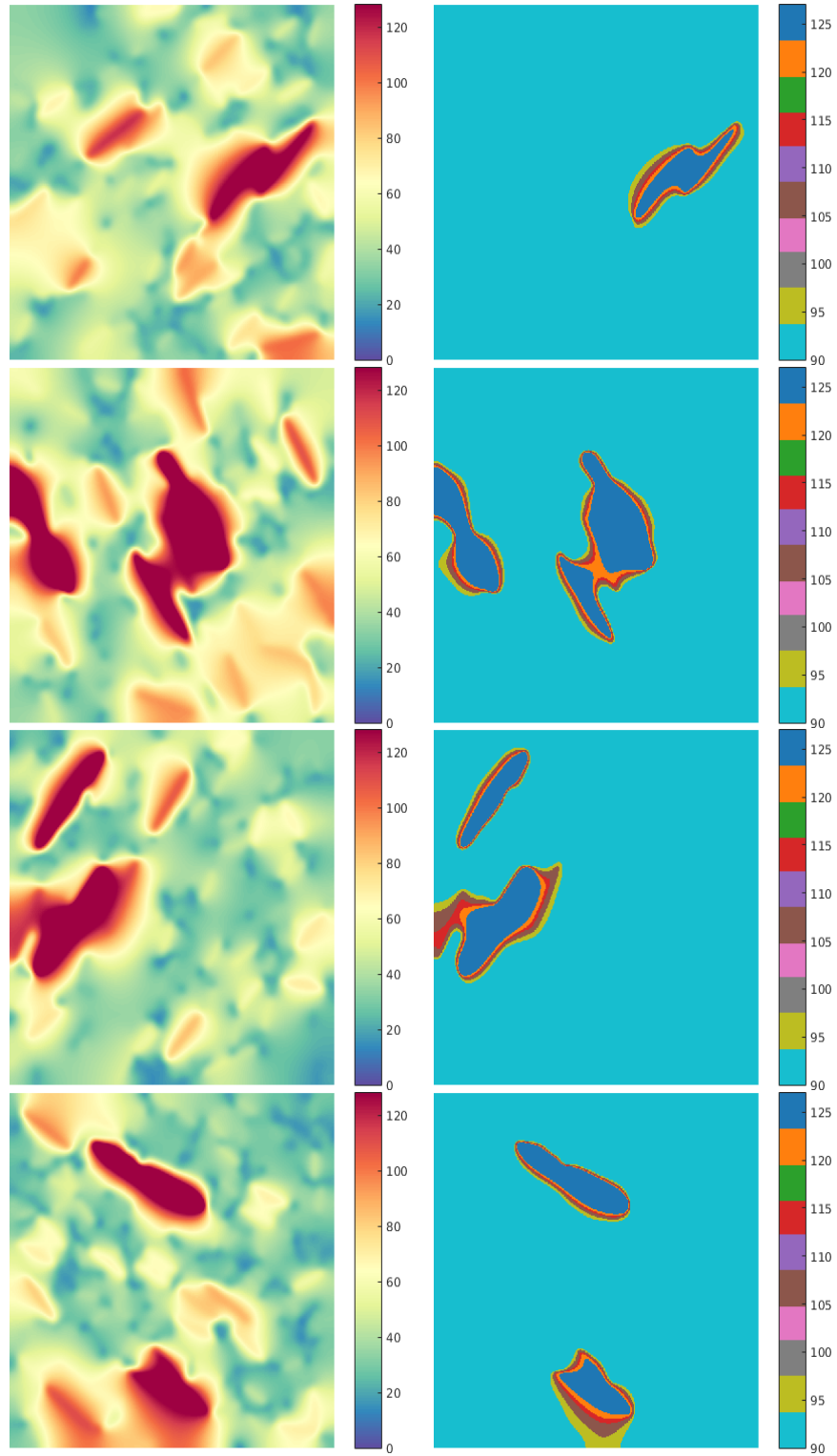


Figure 6.7: Detection of the clusters considered foreign elements to the empirical distributions of maximum cluster intensity $p = 0.05$ within the Fiber Length Map (left column) corresponding to Tumour-like FN B-A+ of 512×512 . The right columns depicts the clusters at different thresholds (indicated in the colorbar).

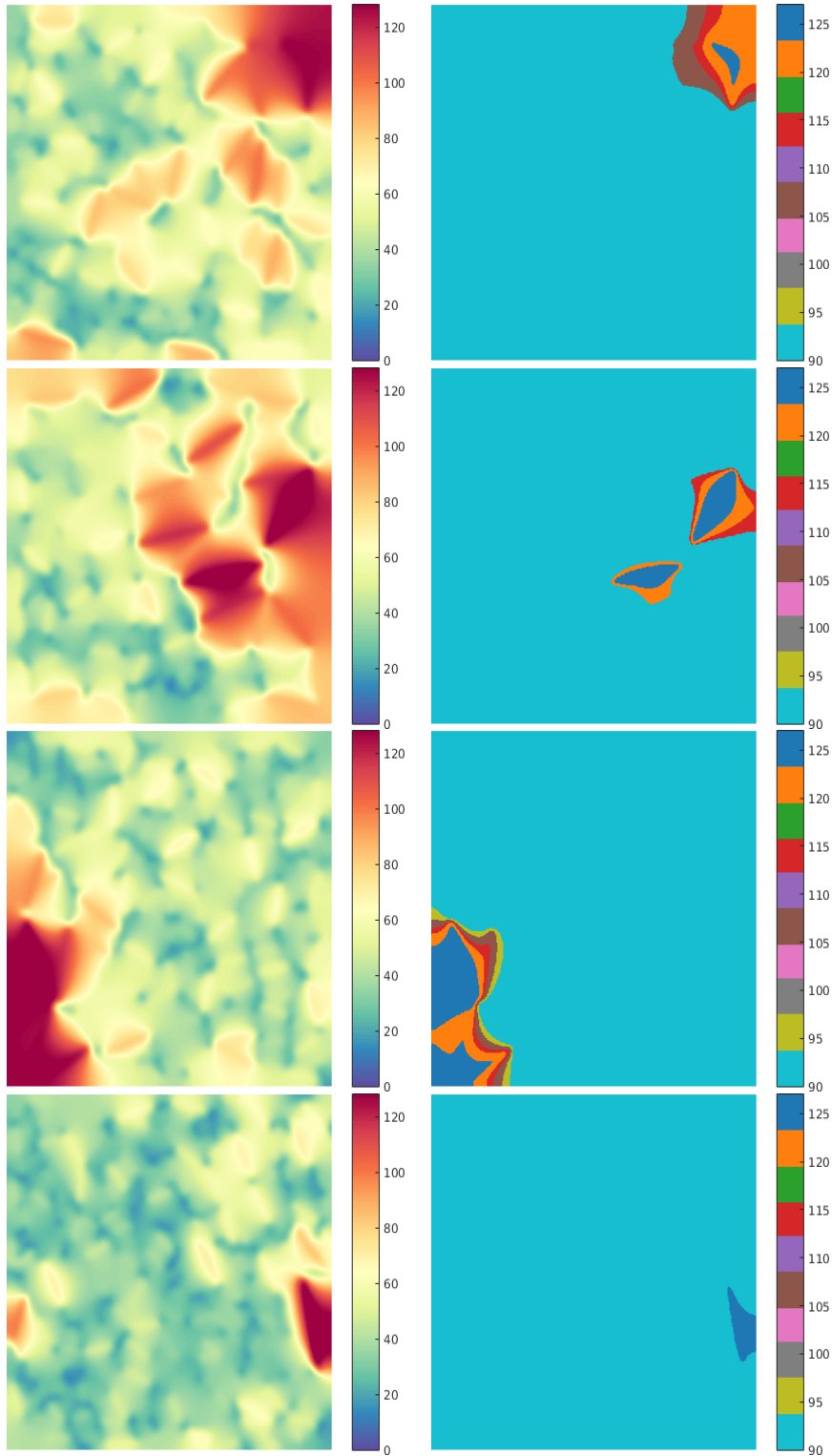


Figure 6.8: Detection of the clusters considered foreign elements to the empirical distributions of cluster surface $p = 0.05$ within the Fiber Length Map (left column) corresponding to Tumour-like FN B-A+ of 512×512 . The right columns depicts the clusters at different thresholds (indicated in the colorbar).

Part V

TOWARD MODELLING

GRAPH MATCHING FOR GRAPH COMPARISON

In this chapter, we present the general background concerning the graph-matching approaches for graph comparison. We focus on the presentation of inexact graph matching techniques, and more precisely on the discrete problems reformulated through continuous relaxations. Finally, we describe a relevant approach, namely, many-to many assignment framework, proposed by [ZBV10], that we have selected for the use in certain applications related to the matching of FN-specific networks, within subsequent chapters.

7.1 GRAPH MATCHING GENERAL BACKGROUND

While the problem of comparing various structures via graphs was known from the 50s in modern chemoinformatics, the use of graphs to characterize or classify complex networks (such as visual patterns), dates back to the 70s [Ven15]. Introduced in order to compensate the weaknesses of the vector-based representations, which are not always adequate to capture patterns with an identifiable structures of different sizes, nowadays, there is a variety of fields that rely on modelling specific problems using graphs (i.e. pattern recognition [Con+07], computer vision [BBV01], network monitoring [SK99], computational biology [YS07; Lia+09], etc.). This long interest in graph-based approaches is not surprising, given the range of different representations in terms of the graph topology, the nature of the nodes and edges (deterministic or stochastic), the type of the attributes (numeric values, symbols, probabilities) [Con+07; Yan+16].

Over the years, the advances in graph-based approaches have focused on a few directions with specific dedicated algorithms such as graph-matching, graph embedding, clustering, or learning. Often across the domains, graphs encode sets of features connected by structural relationships. Most of the applications seek to understand how similar these objects are, or whether equivalent patterns can be found within them. In this regard, the question faced by the graph-based algorithm community, when given two objects represented using graphs, is how best to evaluate the distance between them in a way that reflects their structure in order to either classify them, learn a model representative of a class based on graphs, or to study the variation of certain parameters associated to the features.

This is commonly known as the *graph-matching problem*. Once a graph is associated to the identifiable parts of those objects, comparing them can be achieved by coupling their nodes accordingly, so that similar structures can be associated together. It then becomes essential to develop the proper tools to achieve an accurate matching between two graphs that reflects the structural differences.

Evaluating the alignment or matching between graphs is considered challenging, especially for large graphs, and hence approximate algorithms have been developed to estimate the solution of these problems (hard-combinatorial problems) in a reasonable time.

In order to define the framework of graph-matching, we start by illustrating the following definitions and notations:

NOTATIONS:

Let G be a *graph*, where $G = (V, E)$ is a set of nodes (vertices) connected by the set of edges $E \subset V \times V$. The structure of G can be encoded in a square adjacency matrix, A_G of size $|V| \times |V|$, where $(A_G)_{ij}$ is equal to 1 if node i is connected by an edge to node j , and 0 otherwise, also called a binary adjacency matrix.

We refer to *real-valued adjacencies matrices (weighted)* if $(A_G)_{ij}$ represents the weight assigned to the edge between node i and j .

G is called an *undirected* graph when A_G is a symmetric matrix, i.e. $(A_G)_{ij} = (A_G)_{ji}$.

We refer to the *matching* between 2 graphs as the mapping that denotes the assignment between the nodes: $f : V^G \rightarrow V^H$. Denoting by $N_G = |V_G|$ and by $N_H = |V_H|$, the number of nodes of G and H , respectively, the assignment can be encoded into a binary correspondence matrix $P \in \{0, 1\}^{N_G \times N_H}$, such that $P_{i,j} = 1$ when the i -th node of G and the j -th node of H are matched and 0, otherwise.

ALGORITHMIC ASPECTS AND NOTIONS

- *Combinatorial optimization* is a subset of mathematical optimization whose role is to search for maxima (or minima) of an objective function whose domain is a discrete but with large solution space where exhaustive search is not commonly tractable.
- *Hungarian algorithm* is an instance of an assignment problem (e.g. assigning tasks to agents such that the total cost of the assignment is minimized, knowing that each job is assigned to one worker and each worker is assigned one job) [Hun].
- *Linear programming* [Lin] is a method employed for the modeling of various types of problems (planning, scheduling, assignment) that seeks to optimize a linear objective function under linear equality and linear inequality constraints, formalized as:

$$\text{maximize } c^T x \quad \text{s.t.} \quad Ax \leq b \quad \text{and} \quad x \geq 0 \quad (7.1)$$

where x is the variable vector, c, b represent vectors, A is a matrix of *known* coefficients.

- *Sinkhorn-Knopp algorithm* [Sin] is an iterative method that shows that one can obtain a double stochastic matrix ¹ by alternating the rescaling of the rows and columns of a nonnegative matrix.
- In computational complexity theory, NP (nondeterministic polynomial time) represents the set of decision problems solvable in polynomial time by a non-deterministic Turing machine. Non-deterministic machines are idealized models of computation that have the ability to make perfect optimization.

7.1.1 Exact and inexact matching

A broad classification of the graph-matching problem distinguishes between two main types of approaches: exact and inexact matching techniques. The exact graph matching problem seeks the bijective node mapping that preserve the edges of both graphs with zero distortion. It implies that there exists an exact correspondence between the nodes and edges of the compared graphs. The term of *graph isomorphism* is used to encompass the latter concept, namely whether two graphs are structurally identical. It is known as a NP problem [GJ90], and several works have proposed algorithms to solve the problem in polynomial time [Cor+04; Ull76]. An example is illustrated in Figure 7.1, where the challenge is to determine that the two given graphs with different node numbering, possess in fact, an identical structure.

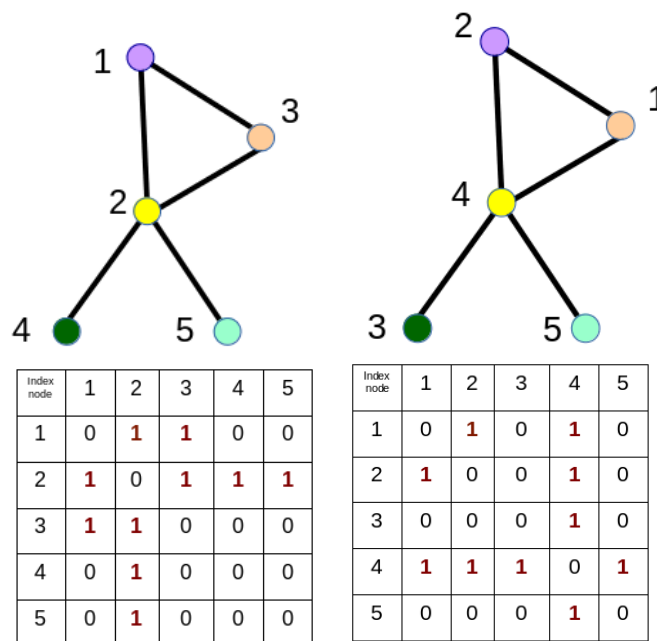


Figure 7.1: Graph matching - Exact matching (Isomorphism) case: top row: two isomorphic graphs, bottom row: adjacency matrices corresponding to the graphs above. Even if the graphs are structurally identical, their adjacency matrices are different due to the different numbering.

¹ A double stochastic matrix is a square matrix of nonnegative real numbers, with the property that each of its rows and columns sums to 1.

Alternatively, *inexact graph matching* problems focus on finding a matching cost as an indicator of similarity between graphs that are structurally different.

7.1.1.1 Edit distance

A class of approaches (optimal inexact matching) computes the edit distance [SF83] to measure the distance between two graphs. The edit distance is the set of graph edit operations with an assigned cost that consists of deletion, insertion and substitution of nodes/edges. Hence the objective is to find the sequence of operations that minimizes the cost of "converting" one graph to another (Figure 7.2). These kind of minimization problems (hard combinatorial) are approached by searching for the solution using a tree search [TF79] in the possible solutions space, e.g. beam search [NRB06], breadth-first, depth-first search, etc. The inconvenient of these methods is the computational complexity, exponential in the number of vertices of involved graphs, which makes it feasible in practice only for reasonably sized graphs.

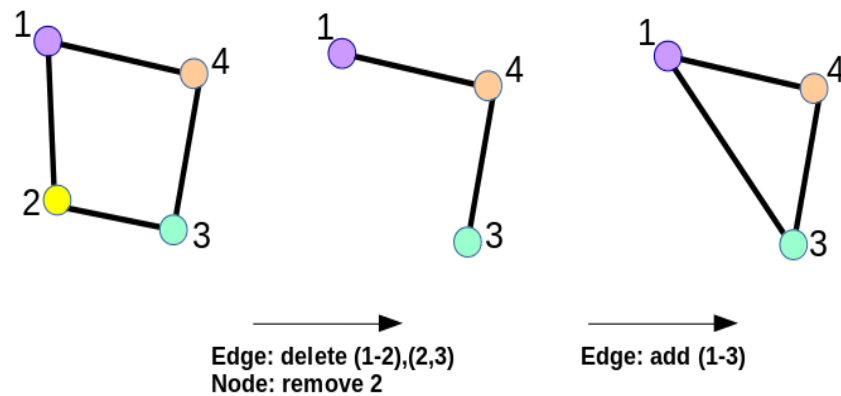


Figure 7.2: Graph matching - Edit Distance obtained through the optimal sequence of the operations (e.g. edge/node insertion/deletion) to "transform" one graph into another

7.1.1.2 Spectral methods

The majority of algorithms for inexact graph-matching are sub-optimal that ensure to reach a local minimum of a cost function and have a lower computational cost [Ven15]. Some of these techniques rely on the reformulation of the discrete optimization problem to a continuous optimization problem, while alternative methods are based on spectral properties of the graphs (i.e. properties related to the eigenvalues and eigenvectors of the adjacency matrix or of other matrices characterizing the graph structure).

Regarding the spectral methods, the underlying concept is based on the fact that the eigenvalues and eigenvectors of the adjacency matrix of a graph are invariant with respect to the node permutations. Matching is done between the nodes of similar spectral coordinates, i.e. rows of eigenvector matrices. Their computation is performed in polynomial time [Ven15], however, a disadvantage of these methods is the fact that the spectral embedding is not uniquely defined [Zas+10]. Some of

the papers proposing algorithms for graph matching using spectral decomposition of the adjacency matrices of the graphs are found in [Ume88; CK04].

7.1.1.3 Continuous relaxation of the objective function

Different approaches rely on the reformulation of the graph matching problem, a discrete optimization one, into a continuous optimization problem that can be efficiently solved. Once the solution is approximated in the continuous domain, it has to be projected back into the discrete domain. Even if the algorithms ensure a local optimum, due to the discretization step, the local optimality is not necessarily guaranteed.

We note here two important contributions in the literature of weighted graph matching, the problem of finding the matching matrix between the nodes of two given graphs, where the objective is to optimize a function over this matrix so as the sum of weights of the preserved edges is as high as possible.

The work in [AD93] proposes to linearize the objective quadratic function to solve it with linear programming methods and then convert the approximate solution back to the discrete domain form using the Hungarian assignment method [Hun] for the assignment problem. Secondly, in [GR96], they employ methods inspired from deterministic annealing to relax the discrete problem into a continuous one, and use the efficient Sinkhorn's algorithm [SK67] to ensure the constraints on the matching (permutation) matrix.

The inexact graph matching problem (one-to-one matching) can be formulated as a quadratic assignment problem [Loi+07; Law63], a NP-hard problem through which combinatorial optimization problems (e.g travelling salesman) can be formulated. Several works based on this formulation can be found at [GR96; ZDLT]. Being a difficult combinatorial problem, the research has been devoted to creating the appropriate algorithms to approximate the solution, which is a rather challenging task due to the non-convexity of the objective function. A recent state of the art paper [ZDLT] improves this formulation by incorporating geometric constraints between nodes and by using better optimization strategies.

7.1.2 Graph kernels and graph embeddings

Alternative approaches to graph comparison are based on the so-called graph embeddings and graph kernels. *Graph embeddings* (Figure 7.3) refer to techniques that map the nodes of the graphs onto points in a vector space, in such a way to preserve the structure (e.g. nodes that are found closer in a neighborhood will be mapped to points that are close.) [FPV14]. In [LWH03], they resorted to spectral graph methods and PCA or multi-dimensional scaling (MDS - whose purpose is to reconstruct a map that preserves distances, given pairwise dissimilarities) based embedding of the spectral features for graph clustering with the purpose of exploring different patterns in graphs. In [BYH04], the authors have applied MDS to a matrix of shortest distance between the graphs to embed them in an Euclidean space and then perform the matching using simple point-pattern matching methods.

Graph kernels are functions that map graphs onto a real number (similar properties with the dot product on vectors). The use of such kernels as a measure of similarity

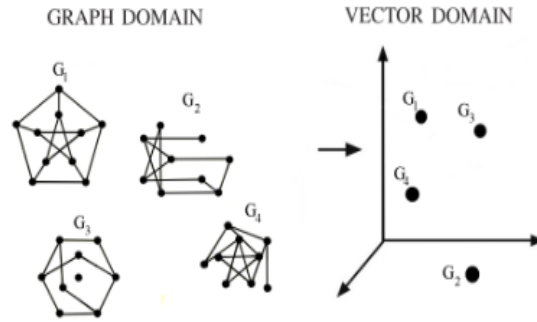


Figure 7.3: Graph embedding into a vector domain. Figure reproduced from [FPV14]

(positive-semidefinite), provides the access to classification tools (support vector machine), clustering, principal component analysis.

There is a growing interest in these methods, for learning, classification, or clustering problems or generally for pattern recognition. However, unlike graph-matching, these approaches are more suitable for applications where there is no direct need of knowing which part of a graph was matched to given nodes of another graph.

7.2 FORMULATION OF THE GRAPH-MATCHING PROBLEM

7.2.1 General matching

If the graph isomorphism problem verifies whether two given graphs are identical, we are interested here in a more realistic setting that can evaluate how different/similar graphs are. Generally, this problem is formulated as an *assignment task*, where in order to determine how similar two given graphs are, we seek to align them, looking for a matching (assignment) between the nodes that achieves this as closely as possible (see Figure 7.4). The graph-matching (GM) problem is generally classified into two general categories: exact matching and inexact matching [Yan+16].

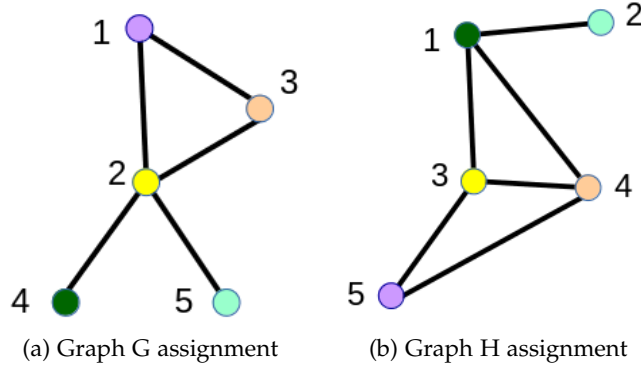
The first category groups the methods that require a bijective node mapping between the two graphs, such as to preserve the edges with no distortion. Less adapted to real settings, where exact matching might be difficult, or even impossible to achieve, and computationally expensive (NP problem), our study was hence motivated to focus on the second category of problems. More specifically, we consider the class of inexact GM, tailored to real-world graphs where we allow a less strict correspondence of the graph nodes.

As already outlined, the graph-matching problem between two given graphs implies finding the assignment between the nodes of the graphs that aligns them as closely as possible. To formalize this definition, we have introduced the concept of a correspondence matrix P that encodes the matching between the graphs G, H , such that $P_{i,j} = 1$ when the node G_i and the node H_j are matched and 0, otherwise.

The least square formulation can be shown to be equivalent to solving an instance of quadratic assignment problem formulated as :

$$\max_P \text{tr}(G^T P H P^T)$$

subject to $P \in \mathcal{P}$
[Zas+10].



Index node	1	2	3	4	5
1	0	0	0	0	1
2	0	0	1	0	0
3	0	0	0	1	0
4	1	0	0	0	0
5	0	1	0	0	0

(c) Matching matrix P

Figure 7.4: Graph matching principle colour coded: Nodes that are assigned together have the same colour. The matching is represented in a binary matrix whose element ij is 1 if node i of graph G is matched to node j of graph H .

For the sake of simplicity, we assume that the two graphs have equal size, given by the nodes number, denoted by N . Thus, $P \in \{0, 1\}^{N \times N}$, is a permutation matrix, indicating the matching with exactly one entry 1 on each row and column.

To further illustrate the *general graph matching framework*, we apply the permutation matrix P to the second graph H , hence obtaining a permuted graph, isomorphic to H , whose adjacency matrix is given by PHP^T (Figure 7.5).

Subsequently, we can evaluate the quality of the alignment resulted from P , by computing the discrepancy between G and the adjacency matrix of the permuted graph H . We denote $\|\cdot\|_F$, the Frobenius norm of the matrices (defined as $\|A\|_F^2 = \text{tr}A^T A = (\sum_i \sum_j A_{ij}^2)$), and compute the measure of discrepancy between the two graphs after alignment as introduced in [Ume88] - least square formulation, as follows:

$$F(P) = \|G - PHP^T\|_F^2 \tag{7.2}$$

Understandably, the lower the discrepancy, the better the quality of the matching is. This leads us to formulate the problem of graph matching as finding the optimal permutation that minimizes the discrepancy between the graphs, as computed as in (7.2). This combinatorial problem requires approximate methods to derive its solutions, as it becomes unfeasible for large graphs. If we consider binary adjacency matrices, then the quantity $0.5 \cdot F(P)$ represents the number of non-overlapping edges, regardless of the matrix norm employed $l_p (1 \leq p \leq \infty)$ [Zas+10].

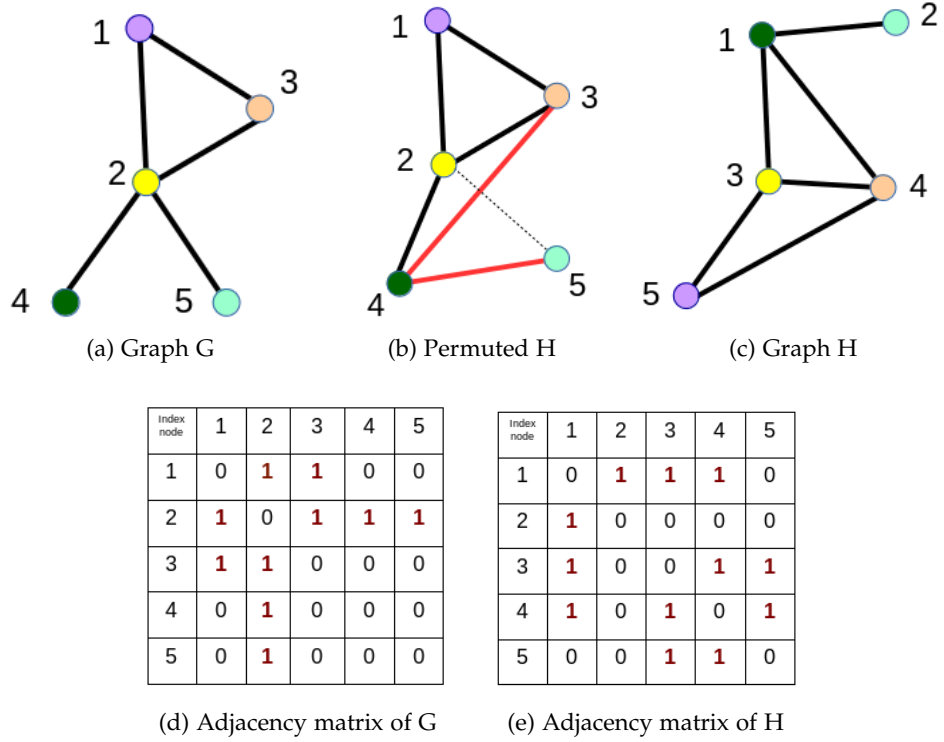


Figure 7.5: Graph matching (to compare G and H), measures the discrepancy between one graph (G) and the optimally permuted version of the second one (H).

7.2.2 Different instances of graph matching

There are different scenarios requiring the assignment to be either a one-to-one correspondence between the graphs, or allowing multiple nodes to be matched to another node (group of nodes). We mention the following cases:

one-to-one matching: We consider G and H the adjacency matrices of two graphs, with N_G, N_H the number of nodes for each graph, respectively. Under one-to-one constraints, $N_G = N_H \stackrel{\text{def}}{=} N$ and the matching matrix becomes a permutation matrix.

The feasible set \mathcal{P} , composed of all permutation matrices is defined as follows: $\mathcal{P} = \{P \in \{0, 1\}^{N \times N} : P\mathbf{1}_N = \mathbf{1}_N, P^T\mathbf{1}_N = \mathbf{1}_N\}$, where $\mathbf{1}_N$ represents the constant N -dimensional vector of all-ones.

Then, the graph matching problem in the one-to-one mapping framework, is defined as a discrete optimization problem whose objective is to find the permutation matrix $P \in \mathcal{P}$ for:

$$\min_{P \in \mathcal{P}} \|G - PHP^T\|_F^2, \quad \text{s.t.} \quad P\mathbf{1}_N = \mathbf{1}_N, \quad P^T\mathbf{1}_N = \mathbf{1}_N \quad (7.3)$$

One possible solution to account for the difference of sizes between G and H is to insert *dummy nodes*, i.e. nodes without any connection to any other nodes in either of the graphs, which formally translates to adding rows and columns of zeros in the adjacency matrix.

In practical applications, the concept of one-to-one matching can be restrictive, either for situations where the dimensions of the graphs to be matched is different, or for the situations where similar parts are represented through different number of nodes in the two graphs. In those cases, it would be helpful to relax the constraints of the original one-to-one matching problem, to allow multiple nodes from one graph to be matched to one (many-to-one matching) or more (many-to-many matching).

many-to-one matching The formulation for assigning at most k_{max} nodes from G to H in a many-to-one matching formulation (Figure 7.6) is based on the optimization problem illustrated in Equation 7.3. We consider two undirected graphs represented by their real-valued adjacency matrices G and H of size $N_G \times N_G$ and $N_H \times N_H$ respectively. The objective is to find the permutation matrix $P \in \{0, 1\}^{N_G \times N_H}$ under the following constraints:

$$\{P^T \mathbf{1}_{N_G} \leq k_{max} \mathbf{1}_{N_H}, \quad P \mathbf{1}_{N_H} = \mathbf{1}_{N_G}, \quad P^T \mathbf{1}_{N_G} \geq \mathbf{1}_{N_H}\} \tag{7.4}$$

where each constraint refers to the manner in which nodes can be matched together, k_{max} representing the upper bound on the number of nodes of graph G , that can be matched to a single node in the graph H .

- $P^T \mathbf{1}_{N_G} \leq k_{max} \mathbf{1}_{N_H}$: at most k_{max} nodes of G can be matched to a single node of H .
- $P \mathbf{1}_{N_H} = \mathbf{1}_{N_G}$: all nodes belonging to G have to be matched to nodes of H , but a single one of G can not be matched to multiple H nodes.
- $P^T \mathbf{1}_{N_G} \geq \mathbf{1}_{N_H}$: all H nodes have to be matched to G nodes.

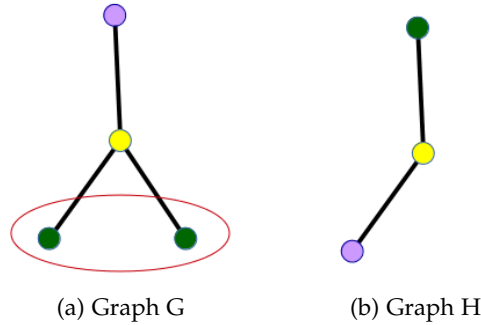


Figure 7.6: Many-to-one assignment between two graphs G and H , where multiple nodes (here 2) of G can be assigned to the same node in H .

7.2.3 Many-to-many-assignment problem

The problem is defined in the following manner: the objective is to find the matrices $P_1 \in \{0, 1\}^{N_k \times N_G}$ and $P_2 \in \{0, 1\}^{N_k \times N_H}$ (which can be regarded as matching matri-

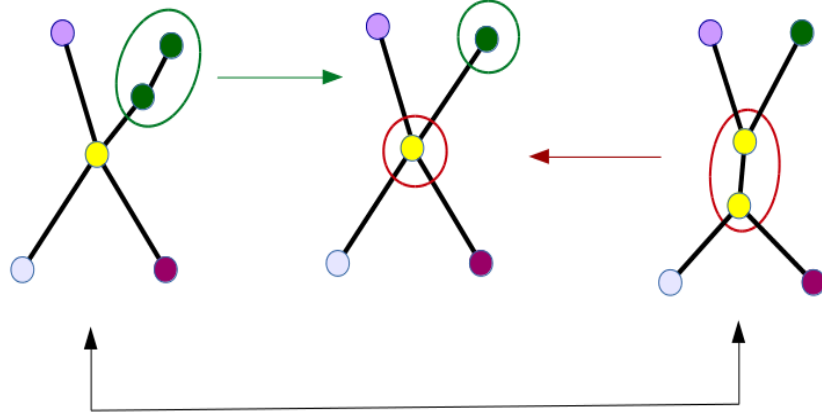


Figure 7.7: Many-to-many assignment between two given graphs seen as two many-to-one matchings from either graphs towards an imaginary intermediate graph

ces between G , H , and a virtual intermediate graph for each matching (Figure 7.7), of size N_K , where N_K is $\min\{N_G, N_H\}$):

$$\begin{aligned} \min_{P_1, P_2} \|P_1 G^T P_1^T - P_2 H P_2^T\|_F^2 \quad \text{s.t.} \\ P_1 \mathbb{1}_{N_G} \leq k_{\max} \mathbb{1}_{N_K}, \quad P_1^T \mathbb{1}_{N_K} = \mathbb{1}_{N_G} \\ P_2 \mathbb{1}_{N_H} \leq k_{\max} \mathbb{1}_{N_K}, \quad P_2^T \mathbb{1}_{N_K} = \mathbb{1}_{N_H} \end{aligned} \quad (7.5)$$

where $\mathbb{1}_N$ represents the constant N -dimensional vector of all-ones. In our experiments, we consider $N_G \geq N_H$. The maximal number of vertices merged together is represented by k_{\max} and the many-to-many matching matrix is given by $P = P_1^T P_2$, where $P \in \{0, 1\}^{N_G \times N_H}$ is the matching matrix between G and H .

The difference between the graphs size (Figure 7.8) is handled either by setting $k_{\max} \geq 2$ (hence allowing at most k_{\max} nodes to be merged), or by setting $k_{\max} = 1$ (hence allowing the implicit choice of nodes that will be assigned as dummy, within the graph having a larger dimension).

The authors in [ZBV10] propose an approximation of the final solution, using a version of the conditional gradient algorithm, based on the continuous relaxation of (7.5). They also reformulate the gradient minimization as a linear assignment problem with a cubic complexity, hence making it feasible for high-dimension graph matching.

7.2.4 Matching of weighted labeled graphs

There are numerous applications where nodes in a graph represent different features (e.g. location) while the edges designate the relationships between those. In this context, certain labels can be assigned to nodes, indicating an additional degree of similarity. Let us consider that the matrix $C \in \mathbb{R}^{N_G \times N_H}$ incorporates the node-to-node dissimilarities from G to H (i.e. the difference between node labels of G and node labels of H). Then, the formulation that takes into account at the same time

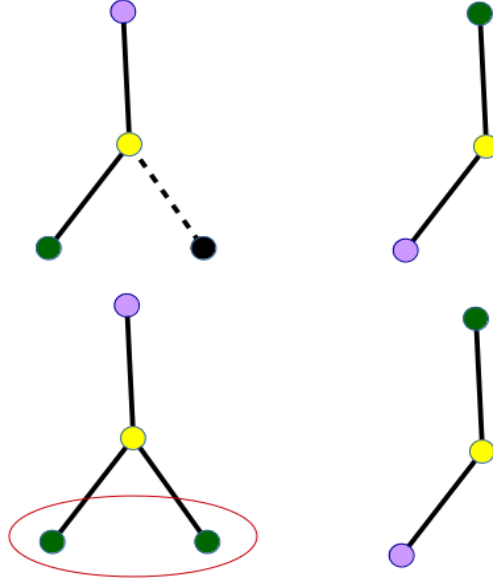


Figure 7.8: The effect of modifying the maximum number of nodes matched together. Top row corresponds to the case where k_{\max} is set to 1, so that one of the nodes in the larger graph is assigned as a dummy node. Bottom row illustrates the case where the value of $k_{\max} = 2$ corresponds to the maximum number of nodes merged together.

the label differences, and the graph structure (the influence of which is controlled by $\lambda \in [0, 1]$), is the following:

$$\begin{aligned} \min_{P_1, P_2} (1 - \lambda) \|P_1 G^T P_1^T - P_2 H P_2^T\|_F^2 + \lambda \text{tr} C^T P_1^T P_2 \quad \text{s.t.} \\ P_1 \mathbb{1}_{N_G} \leq k_{\max} \mathbb{1}_{N_K}, \quad P_1^T \mathbb{1}_{N_K} = \mathbb{1}_{N_G} \\ P_2 \mathbb{1}_{N_H} \leq k_{\max} \mathbb{1}_{N_K}, \quad P_2^T \mathbb{1}_{N_K} = \mathbb{1}_{N_G} \end{aligned} \quad (7.6)$$

While incorporating label (dis)similarities among the nodes is a common practice for graph-matching techniques, our practical setting given by the matching of FN variant graphs, does not include any label information at the nodes. Therefore, for the rest of the applications presented in this manuscript, where the many-to-many assignment framework is used, we consider that λ is 0, and thus rely on [Equation 7.5](#).

7.3 SUMMARY

Since we are interested in computing the global matching between the graph-based representation of FN variants (for subsequent use in different applications), we were motivated to explore the graph-matching techniques to identify an appropriate methodology to compare graphs. Many-to-many assignment framework is capable to evaluate the similarity with respect to the number of mismatched edges between any two graphs and can be applied to reasonably-sized graphs (e.g. several hundred nodes).

In the next chapters ([Chapter 9](#)), we compare the performances of the many-to-many assignment framework to a different methodology that evaluates the

local structural similarity between graphs, through optimal transport. Furthermore, we rely on this formulation to propose various ways to define the representative individual ([Chapter 10](#)) of a set of given graphs, to compute the "average" individual of FN-specific variants in normal and tumoral cases. Finally, in [Chapter 11](#), we show an application of the graph-matching problem for statistical analysis of variation of parameters through FN deformation maps.

OPTIMAL TRANSPORT THEORY - APPLICATION TO HISTOGRAM AND GRAPH MATCHING

In this chapter, we provide general notions regarding the discrete optimal transport theory. We focus on some algorithmic aspects, and subsequently, we describe a methodology based on Gromov-Wasserstein discrepancy between similarity matrices, introduced in [CSP16], that we have selected for the purpose of comparing graphs. Not being explicitly constructed to match graphs, we illustrate the modifications that we proposed here in order to employ this approach in a graph-matching context.

8.1 OPTIMAL TRANSPORT GENERAL BACKGROUND

Optimal transport theory has been established as being a powerful tool in data sciences, capable to compare distributions (expressed as features, descriptors, weights, etc.) and to provide meaningful distances between them while reflecting their underlying geometrical properties.

Dating back to the 18-th century with the Monge's *Mémoire sur la théorie des déblais et des remblais* [Mon81], the optimal transport (OT) has known significant advancement with the contributions of Leonard Kantorovich [Kan06] in 1942 in economics, on the resource allocation problem involving OT, and later, with the works of Brenier [Bre91] showing there is a link to mathematical domains, such as convexity, partial differential equations and statistics. A recent collection of theoretical insights into the optimal transport framework [Vil03; Vil08; San15] is considered a reference point for an expanding community working with optimal transport.

In computer vision, the landmark paper [RTG00] on content-base image retrieval, has introduced the notion of Earth Mover Distance, as a means to compare colour histograms in a space endowed with a ground distance between the bins (i.e. individual distance between the colours). The OT framework has subsequently been used for solving a plethora of applications related to signal and image processing (e.g. texture and colour modelling, shape and image registration), of which we name a few that are relevant to the use of OT throughout this manuscript.

One of these applications concerns the color transfer between images, i.e. modifying an image to match the color distribution of another one, while preserving its geometry. The analogue problem for gray-scale images is the histogram matching problem that corresponds to the application of the 1D optimal transport between the gray levels of the pixels of the images [Del04]. The difficulties of color transfer problem arise when the color distributions have different shapes and the transfer of

colours, being sensitive to outliers, is incoherent for neighboring pixels [DRG09]. Therefore, efforts have been made in order to consider the spatial nature of images in a regularized formulation of the optimal transport [RDG11; Rab+14]. In [Fer+13], the relaxed and regularized transport (Figure 8.1) shares similarities with the *graph matching framework* as the regularization takes into account the spatial information provided by the cloud points represented using graphs (graph-based regularization encodes neighborhood similarity).

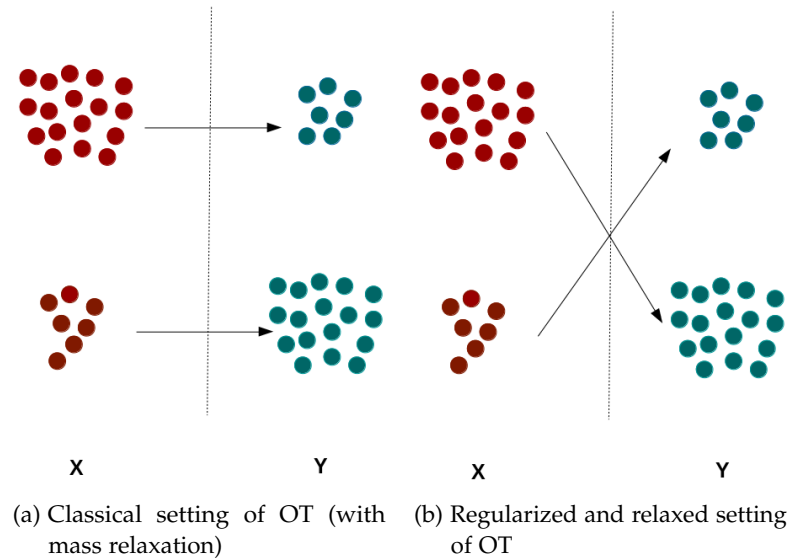


Figure 8.1: Relaxed and regularized OT, figure inspired from [Fer+13]: Given X and Y as 2 point clouds to be matched under the OT framework (with a mass relaxation constraint, e.g. one can allow the mass from X to not entirely be transported to Y). If there is no regularization on the classical formulation then the matching occurs as in a), while introducing regularization in the transport formulation (that takes into account the proximity of the points to one another), allows clusters with similar shapes to be matched together.

Another line of applications concerns the point/cloud/mesh registration for shape analysis and graphics. Shapes or objects (2D or 3D) are represented through sets of cloud points endowed with metric properties. Within the optimal transport methodology applied to compare or register images, it is common to enforce the neighboring points to remain near to each other after the transformation. Often, these methods rely on graph-based representations of the shapes, to embody the structural relationships between the points [Mém11; Mém07; Sol+; CSP16].

Optimal transport is additionally employed to perform texture synthesis and mixing, image denoising and restoration, machine learning and statistics, etc. The range of applications based on the optimal transport framework grows larger every year considering its attractive properties, however, it is not the intention of this work to cover them in detail. For a recent review of the computational properties and applications of optimal transport theory, the reader is referred to [PC19] and [Kol+17b].

One last relevant application of the OT framework concerns the computation of the barycenter (weighted mean) of data points, that can be formulated once a

distance (e.g Wasserstein distance) is obtained. The barycenter provides a representation of the average individual of different sets of objects, shapes, etc, embedded in a metric space. Aside for the shape analysis [CSP16], barycenters can also be found in image processing for texture synthesis and mixing [Jul+11].

In this chapter, we introduce the general optimal transport framework for discrete distributions, such as it is formulated by Monge and reformulated by Kantorovich through linear programming. We then focus on some particular applications of the discrete optimal transport, namely 1D OT for histogram equalization and an extension to the OT theory, Gromov-Wasserstein distance for comparing topological metric spaces.

8.2 OPTIMAL TRANSPORT FOR DISCRETE DOMAINS

NOTATIONS:

Let $S_N = \{s \in [0, 1]^N, \sum_i s_i = 1\}$ be a set of N -dimensional histograms (probability vectors).

$\mu_0 = \sum_{i=1}^N a_i \delta_{x_i}$ and $\mu_1 = \sum_{j=1}^M b_j \delta_{y_j}$, where δ_x is the Dirac distribution at the location x , be two discrete probability measures with (positive) weights a_i, b_j , where $a, b \in S_N$, on the support defined by $\{x_1, \dots, x_N\} \in \mathbb{X}$ and $\{y_1, \dots, y_M\} \in \mathbb{Y}$.

Monge's optimal transport problem between discrete measures [Mon81] searches for the map $T : \{x_1, \dots, x_N\} \mapsto \{y_1, \dots, y_M\}$ that will associate to each x_i , a single y_j , by pushing mass from μ_0 to μ_1 in such a way so that the mass transportation cost is minimized. If we consider the cost function $c(x, y)$ defined for $(x, y) \in \mathbb{X} \times \mathbb{Y}$, then the optimal transport problem between discrete measures, under the mass conservation constraint, is the following [PC19]:

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) : \forall j, b_j = \sum_{i: T(x_i)=y_j} a_i \right\} \quad (8.1)$$

An intuitive way to interpret the mass conservation constraint is the following: Firstly, the mass cannot be *split*, i.e. one x_i cannot be matched to multiple y_j , and secondly, the surjectivity of the mapping implies that every y_j needs to be associated to another x_i . Figure 8.2(a) illustrates an example of a mapping determined by Monge's optimal transport theory in a discrete case.

Defining σ as a permutation function, $\sigma : \{1, \dots, N\} \mapsto \{1, \dots, M\}$, so that $j = \sigma(i)$, the mass conservation constraint can be re-written as:

$$b_j = \sum_{i \in \sigma^{-1}(j)} a_i \quad (8.2)$$

If we consider the particular case where the discrete measures own a support with equal dimension, $N = M$ and a_i, b_j are uniform weights, $a_i = b_j = 1/N$, then the transport map is bijective $T(x_i) = y_{\sigma(i)}$. Additionally, we express the result of a cost function $c(x, y)$ as the element C_{ij} of a matrix $C \in (\mathbb{R}^+)^{N \times N}$.

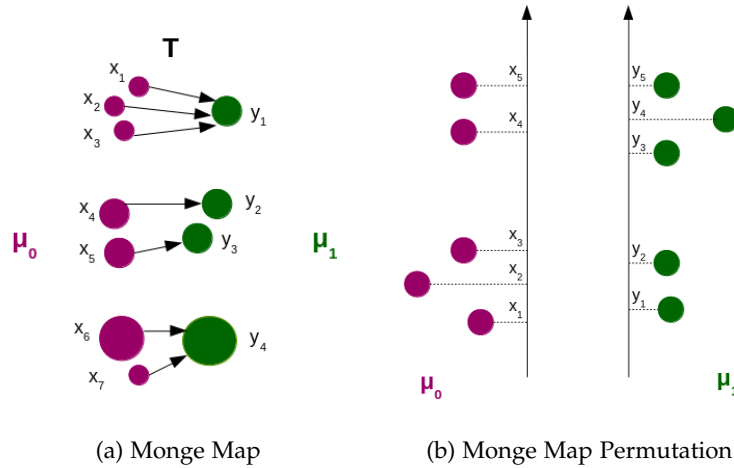


Figure 8.2: (a) Monge map T that associates μ_0 to μ_1 through the map $T: T(x_1) = T(x_2) = T(x_3) = y_1; T(x_4) = y_2; T(x_5) = y_3; T(x_6) = T(x_7) = y_4$. The size of the disks is proportional to the corresponding weights a_i, b_i . (b) The weights are uniform so the map is bijective. One possible map would be: $T(x_i) = y_{\sigma(i)}$, where $\sigma_{\{1, \dots, 5\}} = \{5, 4, 3, 2, 1\}$.

Thus the Monge formulation becomes the following optimal matching problem:

$$\min_{\sigma \in \text{Perm}(N)} \frac{1}{N} \sum_i^N C_{i, \sigma_i} \tag{8.3}$$

This particular case is displayed in Figure 8.2(b), where for two given distributions μ_0, μ_1 , the weights a_i, b_j , associated to discrete measures with size N , are uniform so the optimal transport problem can be formulated as a permutation problem. Being formulated as a permutation problem, (a combinatorial problem), there can be more feasible solutions; one direct application concerns the comparison of equal-sized histograms, which is the subject of Section 8.2.3.

Kantorovich's ideas have an economic interpretation, that, for the sake of analogy, we shortly illustrate here, as a resource allocation problem: Resources found at x_i in the quantity defined by a_i , need to be transported at the warehouses at y_i that demand b_j resources. Knowing the cost to move a unit of resource from a_i to b_j , as C_{ij} , the main idea is to obtain a transport plan (P), that optimizes the total cost of transport.

8.2.1 Optimal Transport with Linear Programming

Regarding the mass constraint, there are certain limitations when comparing discrete probability measures with a different size of the support, i.e. $M \neq N$. More specifically, if $M < N$, the fact that the mass cannot split is too restrictive, as in this case, assignment maps between μ_0 and μ_1 cannot be derived. One illustrating example is given in Figure 8.3, where $\mu_0 = \delta_{x_1}$ and $\mu_1 = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$. Under the assumptions of the formulation in Equation 8.2, one can not find any feasible map between μ_0 and μ_1 .

In the next subsection, we present a relaxation of the mass constraint, an important contribution of Kantorovich [Kan06] to the optimal transport theory, with a strong economical flavour. In the above example, this relaxation allows for the mass at location x_1 to be split towards y_1, y_2 . As already mentioned, Kantorovich's contribution to optimal transport theory concerns the relaxation of the mass conservation constraint. Under Monge's formulation, this constraint states that a point at

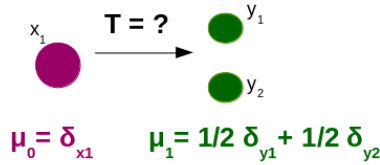


Figure 8.3: Monge mass splitting: no feasible transport map between μ_0 and μ_1 .

location x_i can only be assigned to another y_i . Specifically, in his groundbreaking work, Kantorovich proposes to have the mass at any location x_i , potentially *split* across various locations y_j .

To formalize this aspect, in the framework of optimal transport between two discrete probability measures μ_0 and μ_1 , we define the transport matrix $P \in \mathbb{R}^{N \times M}$, whose element P_{ij} indicates the amount of mass flowing from a_i to b_j .

Then, the Kantorovich optimal transport now reads:

$$F_C = \min_P \sum_{ij} C_{ij} P_{ij} \tag{8.4}$$

where $P \mathbb{1}_N = \mu_0$, $P^T \mathbb{1}_M = \mu_1$ and $P \geq 0$. Finding the optimal transport plan denoted by P of the convex optimization problem in Equation 8.4, can be achieved by solving a linear program, which is a method to obtain the best outcome given constraints that are linear equations of the form: $LQ_{\min} = \{ \sum_{ij} C_{ij} P_{ij} : P_{ij} \geq 0, \sum_j P_{ij} = \mu_0, \sum_i P_{ij} = \mu_1 \}$ [San18].

This formulation is analogue to the Earth Mover Distance, defined in [RTG00], where it has been used in a content-based image retrieval task, leveraging the fact that metric can provide a meaningful comparison between histograms, based on the known distance between the individual bins (encoded in the cost matrix).

For the example above, where $\mu_0 = \delta_{x_1}$ and $\mu_1 = \frac{1}{2} \delta_{y_1} + \frac{1}{2} \delta_{y_2}$, the feasible coupling (transport) matrix that respects the mass constraints of Equation 8.4 is shown in Figure 8.4.

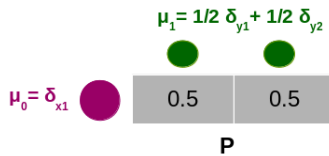


Figure 8.4: Kantorovich mass splitting

As a matter of fact, transportation problems were among the first problems to be approached with linear programs; optimal transport are in fact equivalent to an important class of linear programs, known as minimum cost network flows [KV00].

8.2.2 Metric properties of optimal transport

The question whether OT can provide a distance between histograms (probability measures) has been studied and proven true [Vil03] as long as the cost matrix satisfies certain properties. In [RTG00], the cost matrix C is associated to the *ground distance* between the bins of the histograms μ_0 and μ_1 . We can define W_p , the p -Wasserstein distance on S_N , as:

$$W_p(\mu_0, \mu_1) \stackrel{\text{def.}}{=} F_{C^p}(\mu_0, \mu_1)^{1/p} \tag{8.5}$$

8.2.3 Optimal transport 1D

1D optimal transport can be used to perform histogram equalization, with applications the histogram of a gray-scale images.

PROBLEM FORMULATION: Let us consider $f, g \in \mathbb{R}^N$, where $f = (f_i)_{i=1}^N, g = (g_i)_{i=1}^N$ correspond to pixel intensities of gray-scale images that are characterized by the empirical intensity distributions: $\mu_0 = \frac{1}{N} \sum_{i=1}^N \delta_{f_i}$ and $\mu_1 = \frac{1}{N} \sum_{i=1}^N \delta_{g_i}$, respectively.

Next, we define the "ground cost" between the sets of 2D pixel coordinates: $\forall (x, y) \in \mathbb{R}^2 \times \mathbb{R}^2$, as the L^p Euclidean-norm cost, $C(x, y) = \|x - y\|^p$, where $p > 1$. An optimal matching (assignment) between f and g corresponds to the optimal permutation σ^* that minimizes the difference between f and the permuted (re-ordered) version of g , as follows:

$$\sigma^* \in \arg \min_{\sigma \in S_N} \sum_{i=1}^N C(f_i, g_{\sigma(i)}) \quad (8.6)$$

In order to compute σ^* , we follow the subsequent steps that are illustrated in [Figure 8.5](#):

1. Consider σ_f , the permutation that will sort the pixel indices (enumerated in a predefined order, e.g. column-wise from 1 to N) of f , such that $f_{\sigma_f(k-1)} \leq f_{\sigma_f(k)} \leq f_{\sigma_f(k+1)}, \forall k \in \{2, \dots, N-1\}$. Analogously, we consider σ_g , such that $g_{\sigma_g(k-1)} \leq g_{\sigma_g(k)} \leq g_{\sigma_g(k+1)}, \forall k \in \{2, \dots, N-1\}$.
2. Denote σ_f^{-1} as the inverse permutation, such that $\sigma_f^{-1}(\sigma_f) = \text{Id}$.
3. Consider the mapping (assignment) $f_{\sigma_f(k)} \leftrightarrow g_{\sigma_g(k)}, \forall k \in \{1, \dots, N\}$, that corresponds to an optimal assignment in the sense of the cost function C . It follows that the optimal permutation that is applied to g to "reproduce the closest as possible" the ordering of f is of the form: $\sigma^* = \sigma_g(\sigma_f^{-1})$.

By applying the optimal permutation σ^* to g , we obtain $f_{\text{eq}} = g(\sigma^*)$, that represents the reordering of the pixel indices of g to match as closely as possible f , or, alternatively, the histogram equalization of f , using the histogram of g . The result for the histogram equalization of Image a in [Figure 8.6](#) using the histogram of Image b, is shown in [Figure 8.7](#).

8.3 ENTROPIC REGULARIZATION OF THE OT

The classical formulation of the optimal transport theory, such as it is formulated in [Section 8.2.1](#), is a combinatorial problem, and the algorithms that provide the solution (e.g. network simplex or interior point methods) have a complexity of at least $\mathcal{O}(N^3 \log N)^1$ for histograms of dimension N . The cost can be prohibitive for histograms whose dimension exceeds a few hundreds. A regularization of the

¹ Notation for describing the time complexity (in number of operations) of an algorithm being developed to operate on a set of N elements.

The entropic regularization of the optimal transport is inspired from transportation theory [Erlander and Stewart][ES90], where the properties of entropy as a dispersion measure were considered for traffic patterns modelling. Using an entropic term in the optimal transport problem (thus mitigating sparsity of the couplings) is shown to describe more realistically the traffic patterns.

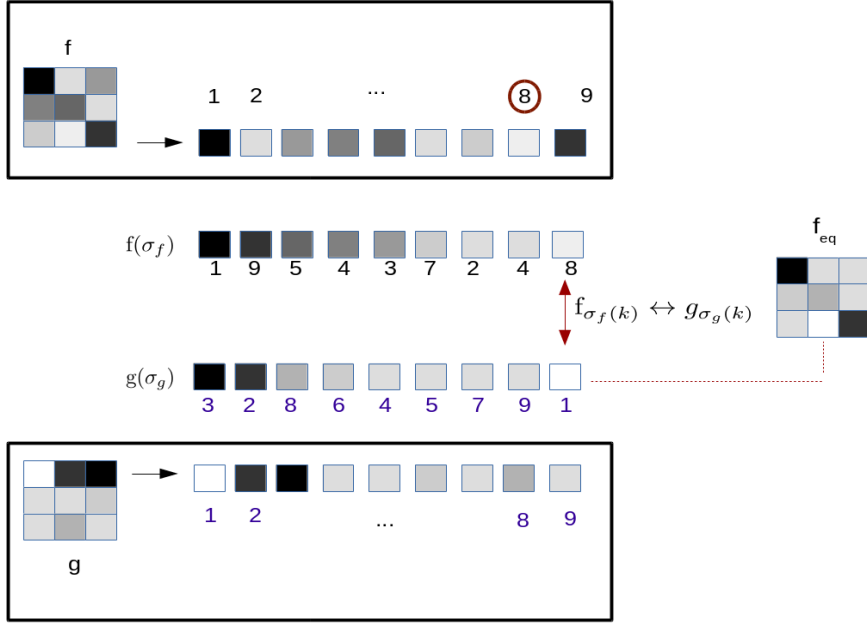


Figure 8.5: OT-1D: Permutation scheme and optimal assignment between f, g (top and fourth rows correspond to pixel intensities of f and g , respectively, displayed in a row enumerated from 1 to 9, starting from the upper-left corner), the middle plots (second and third) correspond to pixel intensities sorted in ascending order with σ_f, σ_g , respectively. When multiple pixels have the same intensity, the permutation that orders the values is not unique. The assignment (that will instill the optimal permutation upon g) is given by $f_{\sigma_f(k)} \leftrightarrow g_{\sigma_g(k)}, \forall k \in \{1, \dots, 9\}$. The scheme explicitly shows that the f_{eq} is formed by the g pixels intensities, reordered (permuted) according to the optimal transport problem, to account for "location" of the corresponding intensity value in f . In other words, f_{eq} "reproduces" the ordering of the pixels in f as closely as possible to resemble f , but with the intensities (hence histogram) belonging to g . f has inherited the histogram of g .

formulation of Equation 8.4 can be proven useful for accelerating the computational time.

The work of [Cut13] showed that by regularizing the classical optimal transportation problem with an entropic term, one can obtain a distance, which can be computed through algorithms that converge at several orders of magnitude faster than the transportation solvers.

The entropy of the coupling matrix (a 1-strongly concave function [PC19]), is defined as follows:

$$H(P) = - \sum_{i,j} P_{i,j} (\log(P_{i,j}) - 1) \quad (8.7)$$

Regularizing the original formulation in Equation 8.4, with an entropic term, leads to the following formulation of the *regularized discrete optimal transport*:

$$F_C^\epsilon = \min_P \sum_{ij} C_{ij} P_{ij} - \epsilon H(P) \quad (8.8)$$



Figure 8.6: OT-1D: (a),(b) Gray-scale images f, g with the corresponding intensity histograms μ_0, μ_1 (10 equal-spaced bins within the range of the f, g) displayed below in (c),(d).

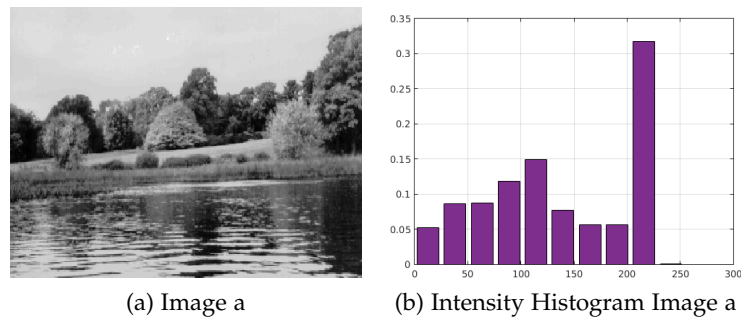


Figure 8.7: OT-1D: (a) Image a is enforced with the histogram corresponding to Image b (b), as a result of mass transport OT -1D between their histograms. In other words, Image a corresponds to the re-ordered (with the optimal permutation) version of Image b.

The objective function in Equation 8.8 is a strongly convex function, that will thus lead to a unique optimal solution. Increasing the value of ϵ will "diffuse" the couplings, i.e. P becomes less sparse. The advantages of this behaviour relate to the speed of convergence of the minimization algorithms, i.e. increasing ϵ and leads to a faster convergence, as shown in [PC19]. Additionally, for a small regularization, the solution converges to the maximum entropy optimal transport coupling, while for a large regularization, the solution converges to the coupling with maximal entropy between two prescribed marginals μ_0, μ_1 , namely the joint probability between two independent random variables distributed following μ_0, μ_1 .

The entropic regularization formulation can be approximated with Sinkhorn's algorithm [SK67]. In [Cut13], it was showed that Sinkhorn-Knopp's matrix scaling

algorithm can perform significantly better on high-dimensional datasets (several orders of magnitude faster) than the algorithms solving the classic formulation of the optimal transport, making this framework feasible for the machine learning community.

8.4 GROMOV-WASSERSTEIN DISCREPANCY FOR OPTIMAL TRANSPORT BETWEEN STRUCTURED OBJECTS

There are applications where it might be interesting to consider the optimal transport framework for shape comparison (represented, for instance, by sets of points embedded in a metric space). Discrete distributions would then be associated to the sets of points. A weakness of the classic discrete formulation in Equation 8.4, is that it doesn't take into account the inner structural dependency of the objects (e.g. neighboring points should stay together when transported/assigned to the other shape).

Within the framework of structured optimal transport, several works [CSP16; AMJJ18; Vay+18; Cou+17] have shown the advantage of incorporating additional geometrical properties into the cost function, for tasks such as domain adaptation, natural language processing, computing graph barycenters or graph clustering. If the approach in [CSP16] includes the intrinsic structure of the objects in the cost formulation, the authors in [Vay+18] present a new class of distances, that incorporates both structural and feature information into its transport cost. They focus on previously labeled structured objects, where for instance, graph edges represent relationships between features (nodes).

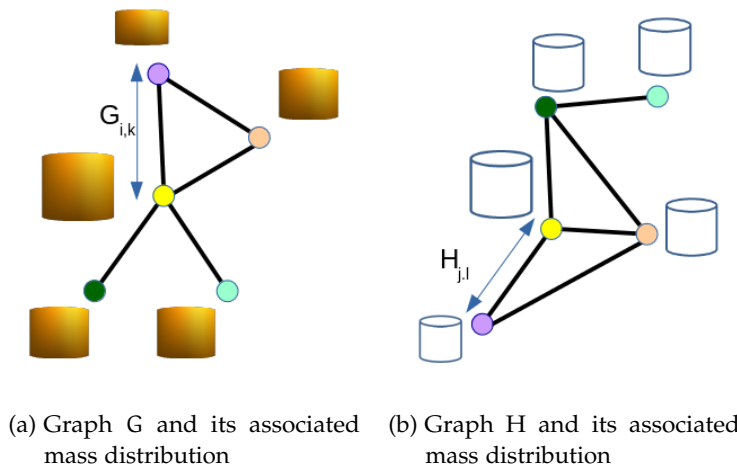


Figure 8.8: OT-Gromov-Wasserstein: (a) Mass is associated to nodes in a graph G that are transported to nodes of graph H (b), after finding the optimal assignment.

8.4.1 Gromov-Wasserstein discrepancy - adaptation to a graph-matching context

Throughout this manuscript, we are mainly interested in comparing unlabeled graphs, where no previous pairwise correspondences between the nodes of two

graphs are known beforehand. Therefore, we selected the work of [CSP16], where the authors have considered a metric called Gromov-Wasserstein, capable of comparing objects (with a defined structure, given for instance by distance matrices) that can lie in spaces with different dimensions. This discrepancy, formulated as a minimization problem of a non-convex objective function, is computed with a fast iterative algorithm based on an entropic regularization of the transportation matrix.

It is worth mentioning that the formulation [CSP16] for computing the Gromov-Wasserstein discrepancy is conceived for *similarity matrices*, i.e. any matrices containing pairwise relationships (not necessarily distance matrices). The change of semantics to suit a graph comparison context is clear: similarity matrices become graph adjacency matrices.

In the following, we consider $(G, \mu_0) \in \mathbb{R}^{N_G \times N_G} \times S_{N_G}$ and $(H, \mu_1) \in \mathbb{R}^{N_H \times N_H} \times S_{N_H}$, where G and H (Figure 8.8) encode the graphs' structure given by either the adjacency matrices, or the shortest path between the graph nodes, and μ_0 and μ_1 are the mass distributions associated to the graph nodes (e.g. uniform distributions, $\mu_0 = \frac{1}{N_G} \mathbb{1}_{N_G}$ and $\mu_1 = \frac{1}{N_H} \mathbb{1}_{N_H}$). The entropic Gromov-Wasserstein discrepancy between (G, μ_0) and (H, μ_1) is defined as follows:

$$\begin{aligned} \min_{P'} \sum_{i,j,k,l} L(G_{i,k}, H_{j,l}) P'_{i,j} P'_{k,l} - \epsilon H(P') \quad \text{s.t.} \\ P' \mathbb{1}_{N_H} = \mu_0, \quad P'^T \mathbb{1}_{N_G} = \mu_1 \end{aligned} \quad (8.9)$$

The transport matrix indicates the matching between the two graphs such as if its term $P'(i, j) > 0$, the node j of graph H is assigned to the node i of graph G . The loss function $L(u, v)$ can be taken as the quadratic loss (e.g. of the form $L(a, b) = \frac{1}{2}|a - b|^2$) or Kullback-Leibler divergence.

Considering a quadratic loss for the Gromov-Wasserstein discrepancy, the formulation can be traced back to the softassign quadratic assignment problem from the inexact graph-matching domain [GR96].

PRACTICAL CONSIDERATIONS: If the size of the graphs to compare is different, i.e. $N_G \neq N_H$, the constraints of Equation 8.9 (that require all the mass from G to be transported to H) will determine a "dispersion" of mass to compensate the difference of mass between individual nodes of both graphs. Let us consider the following example (Figure 8.9), where we are given two graphs with size $N_G > N_H$, and uniform distributions $\mu_0 = \frac{1}{N_G} \mathbb{1}_{N_G}$ and $\mu_1 = \frac{1}{N_H} \mathbb{1}_{N_H}$. It follows that $\mu_0(i) < \mu_1(j)$, implying that under the mass constraints of the problem, at least 2 nodes from H have to "receive" the mass from G , or, in an assignment framework, that a behaviour of many-to-one matching is being triggered. This behaviour is problematic as one cannot control the dispersion of mass, or, analogously, the nodes assignment.

A second practical aspect results from the choice of regularizing the transport of mass with an entropic term. As already pointed out in the previous section, increasing the value of ϵ diffuses the couplings, which is equivalent to saying that the each of the nodes from one graph, will potentially be assigned to all the nodes from the second graph. Despite the algorithmical advantage of regularization, we decide to set $\epsilon = 0$ to favour one-to-one assignments, in which case the non-convex optimization problem can be approximated by a classical solver with supercubical complexity.

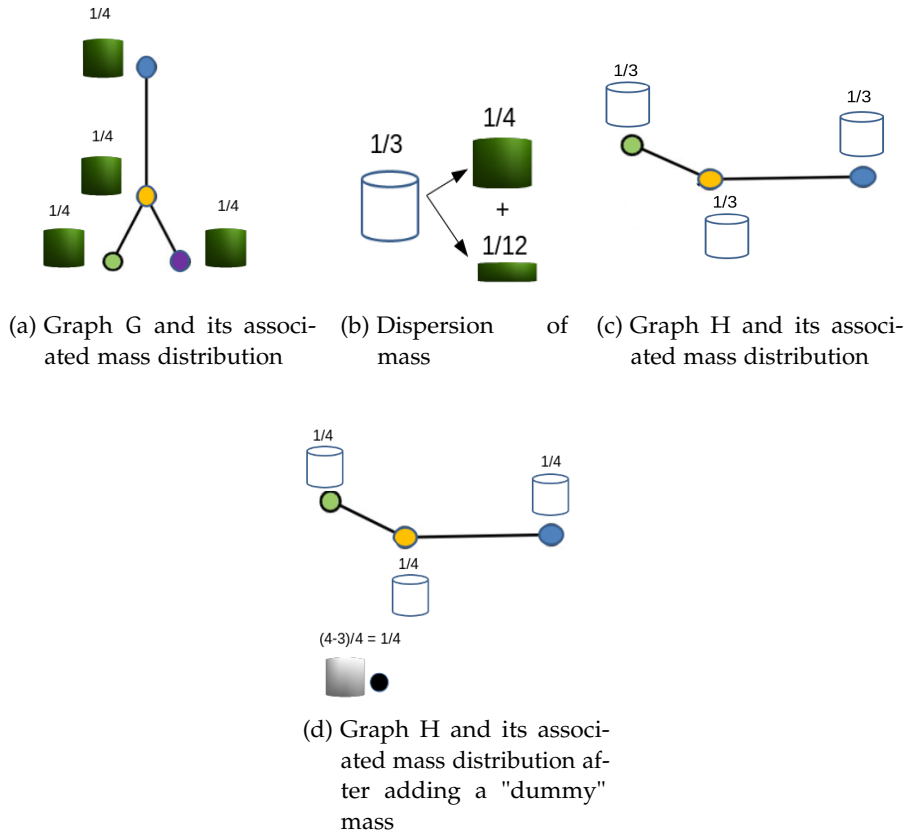


Figure 8.9: Mass dispersion-matching graphs using optimal transport framework with Gromov-Wasserstein discrepancy. The difference in size of two graphs G and H triggers a mass dispersion, i.e. every node of H has to receive mass from 2 nodes of G, to satisfy the mass conservation constraint. The workaround consists in adding a dummy node to compensate for the difference of mass.

8.4.2 Gromov-Wasserstein barycenter

The work of [CSP16] defines Gromov-Wasserstein barycenters for a set of (similarity matrices) that we consider here as graphs represented by the adjacency matrices $G_s \in \mathbb{R}^{N_s \times N_s}$. If we denote by $\mathcal{E}_{G,H}(P')$ the objective function in Equation 8.9, then the barycenter graph of adjacency matrix $G \in \mathbb{R}^{N \times N}$, where N is user-determined and μ is the associated mass distribution, is determined by the following formulation:

$$\min_{G, P'_s} \sum_s \lambda_s (\mathcal{E}_{G, G_s}(P'_s) - \epsilon_H(P'_s)) \quad (8.10)$$

where μ_s are the mass distributions associated to G_s .

The authors in [CSP16] have proposed an iterative minimization scheme, with respect to G and alternatively to P'_s , and have subsequently shown that when using a quadratic loss, the solution of Equation 8.10, with respect to G , converges to $\frac{1}{\mu \mu^T} \sum_s \lambda_s P'_s{}^T G_s P'_s$. $P'_s \in \mathbb{R}^{N \times N_s}$ are the transport matrices between the barycenter and the similarity matrices set. λ_s are considered weights so that $\sum_s \lambda_s = 1$. The

term $\lambda_s P_s'^T G_s P_s'$ recovers a similarity to the formula of the permuted graph from the graph-matching section [Equation 7.3](#), indicating that the solution to the barycenter based on Gromov-Wasserstein discrepancy converges to a sort of the average of realigned matrices G_s .

8.5 SUMMARY

We have introduced the discrete optimal transport framework, and more precisely the methodology based on the Gromov-Wasserstein discrepancy, that we have adapted with the purpose of comparing graphs. An application of this can be found in the next chapter ([Chapter 9](#)), where we perform an analysis of the performance of the optimal transport using Gromov-Wasserstein and many-to-many assignment framework, in a graph comparison context. We also employ an OT-based method, 1D OT transport for "Gaussianization" of an image (i.e. conversion of its native intensity histogram into a normal distribution) ([Chapter 6](#), [Chapter 11](#)). Lastly, the formulation of the Gromov-Wasserstein barycenter serves as inspiration for one of the methodologies proposed in ([Chapter 10](#)) for defining the representative graph of a given set.

COMPARISON OF PERFORMANCES OF OPTIMAL TRANSPORT AND MANY-TO-MANY ASSIGNMENT

In previous chapters, two approaches for graph comparison were described, namely many-to-many assignment and optimal transport-based framework. Before deciding upon the most appropriate method to use in a real setting, i.e. comparison of graph-based FN networks or computation of prototype graphs (average individual of a specific class), it was essential to perform a study of the algorithms' behaviour in a tractable and simpler setting, using a database of generated toy-graphs. In this way, we are provided with a clear expectation of the desired result in terms of node assignment, and subsequently, of the expected cost after graph matching.

In this chapter, we describe our contribution in terms of the analytic framework, that is able to provide a comparison of the performances of the many-to-many assignment and optimal transport-based approach, upon matching randomly generated graphs.

9.1 TOY GRAPHS GENERATION

In order to analyse the the behaviour of the many-to-many assignment and optimal transport frameworks, we have generated a database of graphs, following the steps that we describe here (an illustration of the steps is also shown in [Figure 9.1](#)):

1. Generation of a uniform Poisson point process ¹ with average intensity λ on a bounded rectangle region of size $N \times M$ pixels: the number of points is a Poisson random variable with mean λMN distributed uniformly in the 2D cartesian coordinate system.
2. The previously generated points constitute the seeds of Voronoi diagrams (the partitioning of a plane with N points into convex polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other).
3. Derivation of the skeleton-graph-based of the Voronoi images and modify the intensity λ to change the graph dimension as desired.

¹ Poisson point process [Poi] is characterized by the Poisson distribution: consider $N(B)$, a random variable, as the number of points of a point process N in a region $B \subset \mathbb{R}^2$. The probability that n points belonging to a homogeneous Poisson point process with intensity λ , exist in B , is given by: $P\{N(B) = n\} = \frac{(\lambda|B|)^n}{n!} e^{-\lambda|B|}$, where $|B|$ is the area of B . A uniform Poisson point process is of the form $\Lambda = \nu\lambda$, where ν is a Lebesgue measure and λ is related to the expected number of Poisson points existing in some bounded region.

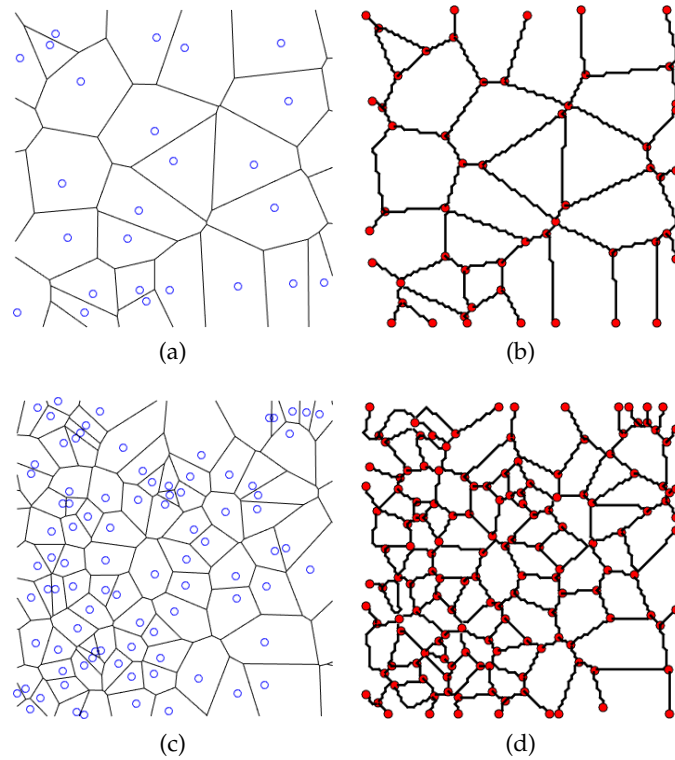


Figure 9.1: Methodology for generating graphs with Voronoi cells and Poisson point process: Uniformly distributed points (\circ), displayed as a Poisson point process on a rectangle of size 100×100 pixels and mean intensity $\lambda = 0.009$ (a,b) and $\lambda = 0.03$ (c,d). Figures (c,d) illustrate the graphs associated to the Voronoi diagram structure built from the seeds-Poisson points.

Following the above graph-generation methodology, we have first generated random graphs of different size (i.e. 16 vertices and 181 vertices), as illustrated in Figure 9.2. These graphs describe thus the structure of Voronoi diagrams, generated

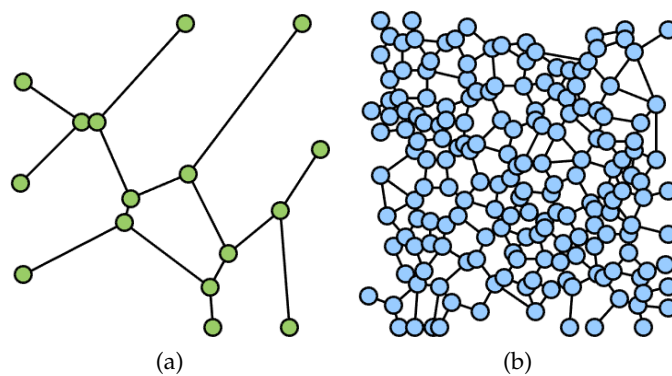


Figure 9.2: Generated toy-graphs: (a) 16 nodes and (b) 181 nodes

from seeds uniformly distributed on a bounded region.

For the purpose of graph matching between pairs of test graphs, several spatial transformations were applied to the previously generated samples:

- *Rotation*: In order to test the ability of the methods to detect the isomorphic graphs, the first modification is related to the rotation (permutation) on the set of nodes. Therefore, the size remains unchanged when applying the modification and the challenge of the matching technique is to detect the isomorphism case (thus, the matching cost should be 0).
- *Dimension change*: Subsequently, we chose to modify the size of the graphs by removing nodes of various degrees (the node degree is given by the number of incident edges). In this case, the expected cost depends on the number of removed edges, as a consequence of the node degree of the extracted nodes.

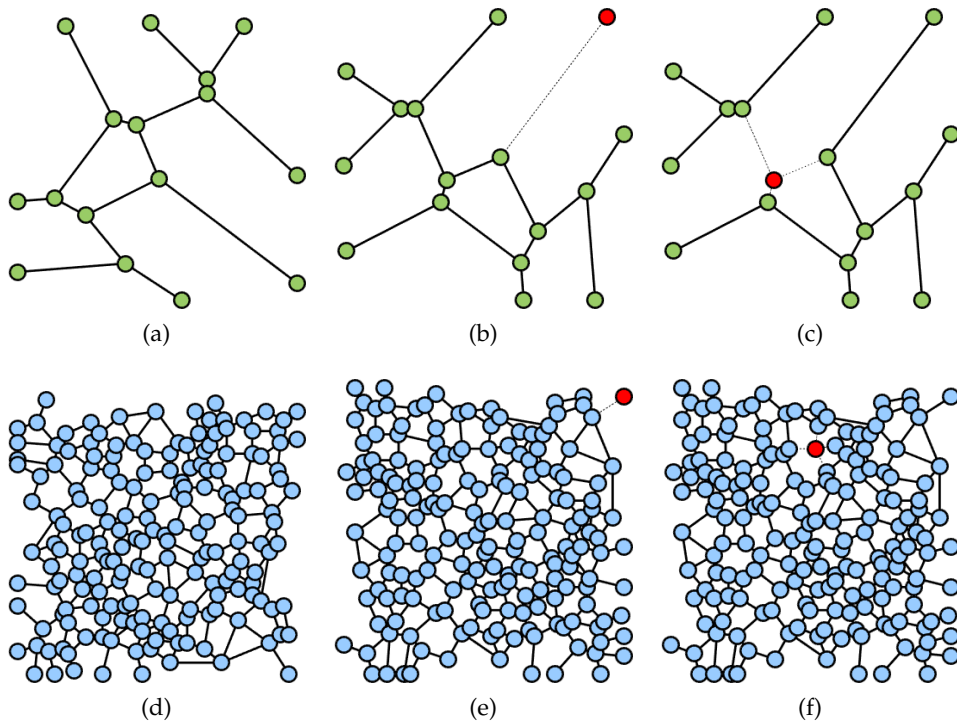


Figure 9.3: Modified toy-graphs: (a,d) Rotation of the node indexes (b,e) Removal of one-degree node, (c,f) Removal of higher-order degree

9.2 METHODOLOGY FOR ESTABLISHING A COMMON FRAMEWORK FOR COMPARISON

For each case determined by the aforementioned spatial modifications, we have, in turn, represented the two graphs to compare, G and H , (see [Figure 9.2](#), [Figure 9.3](#)) by the binary adjacency matrix, or by the length of shortest path between the nodes up to different depths: 2-nd degree, 3-rd degree, and total. Technically, for a k -th ($k \in \mathbb{Z}, k \geq 2$) degree, an element of the adjacency matrix is either 0, if the length of the shortest path between node i and node j is greater than k , or a value from the set $\{1, 2, \dots, k\}$, otherwise. Additionally, we considered the integer values of the shortest path between nodes as well as the subunitary values (i.e. replacing the integer value by its inverse).

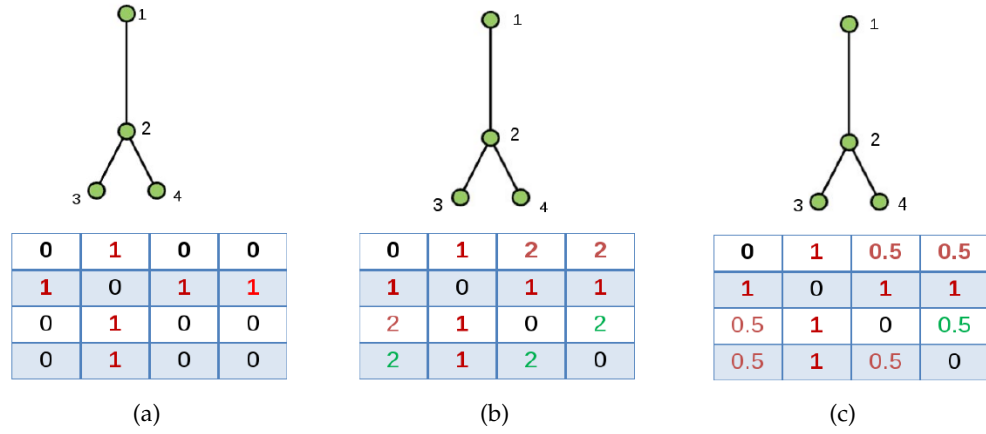


Figure 9.4: Adjacency matrix type for a 4-nodes graph example: (a) binary adjacency matrix, (b) adjacency matrix representing the length (natural integer) of the shortest path between nodes up to the 2nd degree, (c) adjacency matrix representing the length (natural integer) of the shortest path between nodes up to the 2nd degree

9.2.1 Many-to-many assignment framework parameter selection

Denoting by F the objective function in Equation 7.5 and by J_1, J_2 as all-ones matrices of size $N_K \times N_G$, and $N_K \times N_H$, the actual algorithm seeks the matching matrices P_1, P_2 that satisfy:

$$\min_{P_1, P_2} \{F - \lambda_s (\|P_1 - 0.5J_1\|_F^2 + \|P_2 - 0.5J_2\|_F^2 + c)\} \quad (9.1)$$

where λ_s is a sparsity penalization parameter and c is depending on the graphs size. The constraints are identical to those in Equation 7.5. We tested the method for various values of λ_s in the $[0, 1]$ interval, and chose the parameter configuration that resulted in the best matching.

The difference between the graphs size is handled either by setting $k_{max} \geq 2$ (hence allowing at most k_{max} vertices to be merged), or by setting $k_{max} = 1$ (hence allowing the implicit choice of nodes that will be assigned as dummy, within the graph having a larger dimension). Setting $k_{max} \geq 2$ requires careful tuning of λ_s , and doesn't always guarantee a good matching quality, therefore, we kept $k_{max} = 1$.

The initialization of the P_1 and P_2 matrices is extremely important as the non-convex problem is sensitive to it. In our experiments, we kept the initialization proposed by the authors, shown empirically to be a reasonable choice: $P_1 = \frac{1}{N_H} \mathbb{1}_{N_G} \mathbb{1}_{N_H}^T$ and P_2 , the identity matrix I .

9.2.2 Optimal transport framework parameter selection

For the loss function $L(u, v)$, we considered the quadratic version defined in [CSP16], and uniform weights associated to the graph vertices (i.e. $a = \frac{1}{N_G} \mathbb{1}_{N_G}$ and $b =$

$\frac{1}{N_H} \mathbb{1}_{N_H}$). We assume that a vertex i of graph G is matched to a vertex j of graph H , if its entire weight is transported to that of vertex j .

Regularizing the problem with an entropy term is shown to lead to a faster iterative algorithm. However, this also leads to a spread of mass from vertex i from graph G to multiple vertices of graph H . In order to recast this approach in the framework of graph-matching problems, we set $\epsilon = 0$.

Moreover, to avoid the mass spreading as a consequence of the different dimensions of the two graphs (e.g. when removing vertices), we added a dummy vertex (with no connection) to the graph with lower dimension (e.g. graph H) to which we assigned a mass that compensates the mass difference of vertices between the two graphs. Hence, this weight is equal to $1 - \frac{N_H}{N_G}$.

9.3 RESULTS AND INTERPRETATION OF THE COMPARATIVE PERFORMANCE ANALYSIS ON TOY-GRAPHS

In order to evaluate the performances of the two chosen methods (whose implementations are found online ^{2 3}) in a graph matching setting, for each of the previous pairs of graphs ($G - H$), we ran the algorithms for graph matching using many-to-many graph matching and structured optimal transport. In order to compare the performances of the two methods, we computed the cost of matching given by the difference between G , the adjacency matrix of the first graph, and PHP^T , the adjacency matrix of the matched graph, as $\|G - \text{PHP}^T\|_1$, where $\|A\|_1 = (\sum_i \sum_j |A_{ij}|)$.

Since we already know which is the expected assignment between the vertices of the simulated graphs, we computed the cost of the perfect matching for all of the cases mentioned above. [Table 9.1](#) and [Table 9.2](#) contain the matching cost and execution time in the case of the perfect matching (PM), many-to-many method (MM), and the optimal transport (OT).

We notice in the case of many-to-many assignment, that for both graphs, the rotation is handled perfectly for most cases, except when G and H are binary adjacency matrices. Additionally, removal of one degree node returns the expected assignment, while removal of a higher degree node is correctly handled only in the case of larger graphs.




In the case of the approach based on the optimal transport, the results indicate that similarly to the first method, the rotation of the graphs is handled well, except for the case when G and H are the adjacency matrices. A higher order of the shortest-path distance provides a more faithful representation of the graph adjacencies. However, in the case of larger-size graphs, increasing the order might also increase the time needed for the algorithm to converge to an optimal solution. This leads us to believe that a compromise (e.g. consider a 3rd order) might work best, as confirmed by the results.

Generally, the optimal transport fails to provide the expected matching in far more scenarios than the many-to-many assignment. For most of the experiments nonetheless, the OT method finds the result in a shorter time, corresponding to one order of magnitude for the smaller graphs, and up to two orders of magnitude

² <http://projects.cbio.mines-paristech.fr/graphm/mtmgm.html>

³ <https://github.com/gpeyre/2016-ICML-gromov-wasserstein>

Table 9.1: Matching cost for the perfect matching (PM), many to-many matching (MM) and optimal transport (OT). Graphs have 16 nodes and the 3 transformations are: rotation, removal of one-degree node, removal of multiple-degree node. G and H have the following representations: binary adjacency matrices (Int 1), shortest path integer values and subunitary values at order 2,3, total (Int 2, Int 3, Int T, Sub 2, Sub 3, Sub T).




Graph		PM	MM		OT	
Transf	G,H	Cost	Cost	Time(ms)	Cost	Time(ms)
	Int 1	0	16	70	52	30
	Int 2	0	24	50	0	5
	Int 3	0	0	20	0	4
	IntT	0	0	20	0	2
	Sub 2	0	0	20	0	4
	Sub 3	0	0	20	0	3
	SubT	0	0	10	0	2
	Int 1	2	2	20	58	70
	Int 2	10	10	20	86	7
	Int 3	34	34	20	186	6
	IntT	102	142	40	118	4
	Sub 2	4	4	30	32	5
	Sub 3	6.7	6.7	20	18	4
	SubT	10.4	10.4	20	17.5	2
	Int 1	6	14	60	58	40
	Int 2	42	82	90	114	5
	Int 3	124	196	100	232	3
	IntT	538	546	100	550	2
	Sub 2	15	29	60	67	5
	Sub 3	23.6	39	70	44.3	4
	SubT	40.2	50	40	57.6	2

for the largest ones. This may be a considerable advantage over the many-to-many graph matching technique, if considered for the modeling of real graph-networks.

9.4 CONCLUSIONS

Generally, when it comes to comparing undirected and unlabeled graph networks in an assignment framework, the shortest path seems to be a better choice compared to the binary adjacency matrix (G - H representation), as it incorporates more information about their topological structure. In terms of the matching cost, many-to-many matching performs better, as highlighted by the results. However, we

Table 9.2: Matching cost for the perfect matching (PM), many-to-many matching (MM) and optimal transport (OT). Graphs have 181 nodes and the transformations are: rotation, removal of one-degree node (Remove A), removal of multiple-degree node (Remove B). G and H have the following representations: binary adjacency matrices (Int1), shortest path integer values and subunitary values at order 2,3, total (Int2, Int3, IntT, Sub2, Sub3, SubT).

Graph		PM	MM		OT	
Transf	G,H	Cost	Cost	Time(ms)	Cost	Time(ms)
	Int 1	0	424	5600	1048	100
	Int 2	0	0	1800	0	80
	Int 3	0	0	1500	0	30
	IntT	0	0	2700	0	20
	Sub 2	0	0	2000	384	100
	Sub 3	0	0	1500	14.7	50
	SubT	0	0	1700	0	20
	Int 1	2	2	2300	1054	100
	Int 2	14	14	1900	4074	100
	Int 3	50	50	1500	50	40
	IntT	3766	3766	3000	5006	30
	Sub 2	5	5	5900	1000	100
	Sub 3	9	9	2000	29.7	80
	SubT	45.9	45.9	1900	105.2	40
	Int 1	8	8	4800	1044	100
	Int 2	56	56	2800	4324	90
	Int 3	196	196	2000	196	70
	IntT	4928	4928	2900	5916	30
	Sub 2	20	20	18600	108	20
	Sub 3	34.7	34.7	3000	552	100
	SubT	94	94	2900	662	90

found that having to set a sparsity penalization parameter to trigger the expected solution as well as the sensitivity of the non-convex problem to the initialization, to be important drawbacks. The method by optimal transport performs significantly faster. Adding a dummy node to avoid mass splitting allows us to employ this formulation as a one-to-one graph matching problem.

The results that we have obtained have a preliminary character, as we intend to explore these observations for further development of a computational model that is able to compare biological networks. Since the many-to-many assignment framework is more reliable in terms of the assignment quality for most of the scenarios, which is essential for the next steps, we decided to select this method

for further development. More specifically, we intend, based on the many-to-many assignment framework, to propose different methodologies to compute and define the prototype of a given set of graphs (Chapter 10). Additionally, the graph matching setting will be used within an analytic framework, to study the variation of certain fiber parameters between classes, where the matching (registration) of graphs mitigates the impact of the variability within the tissues (Chapter 11).

METHODOLOGIES FOR DERIVING THE REPRESENTATIVE OF A SET OF GRAPHS

In machine and object prototype learning, median computation (estimation) is an important technique for capturing the representative model of a given set of objects (patterns). Such principles are already found in machine learning methods, e.g. clustering algorithms, (k-means clustering, nearest-centroid classification, etc.). These approaches are based on the computation of centroids of sample sets represented as points in a vector space, and require the notion of distance or similarity between the samples to evaluate the dataset centroids. A generalization of these techniques is based on the computation of the Fréchet mean defined as an estimate of the "mean" or barycenter for objects situated in various metric spaces.

When the objects are modeled through graph representations, one of the main concerns becomes the learning of a representative (prototype model) of a set of graphs associated to objects belonging to the same class. For example, in applications where multiple different instances of the same objects are available, the task is to build a model prototype that describes the best the collection of samples.

Generalized median graphs were introduced in [Jia+01] and are defined as the graphs that have the smallest sum of distances to all the graphs in the set. If the previously defined graph is, in turn, a member of the set, then the median graph becomes the set median. The computation of the median is generally exponential in terms of the size of the input graphs (set median) and in terms of the input number of graphs (generalized median). In [Jia+01], the solution of the generalized median graph problem was based on a genetic search algorithm¹ relying on the graph-edit distance between the graphs. A different algorithmic approach (based on the edit distance) was adopted in [Muk+07], for the computation of the generalized median graph in the biological image analysis context, and for building a topological map of all pairs of the human chromosome.

We reiterate some notations in order to formalize the median graph problem:

NOTATIONS:

Let G be a *graph*, where $G = (V, E)$ is a set of nodes (vertices) connected by the set of edges $E \subset V \times V$. The structure of G can be encoded in a square adjacency matrix, A_G of size $|V| \times |V|$, where $(A_G)_{ij}$ is equal to 1 if node i is connected by an edge to node j , and 0 otherwise, also called a binary adjacency matrix.

¹ Genetic search algorithms [Gen] are used to model optimization problems, in a framework inspired by the process of natural selection (population fitness corresponds to the objective function) relying on bio-inspired operations: mutations, crossover, selection.

We refer to *real-valued adjacencies matrices (weighted)* if $(A_G)_{ij}$ represents the weight assigned to the edge between node i and j .

G is called an *undirected* graph when A_G is a symmetric matrix, i.e. $(A_G)_{ij} = (A_G)_{ji}$.

We refer to the *matching* between 2 graphs as the mapping that denotes the assignment between the nodes: $f : V^G \rightarrow V^H$. Denoting by $N_G = |V_G|$ and by $N_H = |V_H|$, the number of nodes of G and H , respectively, the assignment can be encoded into a binary correspondence matrix $P \in \{0, 1\}^{N_G \times N_H}$, such that $P_{i,j} = 1$ when the i -th node of G and the j -th node of H are matched and 0, otherwise.

[] the Iverson [Ive] bracket defined as:

$$[P] = \begin{cases} 1 & \text{if } P \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

PROBLEM FORMULATION: Let $\mathbf{G} = (G_1, G_2, \dots, G_n)$ be a set of weighted undirected and unlabeled graphs, with different dimensions. Given a distance function $d(\cdot, \cdot)$, the generalized median graph \hat{G} minimizes the sum of distances towards each graph G_i as follows:

$$\hat{G} = \operatorname{argmin} \sum_i^n d(\hat{G}, G_i) \quad (10.1)$$

An alternative representation is given by the *set median graph* which is the result of the same formulation in Equation 10.1, with the additional constraint that $\hat{G} \in \mathbf{G}$. $d(\cdot, \cdot)$ is often considered to reflect a similarity measure between a pair of graphs (e.g. the cost of matching provided by graph matching algorithms).

A different line of work (that was previously defined in Section 8.4.2) based on optimal transport [CSP16] defines barycenters for a set of pairwise similarity matrices (not explicitly graphs) using the Gromov-Wasserstein metric to compare and average point clouds. For the sake of completeness, we reiterate that if we consider graphs instead of similarity matrices, represented by the adjacency matrices $G_s \in \mathbb{R}^{N_s \times N_s}$, then the solution of the problem determining the barycenter graph of adjacency matrix $G \in \mathbb{R}^{N \times N}$, where N is user-determined and μ is the associated mass distribution, converges to $\frac{1}{\sum_s \mu_s} \sum_s \lambda_s P'_s{}^T G_s P'_s$, where μ_s are the mass distributions associated to G_s . $P'_s \in \mathbb{R}^{N \times N_s}$ are the transport matrices between the barycenter and the similarity matrices set. λ_s are considered weights so that $\sum_s \lambda_s = 1$.

This formulation returns (even when considering graphs as inputs) a "pairwise similarity matrix" - richer representation of the connection between the nodes which can potentially link all nodes together by weights, rather than an adjacency matrix for describing the graph, so additional steps need to be taken on the generated result to decide how to convert it to a binary adjacency matrix. However, interestingly, the term $\lambda_s P'_s{}^T G_s P'_s$ recovers a similarity to the formula of the permuted graph from the graph-matching section Equation 7.3, indicating that the solution to the barycenter based on Gromov-Wasserstein discrepancy converges to a sort of the

average of realigned matrices G_s . This idea serves thus as inspiration for the first approach proposed here to define the representative graph.

Finally, the authors in [Vay+18] show how barycenters of labeled graphs (e.g. recovering the barycenter of a set of noisy graphs) can be obtained, using a different metric, Fused Gromow-Wasserstein, that takes into account the node label information within its objective function. Our interest throughout this manuscript has been however focused on comparing and computing representative for unlabeled graphs, as the node labels for FN graphs are not explicitly defined.

10.1 METHODOLOGY BASED ON A MAJORITY VOTING AFTER MATCHING TO A COMMON GRAPH

The present methodology to construct a prototype graph G_b for a given collection of graphs, is based on the *many-to-many assignment* framework, defined in Section 7.2.3. We need to ensure equal dimension for the graphs, therefore, for the sake of simplicity, we added the necessary number of dummy nodes to each graph from G .

The basic principle is to start by matching the sequence of graphs in G to one of the graphs G_{init} , where $G_{init} \in G$ (e.g. the set median graph), using the many-to-many matching technique. This will result in a collection of permuted graphs M_i (isomorphic to G_i , but with permuted nodes to match G_{init}). Subsequently, for every possible edge [$G_b(jl) = 1$] between two nodes (j, l) , a decision on whether to preserve it is performed, based on the number of appearances of the same edge $M_i(jl)$ within the permuted graphs.

In [algorithm 1](#), the decision is made following a majority voting rule: if the edge $G_b(jl)$ is found in at least half of the permuted graphs M_i , $G_b(jl)$ is set to 1.

Algorithm 1: Representative graph methodology I

Result: Barycenter G_b
 initialization $G_b = G_{init}; \quad i = 1;$
while $i \leq n$ & $i \neq init$ **do**
 matching G_i to G_{init} : $M_i = P^T G_i P;$
 $i = i + 1;$
end
 $G_b(jl) = \lfloor \frac{\sum_i M_i(jl)}{2} \rfloor;$

The [algorithm 1](#) provides a representative graph G_b for a given set of graphs G . We consider that G is embedded in a 2D Cartesian space, i.e. nodes are located at positions indexed by 2D coordinate vector $v = (x, y)^T$, on a rectangular grid of size $M \times N$. In order to consider the spatial localisation of the nodes for the prototype graph development and, thus, assign G_b a localisation, there are several alternatives.

We can, for instance, after having established the structure of G_b , to consider for every node $G_b\{k\}$, the list of matched nodes (excepting dummy nodes), composed of one corresponding node for each G_i . Essentially, for a given node of the barycenter $G_b\{k\}$, $k \in \{1 \cdots N_G^2\}$, the list of nodes of G_i that were matched to it, is represented as a one-dimensional vector N_k , whose element $N_k(m)$ contains the index of the matched node in the m -th graph, where $m \in \{1, \cdots, n\}$. The spatial localisation can

thus be evaluated, as the median value (2D) of the spatial coordinates assigned to the matched nodes (e.g. for $G_b(k) : v_{G_b(k)} = \text{median}(v_{N_k})$).

However, it is relatively easy to see why such a formulation may lead to incoherent physical representations. The matching problem (Section 7.2.3) does not explicitly take into account the physical localisation of the nodes, it only considers the graph structure in the formulation. Hence, for an example where the graphs to be matched are isomorphic, but the corresponding nodes are found at different locations, the physical embedding of the barycenter resulted from the methodology illustrated above, tends to "cluster" the nodes towards the center of the spatial grid (Figure 10.1).

To overcome this shortcoming, instead of directly computing the median value of the spatial coordinates, one can resort to performing a registration of the node locations with respect with respect to G_{init} , prior to an evaluation of the median value. Algorithms such as Procrustes analysis [Pro], are able to determine a linear transformation (translation, reflection, orthogonal rotation, and scaling) of the coordinates of N_k lists, in order to register them with respect to the nodes of G_{init} , once the graph matching is obtained (hence correspondence node-to-node is known). Such a linear transformation is performed in an attempt to optimize a goodness-of-fit criterion (the sum of squared errors between the location of $G_{\text{init}}\{i\}$ and corresponding node of $G\{i\}$). Consequently, registering the coordinates of the matched nodes in the graph set onto the initial graph G_{init} mitigates the problem of different spatial localisations when producing the physical embedding of the prototype graph in the 2D space (Figure 10.1).

DISCUSSION: The quality of the representative graph can be linked to the matching quality and several examples (Figure 10.2, Figure 10.3, Figure 10.4, Figure 10.5) illustrate this fact. As resulted from Section 9.3, the many-to-many assignment framework is preferred here, as it takes into account the global structure of the graphs during matching, and additionally, provides overall better results during the tests scenarios illustrated in Chapter 9. Regarding the estimation of the spatial localisation of the representative graph, it is important to register the positions of the matched nodes with respect to the initial graph G_{init} before computing their average or median positions. Intuitively, this methodology expresses the representative graph as a subset of G_{init} , or alternatively, the cluster of connected edges from G_{init} that are matched in the other graphs from G . Inevitably, the structure and shape of G_b depends on that of G_{init} (Figure 10.6, Figure 10.7).

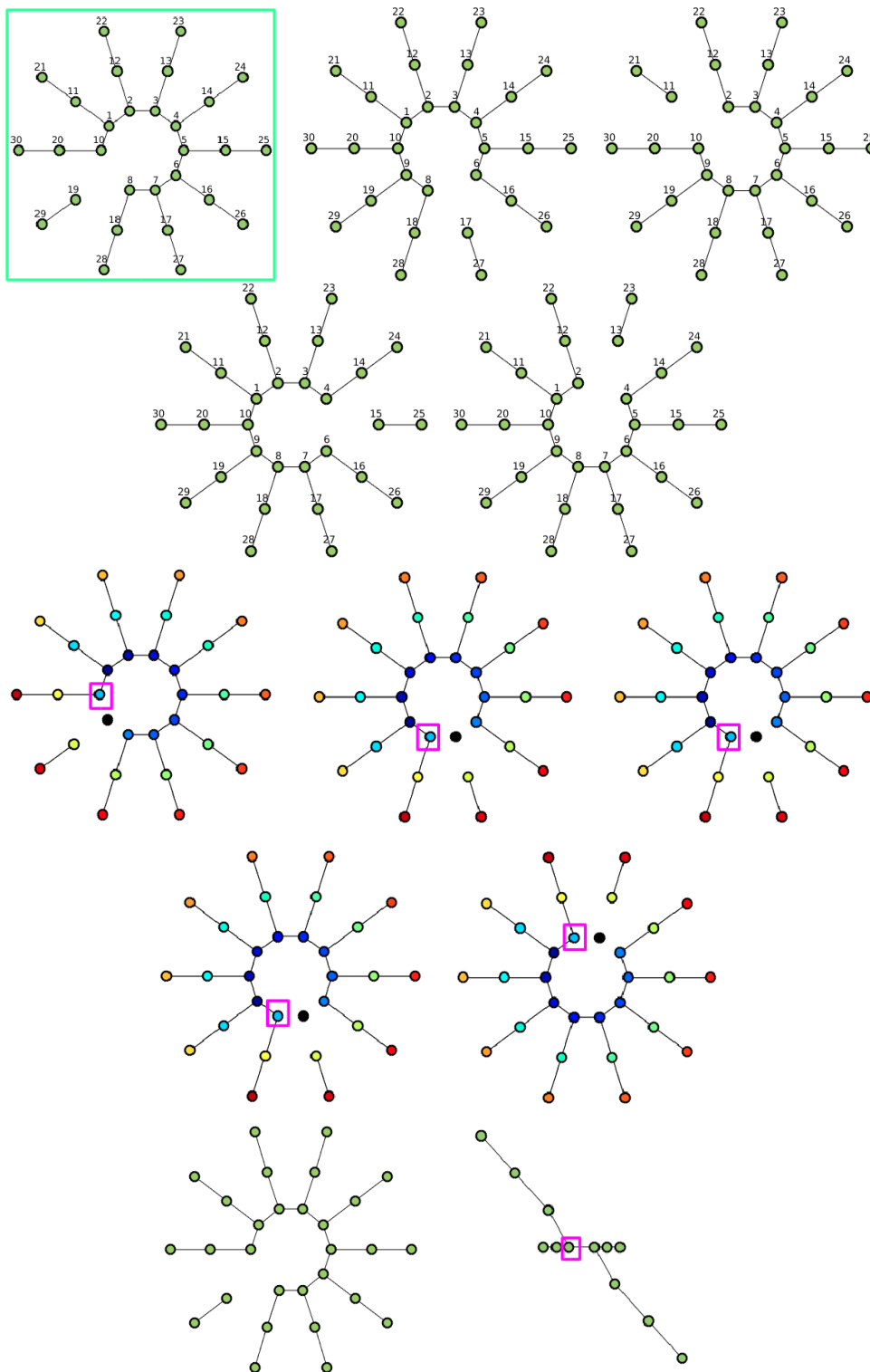


Figure 10.1: Prototype for isomorphic toy-graphs with different spatial localisation: First two rows: set of 5 toy-graphs. 3rd and 4th rows illustrate the toy-graphs matched onto the initial graph, which is the set median marked in a green rectangle. The nodes that are assigned together have the same colour. Last row (left) shows the prototype graph after physical registration of the nodes, the right side shows the same graph with a different physical embedding due to not having registered the nodes localisations onto the initial one. For a node marked with magenta rectangle, the localisation of its corresponding matching across the graphs is different, thus computing the median value returns a location towards the center.

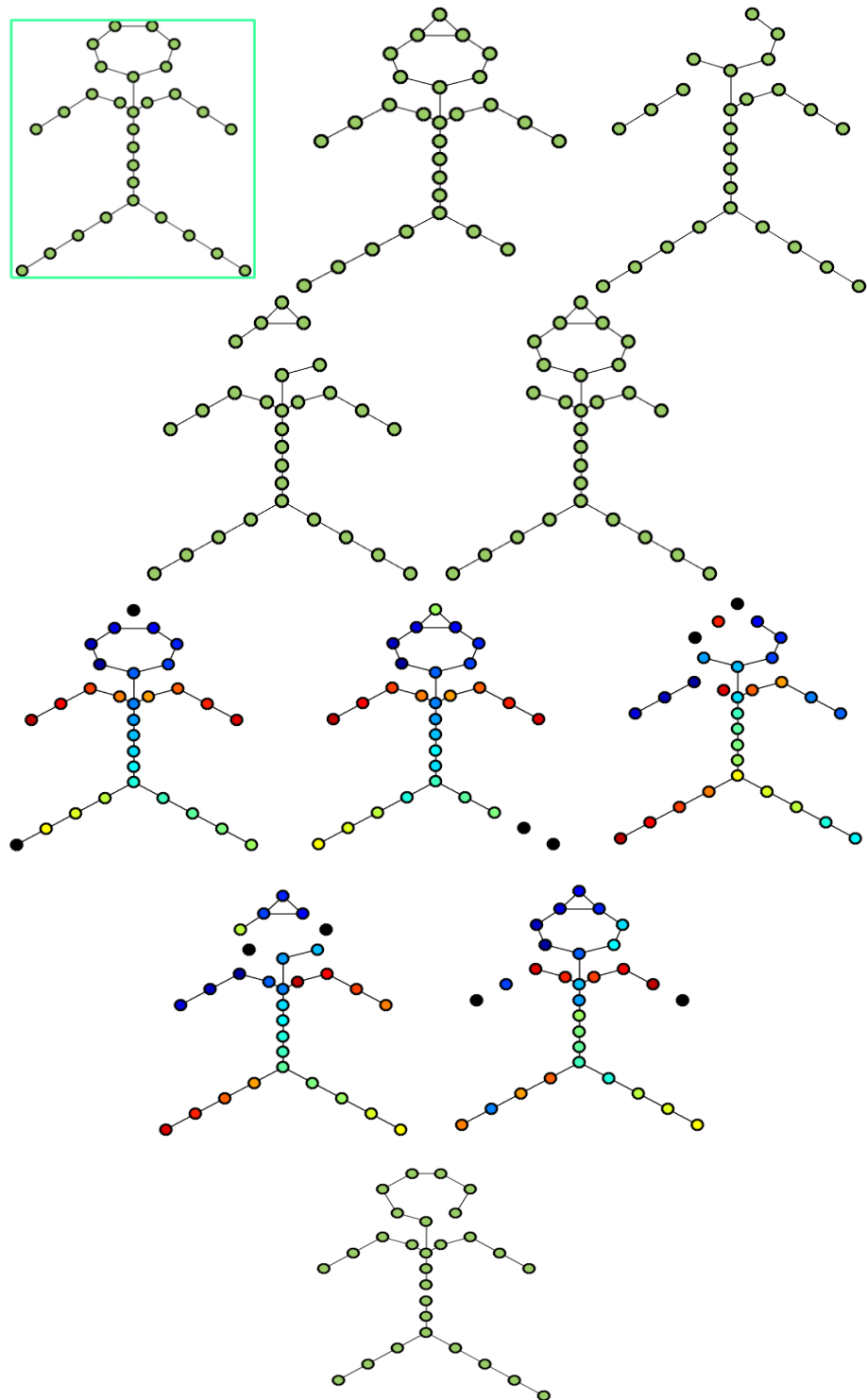


Figure 10.2: Prototype for random toy-graphs: First row: set of 5 toy-graphs. The second rows illustrates the toy-graphs matched onto the initial graph, which is the set median marked in a green rectangle. The nodes that are assigned together have the same colour. Last row shows the prototype graph after physical registration of the nodes. Parameters: $\lambda = 0$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are binary. Matching cost between G_i and G_b : 6, 14, 28, 22, 28.

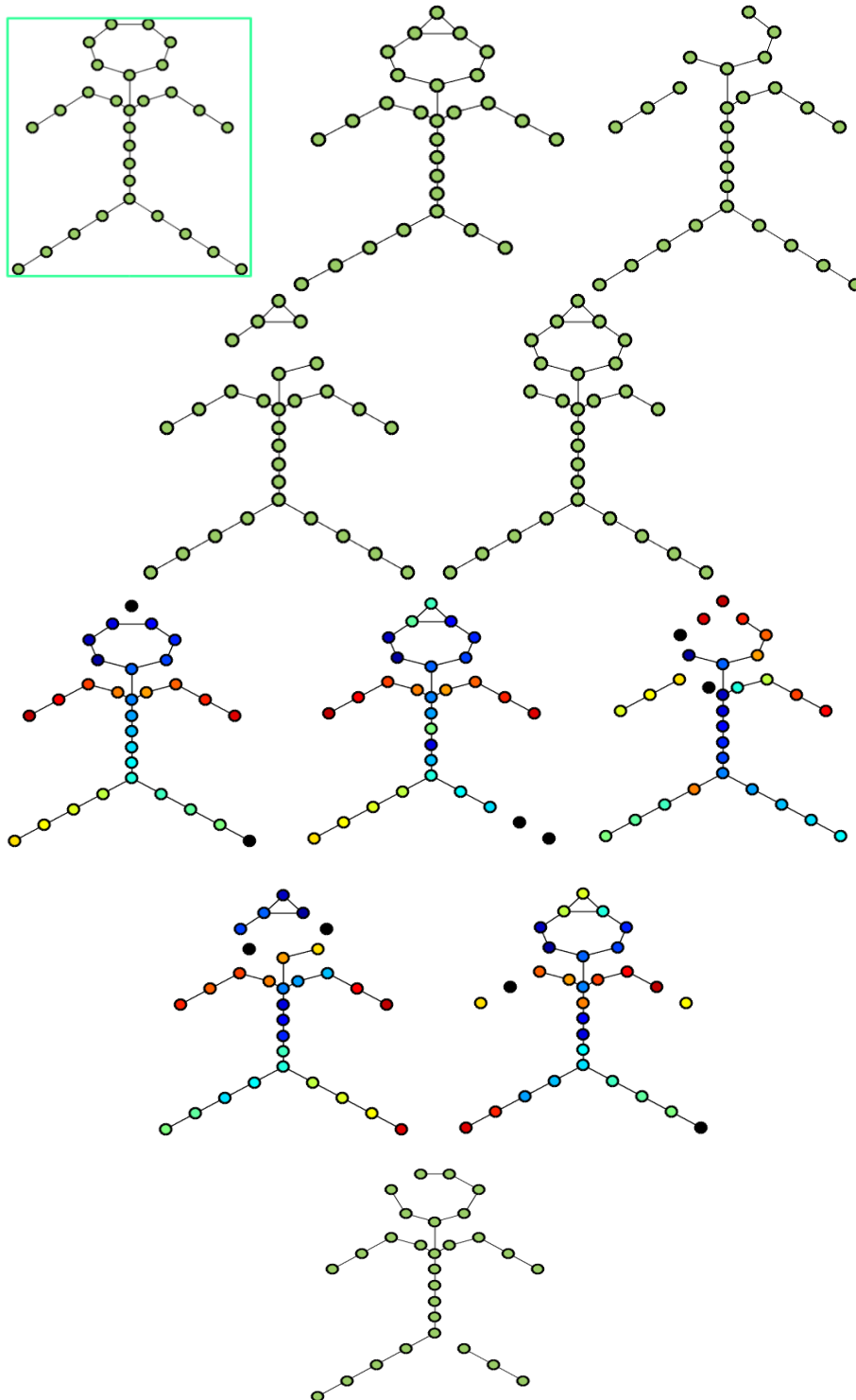


Figure 10.3: Prototype for random toy-graphs: First row: set of 5 toy-graphs. The second rows illustrates the toy-graphs matched onto the initial graph, which is the set median marked in a green rectangle. The nodes that are assigned together have the same colour. Last row shows the prototype graph after physical registration of the nodes. Parameters: $\lambda = 0.5$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are binary. Matching cost between G_i and G_b : 6, 22, 32, 34, 32.

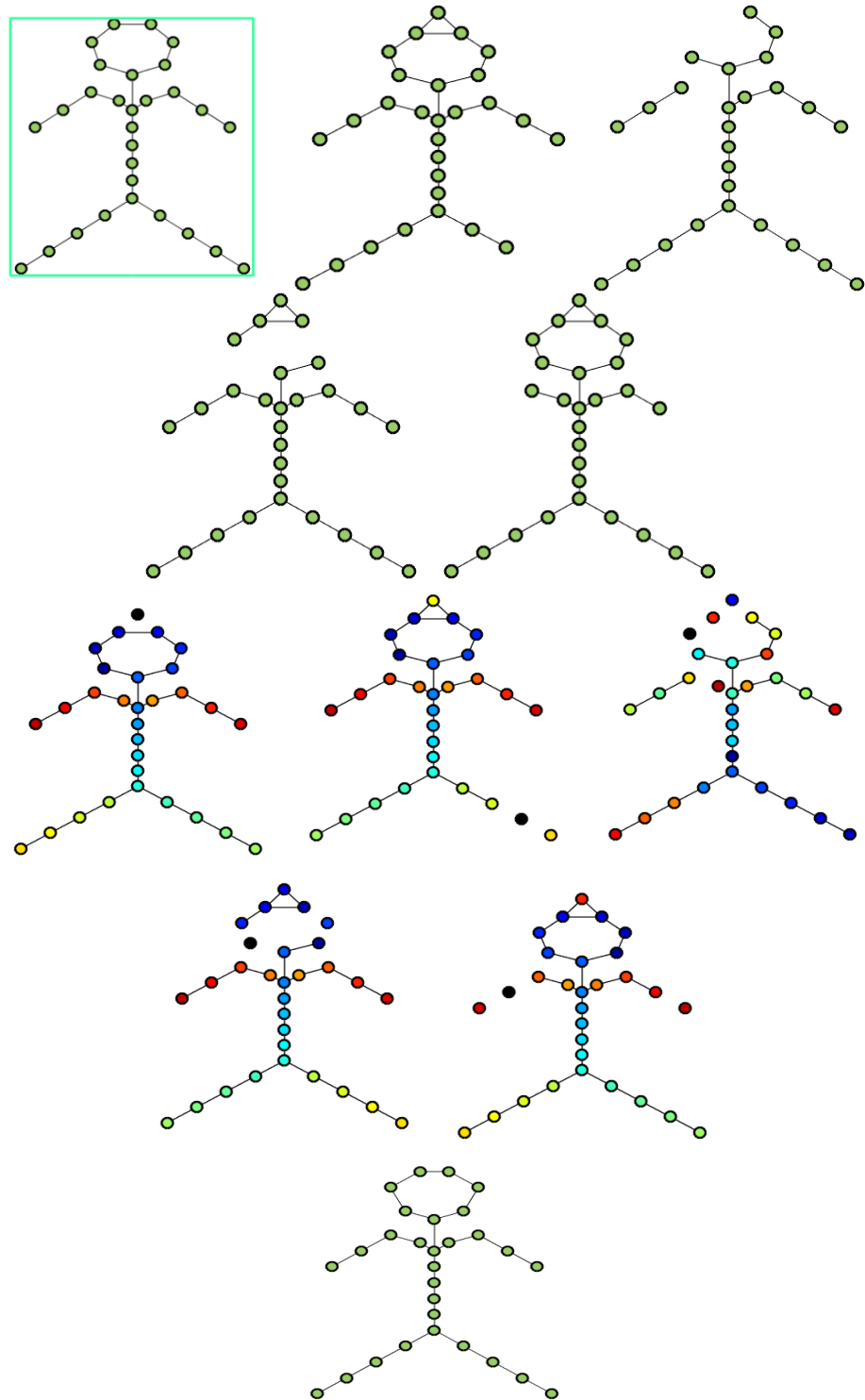


Figure 10.4: Prototype for random toy-graphs: First row: set of 5 toy-graphs. The second rows illustrates the toy-graphs matched onto the initial graph, which is the set median marked in a green rectangle. The nodes that are assigned together have the same colour. Last row shows the prototype graph after physical registration of the nodes. Parameters: $\lambda = 0.5$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are represented as the length of she shortest path of 3rd order. Matching cost between G_i and G_b : 0, 8, 58, 8, 10.

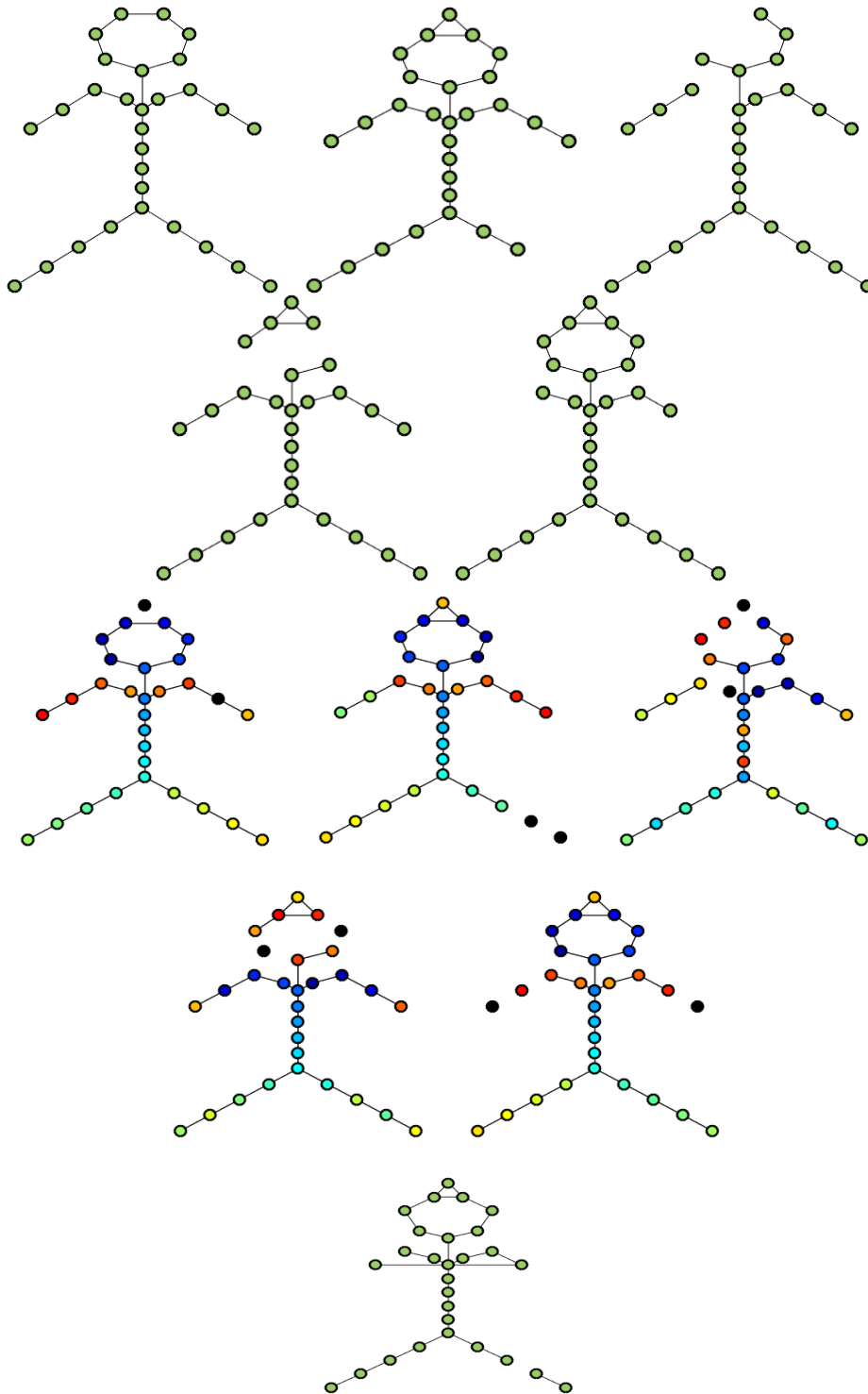


Figure 10.5: Prototype for random toy-graphs: First row: set of 5 toy-graphs. The second rows illustrates the toy-graphs matched onto the initial graph, which is the set median (fifth graph). The nodes that are assigned together have the same colour. Last row shows the prototype graph after physical registration of the nodes. Parameters: $\lambda = 0$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are represented as the length of she shortest path of 3rd order. Matching cost between G_i and G_b : 10, 2, 68, 54, 4..

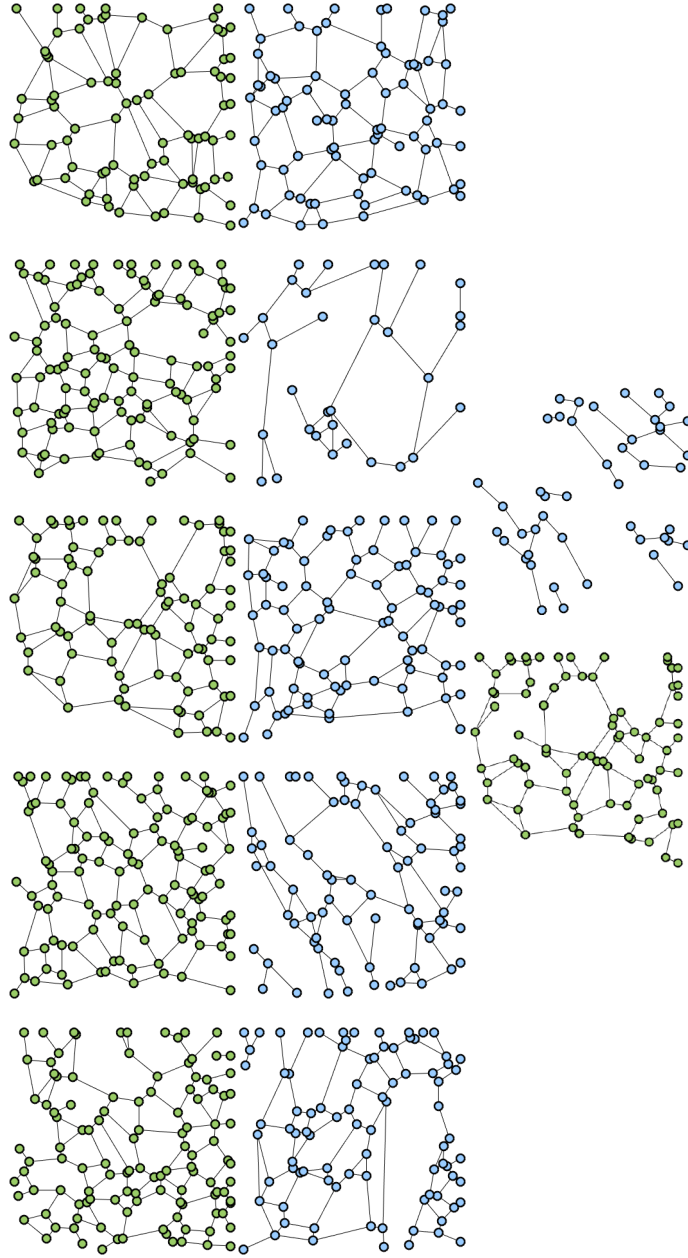


Figure 10.6: Prototype for FN graphs - FN B-A+ (Normal) ECM of 200×200 pixels FN confocal images. Last row shows the prototype graph for the "Normal ECM" (left, when the set median is the third graph) and for the "Tumour-like ECM" (right- when the set median is the second graph) after physical registration of the nodes. Parameters: $\lambda = 0$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are represented as the length of she shortest path of 3rd order.

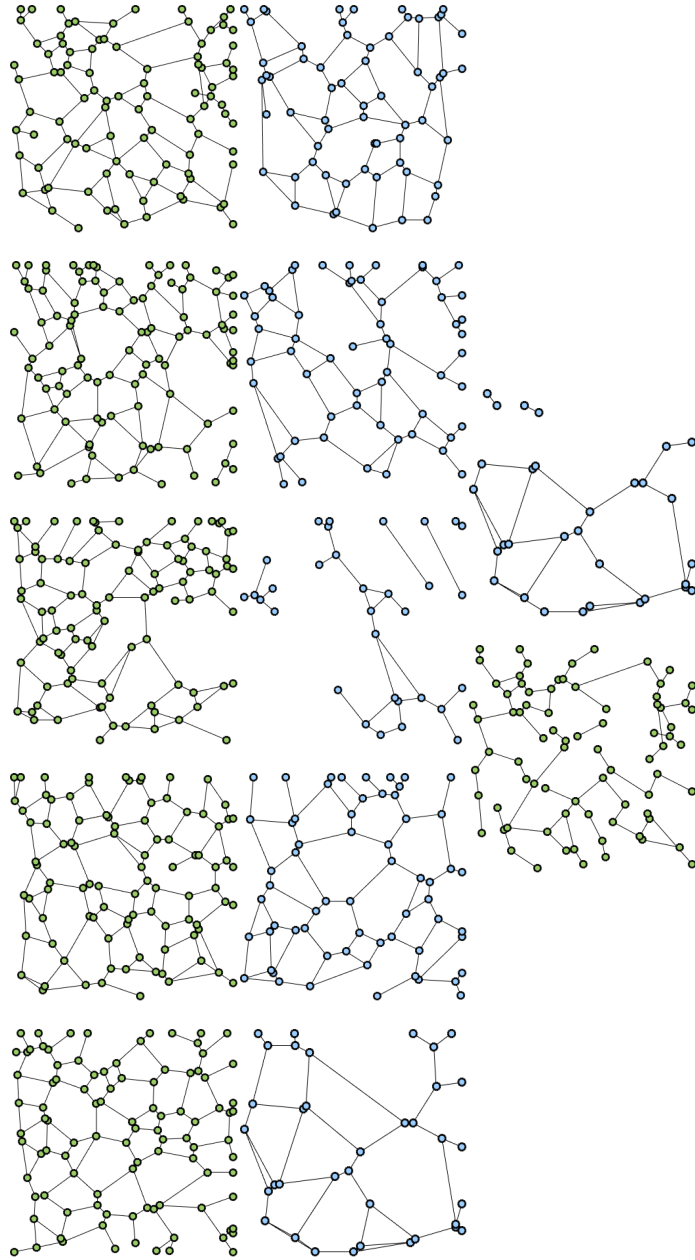


Figure 10.7: Prototype for FN graphs - FN B-A+ (Normal) ECM of 200×200 pixels FN confocal images. Last row shows the prototype graph for the "Normal ECM" (left, when the set median is the fifth graph) and for the "Tumour-like ECM" (right- when the set median is the first graph) after physical registration of the nodes. Parameters: $\lambda = 0$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are represented as the length of the shortest path of 3rd order.

10.2 METHODOLOGY BASED ON A HEURISTIC DERIVED FROM THE LONGEST CHAINS OF MATCHED NODES

The literature dedicated to matching multiple graphs among them, contains various methods that are bound to search for global consistent correspondences across the graph set [Yan+]. These methods have certain applications in shape matching, object recognition, etc. The principle of the present proposed heuristic aims at finding consistent mappings across the graphs with the purpose of defining a "median" graph. Therefore the objective is to find chains (cycles) of nodes that are connected to each other and furthermore, to nodes in the prototype, once the matching between every pair of graphs, individually and independently, is performed. We illustrate an alternative approach for the computation of the prototype graph, which similarly to the first method, is based on the *many-to-many assignment* framework, which serves as a tool to match every graph from \mathcal{G} , to each other.

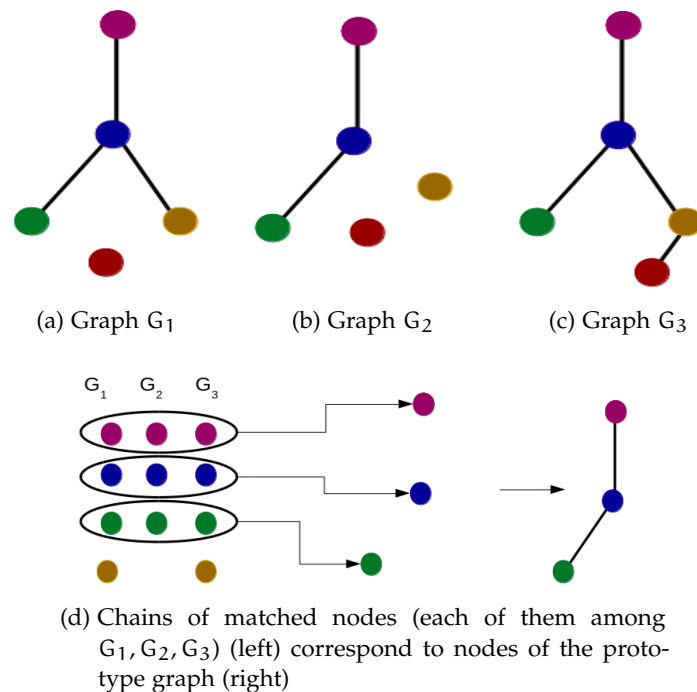


Figure 10.8: Methodology for deriving the representative graph based on chains of nodes connected among them across graphs G_1, G_2, G_3 (a,b,c): the nodes that are matched together between graphs have the same colour. The chains of nodes that are matched together, are pictured in (d): each chain corresponds to one node in the prototype graph. Edges are assigned if they are found in the majority of graphs between the corresponding nodes. Finally, the prototype graph of G_1, G_2, G_3 is displayed in (d) right.

The basic principle relies on identifying the longest chains of nodes (a chain has maximum one node per graph) (Figure 10.8) such that each node is connected to the others in the chain. For instance, if we consider the following 3-nodes sequence $\{G_i(a), G_j(b), G_k(c)\}$, where $(G_i, G_j, G_k) \in \mathcal{G}$, if the result of graph matching indicates the following assignment: $G_i(a)$ is matched to $G_j(b), G_k(c)$ and $G_j(b)$ is in turn matched to $G_k(c)$, then the sequence becomes a chain.

Once the complete list of chains is obtained (of length greater than 2 nodes, and in which nodes appear only once in the list of chains), the next step is to consider that each found chain will correspond to a node belonging to G_b , the representative graph of the set. Consequently, the dimension of the prototype graph can be determined as the total number of chains (in decreasing order of length) that are deemed characteristic enough (e.g. whose length is greater than 2 nodes).

The next steps are devoted to deciding the structure of this depicted median graph (i.e. the adjacencies between the nodes) followed by an assignment of the physical localisation to the nodes. Regarding the structure of the prototype graph, at this stage it is only defined through the presence of nodes corresponding to the selected chains. In order to assign the edges, a decision for a possible connection between 2 given nodes [$G_b(jl) = 1$] is taken, based on the number of appearances of this edge among the chains' nodes attributed to $G_b(j)$ and $G_b(l)$. In other words, for two given nodes, one needs to check if the nodes in the chains are themselves matched by an edge in the respective graphs, and if so, they are counted as one. Finally, the edges that are found in the majority of occurrences are kept as valid for determining the prototype structure. This procedure ensures a certain consistency with respect to the nodes and the edges of the collection of graphs.

As for the localization of the graphs' nodes, we proceed similarly as in the previous representative graph methodology, by performing a registration of the node locations with respect to G_{init} , ($G_{init} \in \mathcal{G}$ is the set median graph, for instance). To produce the physical embedding of the representative graph, a final step is the evaluation of the median value of the node localisations after their registration.

DISCUSSION: This methodology that essentially seeks to capture consistent chains of nodes across a given set of graphs, in order to represent the class prototype, is sensitive for graphs that present symmetries in their structure (Figure 10.9). It can, however, detect isomorphic graphs and provide the expected barycenter, as illustrated in (Figure 10.10).

10.3 CONCLUSIONS

Based on the many to many assignment framework, we proposed two different approaches to define and compute the prototype individual of a given set of graphs. The first one relies on the matching of all the graphs from the set to an initial one (e.g. set median), by keeping the edges that are matched in the majority of the graphs. This methodology provides satisfactory results even for FN-specific graphs, but is strongly dependant upon the choice of the initial graph.

The second methodology explores a different idea, so that the result is not dependant anymore on a certain graph. Briefly, it associates the nodes of the prototype graph, to the sequences of nodes that are matched together across graphs. This approach can construct a relevant prototype for isomorphic and relatively small-sized graphs. The main drawback is represented by its sensitivity to symmetries in graphs, therefore its adaptation to larger graphs remains still a work in progress.

Future perspectives include the classification of FN tissues, based on their graph representation, after defining a relevant prototype graph. Classifying a new FN sample would, in this case, involve a comparison (matching) to the different variants

representative individuals, and deciding, based on the similarity measure obtained after matching, the class to which the new sample will belong.

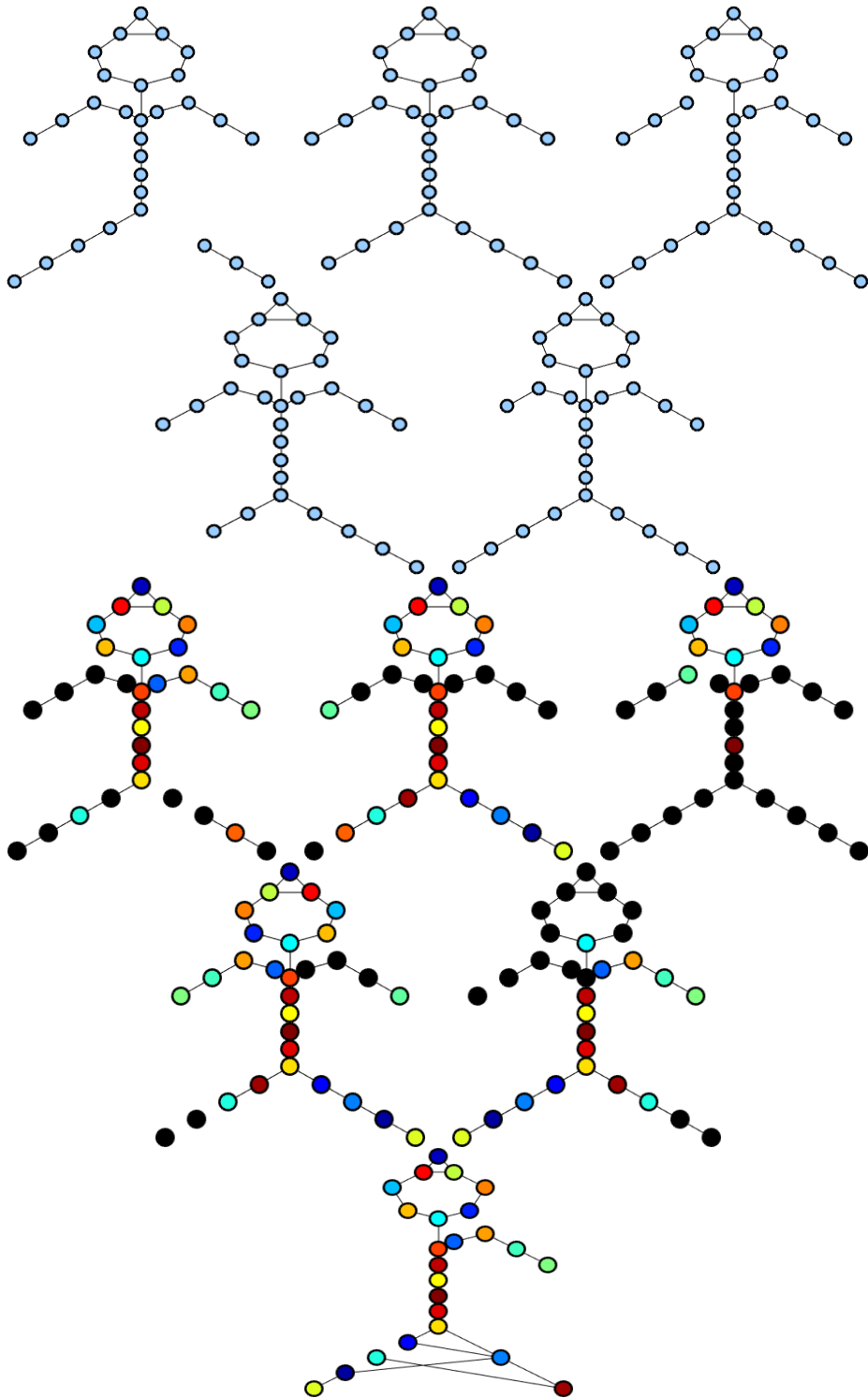


Figure 10.9: Prototype for random toy-graphs -Man Outline- Method based on the longest chains of matched nodes: First 2 rows: set of 5 toy-graphs. The third and fourth rows illustrate the chains of nodes that are matched all together (having the same colour across the graphs). Black nodes are either dummy or not part of any chain longer than 2 nodes). These chains have been selected after previous graph matching of every independent pair. The last row shows the prototype graph corresponding to the longest chains of nodes. Node registration was performed with respect to the first graph. Parameters for matching: $\lambda = 0$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are represented as the length of she shortest path of 3rd order.

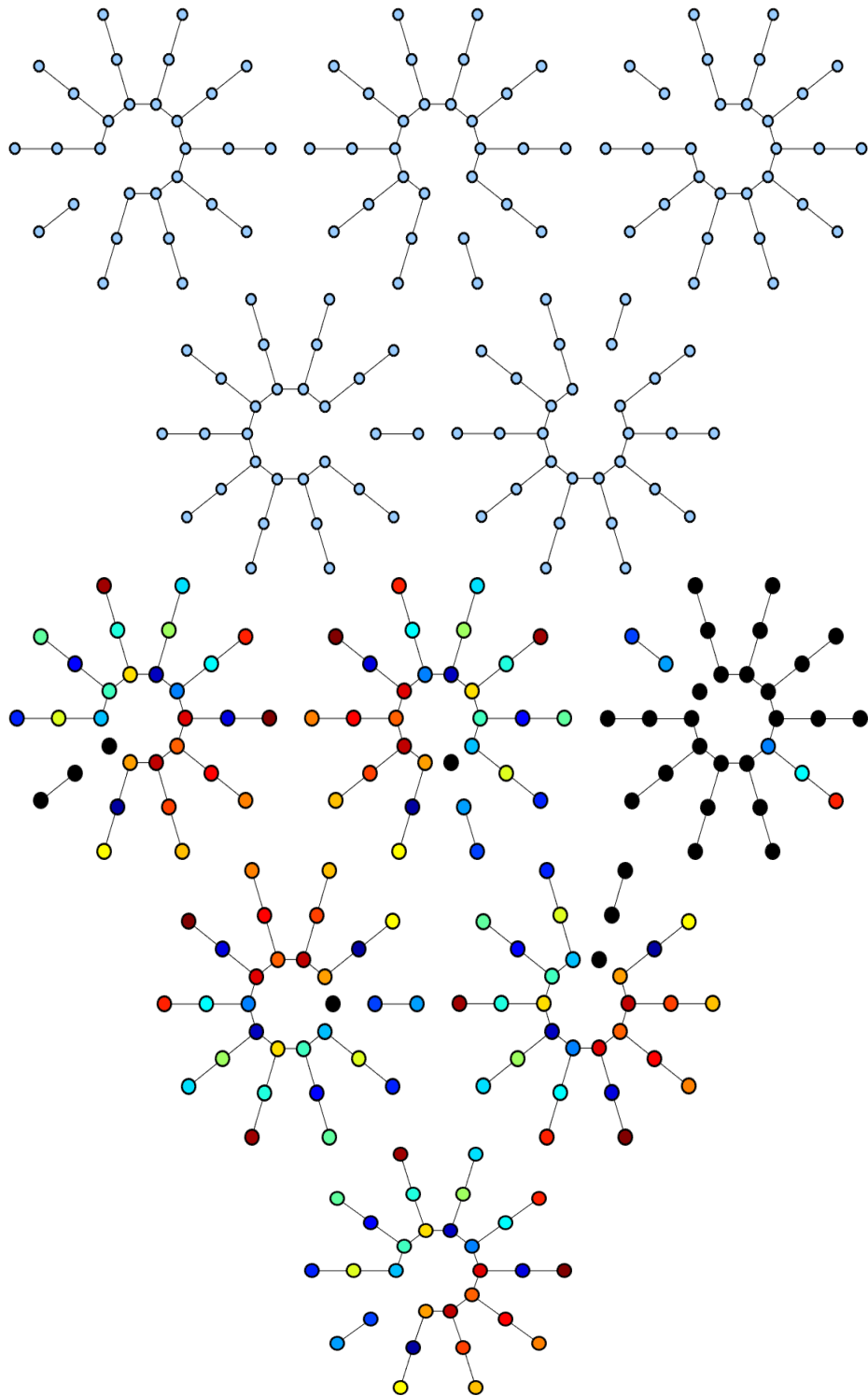


Figure 10.10: Prototype for random toy-graphs - Method based on the longest chains of matched nodes: First 2 rows: set of 5 toy-graphs. The third and fourth rows illustrate the chains of nodes that are matched all together (having the same colour across the graphs. Black nodes are either dummy or not part of any chain longer than 2 nodes). These chains have been selected after previous graph matching of every independent pair. The last row shows the prototype graph corresponding to the longest chains of nodes. Node registration was performed with respect to the first graph. Parameters for matching: $\lambda = 0$, $\lambda_s = 0.5$, sparsity penalization parameter, adjacency matrices are represented as the length of the shortest path of 3rd order.

 STATISTICAL ANALYSIS OF PARAMETRIC DEFORMATION MAPS

In this chapter, we revisit the statistical framework based on the random field theory, that can provide a comparison between the "normal" and "tumour-like" state FN variants, detailed in [Chapter 6](#), in the scope of studying deformation maps between the two classes. We remind the reader that the objective of the statistical analysis is to identify the foreign regions with respect to the GRF within both normal and tumoral deformation maps under the null hypothesis, and subsequently compare their properties (e.g. number, size).

The difference within this context, is that instead of studying the absolute parametric map (e.g. fiber length), we are interested in the analysis of fiber deformation maps that indicate the local differences between pairs of registered fibers (the result of matching their graph representations), with respect to their properties (e.g. length). The reason is that the relative analysis based on matching can lessen the impact of the existing variability within the same tissue.

Similarly to the approaches [Chapter 6](#), we implement two methodologies described in the referred chapter, the first one based on the theory of GRF, and the second relying on the computation of empirical distributions and showing the differences between the two classes, quantitatively and qualitatively.

11.1 DEFORMATION MAPS BASED ON GRAPH MATCHING

An interesting application of the graph-matching framework described in [Chapter 7](#), which provides a means to realign the fiber graphs upon matching of the nodes, is the possibility to compare the properties of the "matched fibers" between the realigned graphs. The graph matching technique that was used for this purpose is the many-to-many assignment framework.

We illustrate the methodology for computing the deformation maps between any 2 given FN specific graphs, in the simplified form (collection of nodes connected by edges) derived from the morphological skeleton extracted from the confocal images (see [Chapter 4](#) for further details on how to obtain this representation).

- Perform the graph matching between G and H , that will return a matching matrix P .
- If we consider $\{k, l\}$ two random nodes in H , such that H_{kl} is an edge (i.e. $H_{kl} = 1$) and $\{i, j\}$ the two corresponding matched nodes in G (not dummies), such that $P(i, k) = 1$ and $P(j, l) = 1$:

- if $G_{ij} \in \{1, 2, 3\}$ (upon considering the shortest path of order 3 between the nodes), we can compute the deformation d_l in terms of edge lengths as : $d_l = \text{length}(H_{kl}) - \text{length}(G_{ij})$ (see Figure 11.1).
- if $G_{ij} > 3$, or in the case of matching to dummy nodes, we set $d_l = 0$.
- Identification of the 2D pixel coordinates that approximate the straight line between the nodes and replacement of the pixels at the concerned locations with d_l .
- Extrapolation of the values of the length deformation map and smoothing with a Gaussian kernel (Figure 11.2, Figure 11.3).

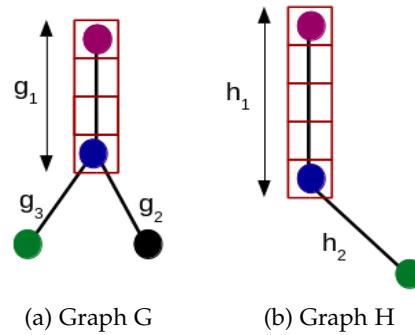


Figure 11.1: Computation of the fiber length difference after graph matching between G and H graphs: Nodes that are matched together have the same colour. The edge denoted as h_1 in the graph H, is thus matched to the edge g_1 in the graph G and the pixels corresponding to h_1 will be set to the value that is equal to $\text{length}(h_1) - \text{length}(g_1)$.

11.2 STATISTICAL FRAMEWORK TO QUANTIFY THE PARAMETER VARIATION

11.2.1 Statistical analysis of the deformation maps based on GRF

The methodology described in Section 6.2.1 was applied for the quantitative and qualitative comparison of the fiber length deformation map, for one variant FN B-A+ in both normal and tumoral-like state, under the following conditions:

- The matching is performed for both the normal FN and the tumour-like FN, between the corresponding graphs and one "Normal" FN graph sample, the result of which leads to the generation of the deformation maps (Normal-Normal and Normal-Tumoral).
- Learning dataset : 30 deformation maps (Normal)
- Test set: 60 fiber deformation maps (Tumoral) and 20 (Normal)
- Thresholds of intensity $T = [45 \ 60 \ 70 \ 80 \ 100]$
- $p = 0.05$. Clusters are considered as foreign to GRF if either P_S or P_H are less than equal to p .

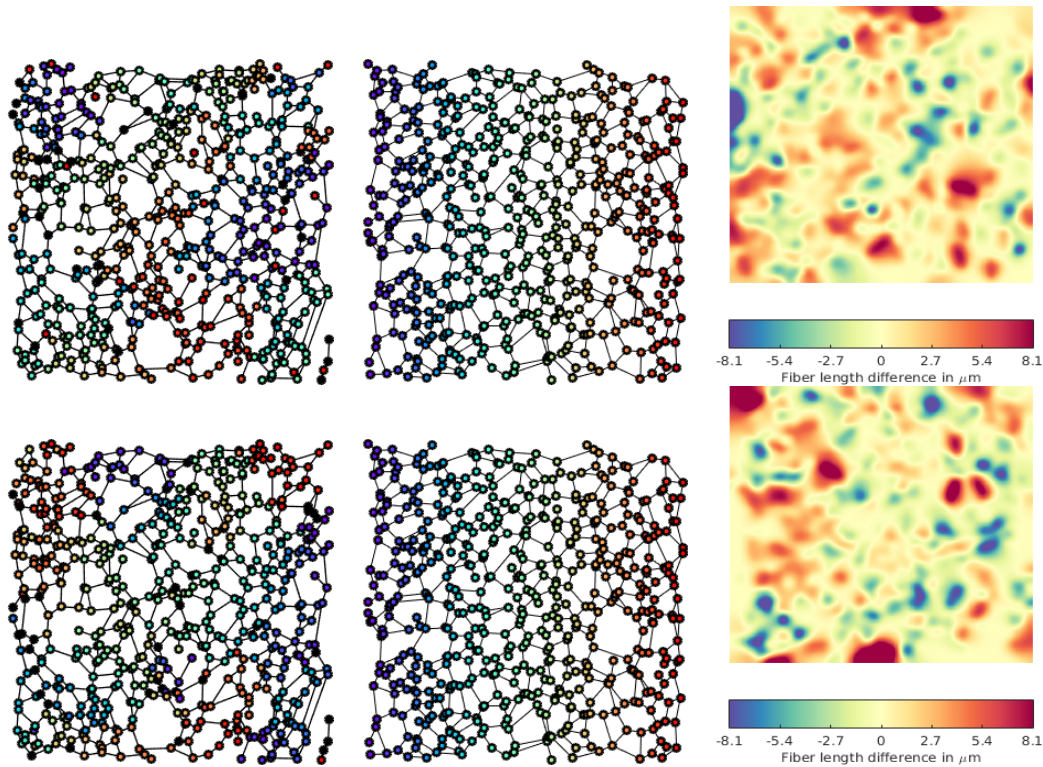


Figure 11.2: Computation of the fiber length difference after graph matching between graph representations of Normal-Normal FN B-A+ (left-right) of 512×512 pixels. Nodes that are matched together have the same colour. Nodes that are assigned as dummy (no connection) are set in black. The deformation map (whose legend is displayed in μm , pixel size is $0.27\mu\text{m}$), is displayed in the right column.

The results in [Table 11.1](#) and [Table 11.2](#) illustrate for every threshold, the average number of identified foreign clusters per image as well as the average cluster area. By comparison, within the normal-normal deformation maps of the test dataset, the method identifies less (foreign) clusters per image with a smaller average surface. The results in [Figure 11.4](#) and [Figure 11.5](#) illustrate several examples of detected clusters for a given p-value equal to 0.05.

Table 11.1: Average number and area of clusters per normal-tumoral deformation map FN B-A+ 512×512 identified as foreign to a GRF (based on the maximum intensity and cluster surface)($p = 0.05$), taken at various thresholds

Thresholds	45	60	70	80	100
Max Intensity - Avg nb/im	1.18	0.92	0.57	0.38	0.15
Max Intensity - Avg area/im	2949.90	1174.49	735.80	487.83	246.65
Surface- Avg nb/im	1.63	1.07	0.75	0.50	0.22
Surface - Avg area/im	2553.49	1134.85	729.10	479.16	246.37

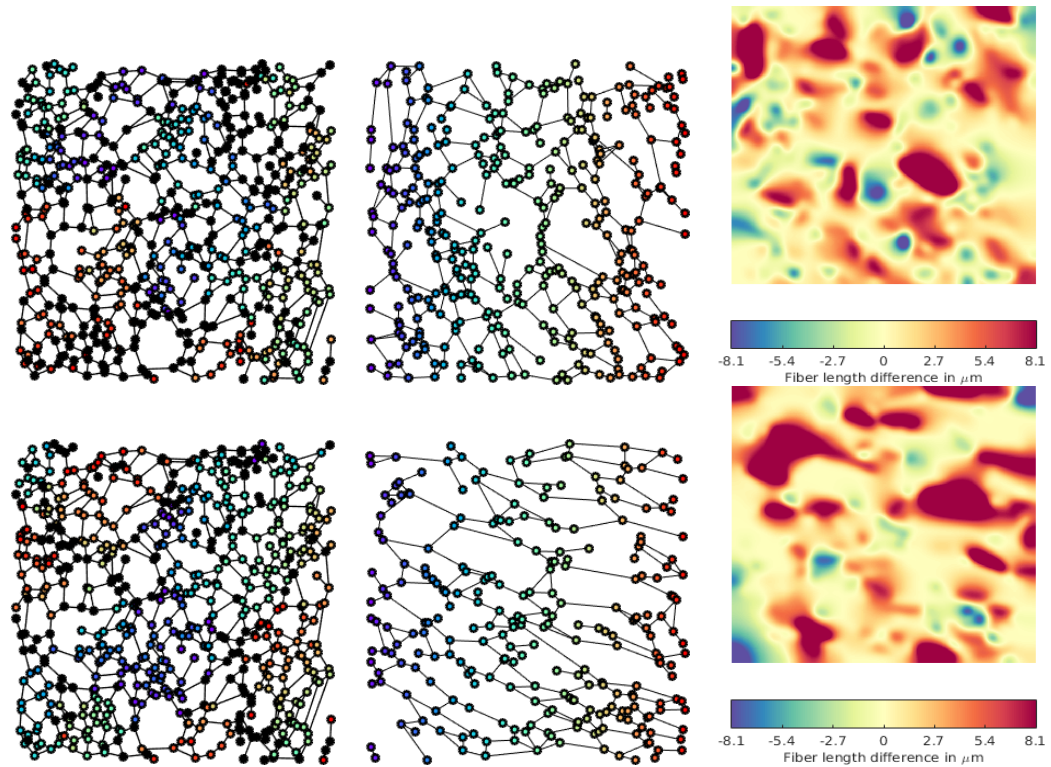


Figure 11.3: Computation of the fiber length difference after graph matching between graph representations of Normal-Tumoral FN B-A+ (left-right) of 512×512 pixels. Nodes that are matched together have the same colour. Nodes that are assigned as dummy (no connection) are set in black. The deformation map (whose legend is displayed in μm , pixel size is $0.27\mu\text{m}$), is displayed in the right column.

Table 11.2: Average number and area of clusters per normal-normal deformation map FN B-A+ 512×512 identified as foreign to a GRF (based on the maximum intensity and cluster surface) ($p = 0.05$), taken at various thresholds

Thresholds	45	60	70	80	100
Max Intensity - Avg nb/im	0.70	0.35	0.30	0.10	0.10
Max Intensity - Avg area/im	1570.56	558.55	255.00	123.95	19.70
Surface- Avg nb/im	1.30	0.50	0.35	0.20	0.10
Surface - Avg area/im	1483.13	556.28	244.03	117.08	19.00

11.2.2 Statistical analysis of the deformation maps based on the empirical distributions

The methodology, fully described in Section 6.2.2, will compute, for a given threshold t , the empirical cumulative histogram of maximum cluster intensities/surfaces for all the images in the learning set. This, in turn, will provide a certain threshold regarding either the cluster area or the cluster intensity that depends on the chosen p -value, above which the clusters from the test set taken at threshold t , if they exist, will be considered as foreign elements.

The results in [Table 11.3](#) and [Table 11.4](#) illustrate a similar pattern observed in the previous methodology: there is a higher number of detected clusters per image with a larger average area within the normal-tumoral deformation maps than in normal-normal.

- The matching is performed for both the normal FN and the tumour-like FN, between the corresponding graphs and one "Normal" FN graph sample, the result of which leads to the generation of the deformation maps (Normal-Normal and Normal-Tumoral).
- Learning dataset : 30 deformation maps (Normal)
- Test set: 60 fiber deformation maps (Tumoral) and 20 (Normal)
- Thresholds of intensity $T = [69\ 79\ 89\ 94\ 99]$
- $p = 0.05$.

Table 11.3: Average number and area of clusters per normal-tumoral deformation map FN B-A+ 512×512 identified as foreign ($p = 0.05$) to the empirical distributions of clusters (size and intensity), taken at various thresholds

Thresholds	69	79	89	94	99
Max Intensity - Avg nb/im	0.35	0.35	0.37	0.33	0.33
Max Intensity - Avg area/im	5303.18	4538.89	2568.84	2306.27	1562.82
Surface- Avg nb/im	0.38	0.48	0.40	0.33	0.33
Surface - Avg area/im	3939.29	3225.60	2119.56	2306.27	1562.82

Table 11.4: Average number and area of clusters per normal-normal deformation map FN B-A+ 512×512 identified as foreign, ($p = 0.05$) to the empirical distributions of clusters (size and intensity) taken at various thresholds

Thresholds	69	79	89	94	99
Max Intensity - Avg nb/im	0.15	0.15	0.15	0.15	0.05
Max Intensity - Avg area/im	2470.40	2050.78	1403.03	1046.23	653.85
Surface- Avg nb/im	0.15	0.15	0.15	0.15	0.05
Surface - Avg area/im	2470.40	2050.78	1403.03	1046.23	653.85

11.3 CONCLUSIONS

In this chapter we illustrated an application of the analytic framework described in [Chapter 6](#), for the study of FN fiber deformation maps (with respect to the fiber length) between normal and tumor-like FN states. We managed to show using both approaches (based on the GRF theory and on the computation of empirical distributions), that the tumoral aspect can be differentiated with respect to the normal-like state for one specific FN variant, and statistically characterized based on the fiber length.

The matching (alignment) between the images and the relative analysis are important to mitigate the effect of non-stationarity in different tissues (variability within the tissues). Future perspectives can include a more comprehensive study that analyses larger image samples and considers different FN fiber parameters.

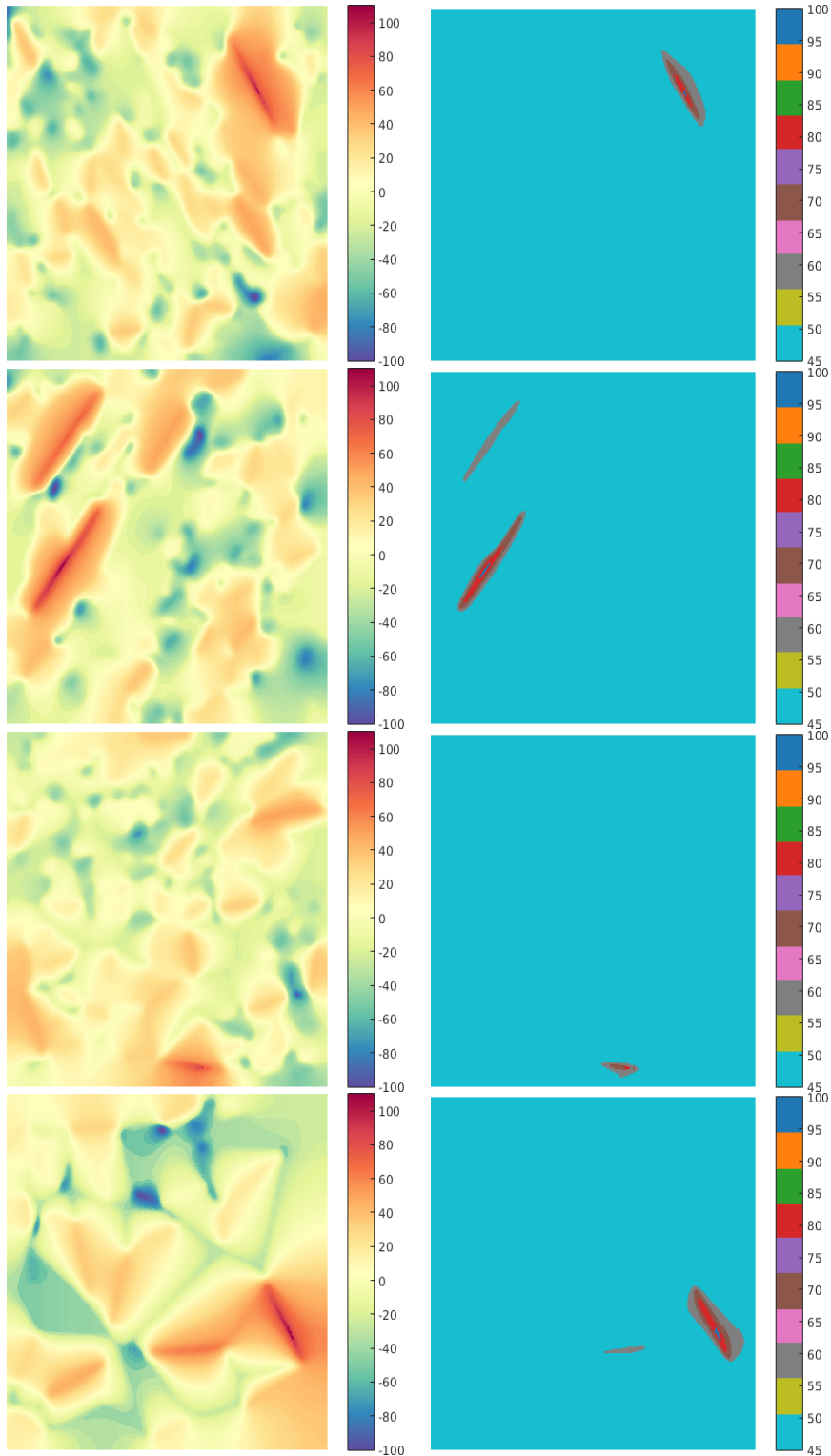


Figure 11.4: Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the maximum cluster intensity), within the Fiber Length Deformation Map Normal-Tumoral (left column) corresponding to Tumour-like FN B-A+ of 512×512 . The right columns depicts the clusters at different thresholds (indicated in the colorbar).

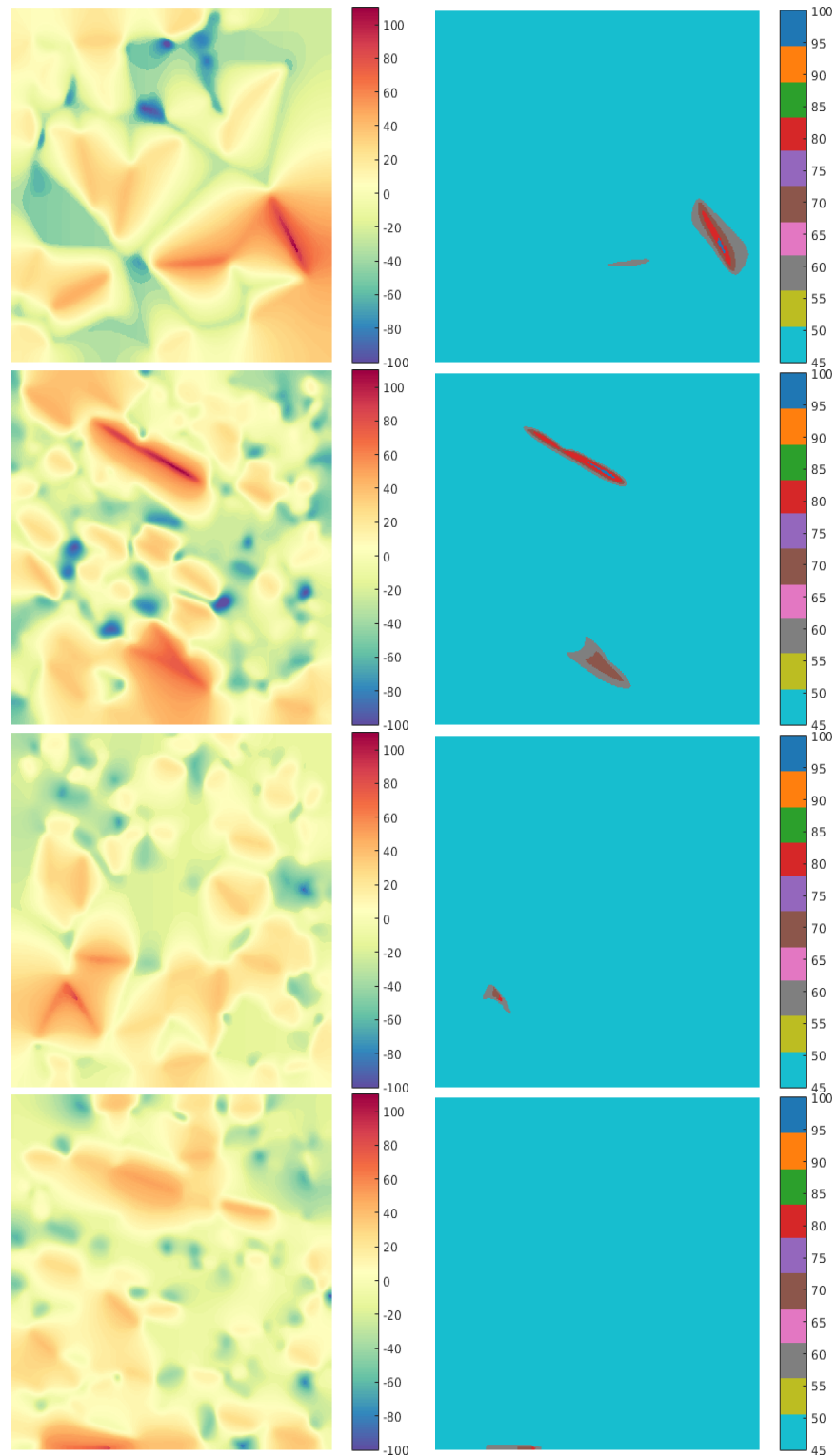


Figure 11.5: Detection of the clusters considered foreign elements to a Gaussian Random Field when $p = 0.05$ (based on the surface of the cluster), within the Fiber Length Deformation Map Normal-Tumoral (left column) corresponding to Tumour-like FN B-A+ of 512×512 . The right columns depicts the clusters at different thresholds (indicated in the colorbar).

Part VI

CONCLUSIONS AND PERSPECTIVES

CONCLUSIONS AND PERSPECTIVES

This manuscript presented our contributions in terms of computational approaches, designed in order to address a biological problem. These contributions were thus developed as the result of a collaboration within an interdisciplinary context.

More specifically, we have presented a set of numerical approaches that provide visual and quantitative characterization of FN variant-specific matrix architecture, in normal and tumor-like states. This image analysis approach is not only important for understanding the dynamics of matrix assembly and remodeling during tumor progression but it lays the foundation for development of improved diagnostic and predictive models of stromal features of tumor tissue for clinical purposes.

Fiber-specific attributes concerning the geometry and topology of the FN networks were extracted from the confocal images. Having derived a graph-based fiber representation, enabled a powerful characterization both locally (through features extracted from the graphs), as well as on a global level. Concerning the latter, a global characterization was obtained using a statistical analytic framework for FN class comparison (between normal and tumoral state) through a parameter variation map study. Additionally, a similar analysis was performed to study the parameter variation across deformation maps, facilitated by a graph-matching setting.

In the following sections, we present in detail our contributions that have, in turn, generated new perspectives and ideas that we briefly attempt to summarize subsequently.

12.1 CONCLUSIONS

CLASSIFICATION: First, we showed through a proposed classification pipeline based on curvelet features and alternatively, using a deep-learning method, that the confocal images contain enough information of the FN variants to be able to differentiate them. The performances were similar to that of a trained specialist, validating the choice of curvelets as feature extractors. A part of these results were published in [Gra+18] and [Eft+].

GRAPH-BASED FIBER REPRESENTATION: We constructed a graph-based representation of the FN networks, based on Gabor filter detection and subsequent morphological operations that resulted in a fiber skeleton. We additionally proposed an approach to improve the graph representation, by reconnecting the missing fibers in the skeleton due to image noise or imperfect skeletonization.

LOCAL CHARACTERIZATION OF THE FN FEATURES: We established a characterization of the FN networks based on local geometrical features, extracted from the graph representation (e.g. fiber network connectivity -node degree, fiber length, median pore dimension, etc.), Gabor features (e.g. fiber local thickness), showing that the FN variants can be compared and distinguished both from the PCA analysis and local parameter distributions across classes, as well as feature classification. Indeed, the graph-based representation embeds relevant and meaningful information about the fibers. These results were integrated in a biology-related article [Eft+].

GLOBAL CHARACTERIZATION OF THE FN PARAMETRIC MAPS: To study the variation of certain parameters within a given FN class (e.g. fiber length) both in normal state and tumour-like state, we applied a statistical analysis framework based on Gaussian random fields. We managed to show the differences between the two classes, quantitatively and qualitatively.

COMPARISON BETWEEN TWO APPROACHES FOR GRAPH MATCHING: Methods based on the many-to-many assignment framework and the discrete optimal transport were identified as being relevant to our graph comparison/alignment purpose. Therefore, we conducted a performance analysis of these approaches, adapted to a graph matching setting, and using randomly generated graphs. The results, which are relevant for understanding how these methods can be used for FN graph comparison, were published in [Gra+19].

PROTOTYPE (REPRESENTATIVE) GRAPH: Based on the many-to-many assignment framework, two different methodologies for defining and computing the prototype of a given set of graphs were proposed, in the hope of defining the representative graph for a given FN class. Although the proposed methods provide a schematic for a relevant prototype graph, the question of determining the representative graph remains largely a work in progress.

GLOBAL CHARACTERIZATION OF THE FN DEFORMATION MAPS: Under the same statistical framework used for the study of FN parametric maps, we analyzed the variation of a given parameter, e.g. fiber length between the variants in normal and tumour-like states. Graph matching was performed before to account for the variability of fiber organization within the same image. We obtained quantitative/qualitative differences between normal-normal and normal-tumoral FN deformation maps with respect to the fiber length, illustrating that the variability of the fiber length is mostly due not to variance within the same class, but to the differences between the classes.

12.2 PERSPECTIVES

GRAPH-BASED REPRESENTATION The methodology follows a certain pipeline starting with fiber detection, morphological skeletonization, graph-association and missing fiber reconnection. Visually, the graphs seems to extract the architecture of the 2D fibers, without taking into account the 3D structure (as the depth of the

tissue is not relevant here). That generates a representation that is not completely accurate for a 3D structure, that will, in turn affect the fiber reconnection.

On one hand, since the confocal microscope is capable of accessing the 3D structure, new ideas could be envisaged for building a 3D model of the fibers, and adapt the whole pipeline to 3D images. On the other hand, for the 2D model, the graph representation might be further improved in the post-processing step, by designing better thresholds when deciding to reconnect the fibers, taking into account the local average orientation of the fibers, image gradient magnitude (smoothed) instead of image local intensity as guideline for reconstruction, etc.

LOCAL AND GLOBAL CHARACTERIZATION OF PARAMETRIC MAPS The results of the statistical analysis of parametric maps (fiber length) were encouraging with respect to both qualitative and quantitative differences among FN variants in normal and tumor-like cases. However, a more comprehensive analysis would be interesting, by extending this study to the analysis of larger samples, and considering other local parameters specific of fibers (e.g. average fiber thickness).

GRAPH MATCHING The literature concerning graph matching contains approaches (that were not yet investigated during this thesis) that are promising with respect to the computational time and matching quality [ZDLT]. An exhaustive study of graph matching methods could be envisaged to choose the most suitable technique for FN graph-based comparison.

PROTOTYPE GRAPH Several works in a biomedical/biological context have already proposed various methodologies [Jia+01; Muk+07] based on different graph-matching distances, that might be explored or adapted to the problem of determining the prototype graph of a given set of FN graphs. Once a good model is determined, deformation maps can be computed with respect to this representative graph. Additionally, graph classification can be envisaged (by comparison of the test samples to the prototype graph).

OTHER APPLICATIONS OF FIBER NETWORK MODELLING The set of methodologies presented in this manuscript have the potential to be applied for the analysis and numerical characterization of different types of fibrillar networks encountered in various fields (biology, material science, etc.), and thus compared to existing methods analysing properties of fibrillar networks in biomedical sciences ([SB09], [Bou+14], [Hot+15]). Depending on the specific application, or image dataset type (acquisition mode, dimension), some steps of the pipeline need to be better adapted to construct a representation that captures relevant information (e.g. Gabor filtering might need to be preceded by extra-filtering, binarization, skeletonization and fiber reconnection tool steps have to be properly adjusted, etc.).

BIBLIOGRAPHY

- [Adl] Robert J Adler. *Some Expectations 5.1 A general result* (2010). *The Geometry of Random Fields*, 93–121. Tech. rep. URL: <http://www.siam.org/journals/ojsa.php>.
- [AD93] H. Almohamad and S. Duffuaa. “A Linear Programming Approach for the Weighted Graph Matching Problem.” In: *IEEE Transactions on pattern analysis and machine intelligence* (1993).
- [AMJJ18] David Alvarez-Melis, Tommi S. Jaakkola, and Stefanie Jegelka. “Structured Optimal Transport.” In: *International Conference on Artificial Intelligence and Statistics* 84 (2018), pp. 1771–1780.
- [Hun] *Assignment problem*. https://en.wikipedia.org/wiki/Assignment_problem. Accessed: 2019-11-29.
- [BYH04] Xiao Bai, Hang Yu, and Edwin R. Hancock. “Graph matching using spectral embedding and alignment.” In: *Proceedings - International Conference on Pattern Recognition*. Vol. 3. 2004, pp. 398–401. ISBN: 0769521282. DOI: 10.1109/ICPR.2004.1334550.
- [BBV01] Stefano Berretti, Alberto Del Bimbo, and Enrico Vicario. “Efficient Matching and Indexing of Graph Models in Content-Based Retrieval.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2001).
- [Bou+14] Arezki Boudaoud, Agata Burian, Dorota Borowska-Wykrk, Magalie Uyttewaal, Roman Wrzalik, Dorota Kwiatkowska, and Olivier Hamant. “FibrilTool, an ImageJ plug-in to quantify fibrillar structures in raw microscopy images.” In: *Nature Protocols* 9.2 (2014), pp. 457–463. ISSN: 17502799. DOI: 10.1038/nprot.2014.024.
- [Bre91] Yann Brenier. “Polar Factorization and Monotone Rearrangement of Vector-Valued Functions.” In: *Communications on Pure and Applied Mathematics* (1991). DOI: <https://doi.org/10.1002/cpa.3160440402>.
- [Bre] *Bresenham’s line algorithm*. https://en.wikipedia.org/wiki/Bresenham's_line_algorithm. Accessed: 2020-28-01.
- [CK04] Terry Caelli and Serhiy Kosinov. “An Eigenspace Projection Clustering Method for Inexact Graph Matching.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2004).
- [Can+05] Emmanuel Candès, Laurent Demanet, David Donoho, and Lexing Ying. *Fast Discrete Curvelet Transforms*. Tech. rep. 2005. URL: <http://www.curvelet.org>.
- [CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A Library for Support Vector Machines*. Tech. rep. 2001. URL: www.csie.ntu.edu.tw/.
- [CM06] J. Cho and D.F. Mosher. “Role of fibronectin assembly in platelet thrombus formation.” In: *Journal of Thrombosis and Haemostasis* (2006).

- [Con] *Confocal Microscope*. <https://micro.magnet.fsu.edu/primer/techniques/confocal/confocalintroduction.html>. Accessed: 2020-07-01.
- [Con+11] Matthew W. Conklin, Jens C. Eickhoff, Kristin M. Ricking, Carolyn A. Pehlke, Kevin W. Eliceiri, Paolo P. Provenzano, Andreas Friedl, and Patricia J. Keely. "Aligned collagen is a prognostic signature for survival in human breast carcinoma." In: *American Journal of Pathology* 178.3 (2011), pp. 1221–1232. ISSN: 15252191. DOI: 10.1016/j.ajpath.2010.11.076.
- [Con+07] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. "How and Why Pattern Recognition and Computer Vision Applications Use Graphs." In: *Applied Graph Theory in Computer Vision and Pattern Recognition. Studies in Computational Intelligence* (2007). URL: <https://doi.org/10.1007>.
- [Cnn] *Convolutional neural networks*. <http://cs231n.github.io/convolutional-networks/>. Accessed: 2020-07-02.
- [Cor+04] Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. "A (sub)graph isomorphism algorithm for matching large graphs." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.10 (Oct. 2004), pp. 1367–1372. ISSN: 01628828. DOI: 10.1109/TPAMI.2004.75.
- [Cou+17] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. "Optimal Transport for Domain Adaptation." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).
- [Cse+10] Botond Cseh, Samantha Fernandez-Sauze, Dominique Grall, Sébastien Schaub, Eszter Doma, and Ellen Van Obberghen-Schilling. "Autocrine fibronectin directs matrix assembly and crosstalk between cell-matrix and cell-cell adhesion in vascular endothelial cells." In: *Journal of Cell Science* 123.22 (Nov. 2010), pp. 3989–3999. ISSN: 00219533. DOI: 10.1242/jcs.073346.
- [Csu+] Gabriella Csurka, Christopher R Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. *Visual Categorization with Bags of Keypoints*. URL: <https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/csurka-eccv-04.pdf>.
- [Cut13] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In: *Advances in Neural Information Processing Systems* 26 (2013). URL: <http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf>.
- [CSP16] Marco Cuturi, Justin Solomon, and Gabriel Peyré. *Gromov-wasserstein averaging of kernel and distance matrices*. Tech. rep. 2016. URL: <http://dl.acm.org/citation.cfm?id=3045671http://hdl.handle.net/1721.1/112918>.
- [Dar+90] Thierry Darribere, Kareen Guida, Hannu Larjava, Kurt E Johnson, Kenneth M Yamada, Jean-Paul Thiery, and Jean-Claude Boucaut. "In Vivo Analyses of Integrin/ 1 Subunit Function in Fibronectin Matrix Assembly." In: *J Cell Biol.* (1990). URL: doi:10.1083/jcb.110.5.1813.

- [Dau85] John G Daugman. “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters.” In: *J. Opt. Soc. Am. A* 2.7 (1985).
- [Del04] Julie Delon. “Midway Image Equalization.” In: *Journal of Mathematical Imaging and Vision* 21 (2004), pp. 119–134.
- [DRG09] Julie Delon, Julien Rabin, and Yann Gousseau. “Transportation Distances on the Circle and Applications.” In: *Journal of Mathematical Imaging and Vision* (June 2009).
- [Eft+] Georgios Efthymiou et al. *The presence of alternatively spliced Extra Domains B and A of cellular fibronectin confers topographically and functionally distinct features.*
- [ELGH93] Ramila S. Patel-King Helen Rayburn Elizabeth L. George Elisabeth N. Georges-Labouesse and Richard O. Hynes. “Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin.” In: *Development* (1993).
- [ES90] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions.* Tech. rep. 1990.
- [Ext] *Extrapolation method.* https://www.mathworks.com/matlabcentral/fileexchange/4551-inpaint_nans. Accessed: 2020-28-01.
- [FS09] Jalal Fadili and Jean-Luc Starck. “Curvelets and Ridgelets.” In: *Encyclopedia of Complexity and Systems Science.* New York, NY: Springer New York, 2009, pp. 1718–1738. DOI: 10.1007/978-0-387-30440-3_{_}111. URL: http://link.springer.com/10.1007/978-0-387-30440-3_111.
- [Fer+13] Sira Ferradans, Nicolas Papadakis, Julien Rabin, Gabriel Peyré, and Jean François Aujol. “Regularized discrete optimal transport.” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 2013. ISBN: 9783642382666. DOI: 10.1007/978-3-642-38267-3_{_}36.
- [FPV14] Pasquale Foggia, Gennaro Percannella, and Mario Vento. “Graph matching and learning in pattern recognition in the last 10 years.” In: *International Journal of Pattern Recognition and Artificial Intelligence* 28.1 (2014). ISSN: 02180014. DOI: 10.1142/S0218001414500013.
- [Fri+94] KJ Friston, KJ Worsley, RSJ Frackowiak, JC Mazziotta, and Ac Evans. “Assessing the Significance of Focal Activations Using Their Spatial Extent.” In: *Human Brain Mapping* 1 (1994), pp. 210–220.
- [Fuj19] Kensuke Fujinoki. *A Note on Curvelets and Multiscale Directional Transforms.* Tech. rep. 2019. URL: <http://www.kurims.kyoto-u.ac.jp/~kyodo/kokyuroku/contents/pdf/2102-02.pdf>.
- [Gab47] D. Gabor. “Theory of communication.” In: *Journal of the Institution of Electrical Engineers - Part I:General* 94.73 (1947), pp. 58–. DOI: 10.1049/ji-1.1947.0015.
- [GJ90] M.R. Garey and D.S. Johnson. *Computers and Intractability; A guide to the theory of NP-Completeness.* Tech. rep. 1990.

- [Gen] *Genetic algorithms*. https://en.wikipedia.org/wiki/Genetic_algorithm. Accessed: 2020-07-01.
- [GR96] Steven Gold and Anand Rangarajan. "A Graduated Assignment Algorithm for Graph Matching." In: *IEEE Transactions on pattern analysis and machine learning* 18.4 (1996).
- [GR11] F. Gómez and E. Romero. "Rotation invariant texture characterization using a curvelet based descriptor." In: *Pattern Recognition Letters*. Vol. 32. 16. Dec. 2011, pp. 2178–2186. doi: 10.1016/j.patrec.2011.09.029.
- [Gop+17] Sandeep Gopal et al. "Fibronectin-guided migration of carcinoma collectives." In: *Nature Communications* 8 (Jan. 2017). issn: 20411723. doi: 10.1038/ncomms14105.
- [Gra+18] A.-I. Grapa, R. Meunier, L. Blanc-Feraud, G. Efthymiou, S. Schaub, A. Radwanska, E. Van Obberghen-Schilling, and X. Descombes. "Classification of the fibronectin variants with curvelets." In: *Proceedings - International Symposium on Biomedical Imaging*. Vol. 2018-April. 2018. isbn: 9781538636367. doi: 10.1109/ISBI.2018.8363723.
- [Gra+19] Anca-Ioana Grapa, Laure Blanc-Feraud, Ellen Van Obberghen-Schilling, and Xavier Descombes. "Optimal Transport vs Many-to-many assignment for Graph Matching." In: (2019). url: <https://hal.archives-ouvertes.fr/hal-02279634/document>.
- [Bin] *Graphical representation of binary SVM*. www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf. Accessed: 2019-11-27.
- [Ham13] William Hamilton. *Biologically Inspired Object Recognition using Gabor Filters*. Tech. rep. 2013. url: <https://pdfs.semanticscholar.org/ce3a/3ff8be6a6853355d2aa5bdb491e650b37663.pdf>.
- [Hot+15] Nathan A. Hotaling, Kapil Bharti, Haydn Kriel, and Carl G. Simon. "DiameterJ: A validated open source nanofiber diameter measurement tool." In: *Biomaterials* 61 (Aug. 2015), pp. 327–338. issn: 18785905. doi: 10.1016/j.biomaterials.2015.05.015.
- [HL02] Chih-Wei Hsu and Chih-Jen Lin. "A Comparison of Methods for Multi-class Support Vector Machines." In: *IEEE Transactions on Neural Networks* (2002).
- [Hyn09] Richard O. Hynes. "The extracellular matrix: Not just pretty fibrils." In: *Science* 326.5957 (Nov. 2009), pp. 1216–1219. issn: 00368075. doi: 10.1126/science.1176009.
- [Imm] *Immunofluorescence*. <https://en.wikipedia.org/wiki/Immunofluorescence>. Accessed: 2019-18-09.
- [IZ09] Md Monirul Islam and Dengsheng Zhang. "Rotation invariant curvelet features for texture image retrieval." In: *2009 IEEE International Conference on Multimedia and Expo* (2009).
- [Ive] *Iverson bracket*. https://en.wikipedia.org/wiki/Iverson_bracket. Accessed: 2019-12-04.

- [Jia+01] Xiaoyi Jiang, Andreas Mu, È Nger, and Horst Bunke. "On Median Graphs: Properties, Algorithms, and Applications." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2001).
- [Jul+11] Rabin Julien, Gabriel Peyré, Julie Delon, and Bernot Marc. "Wasserstein Barycenter and its Application to Texture Mixing." In: *SSVM* (2011), pp. 435–446. URL: <https://hal.archives-ouvertes.fr/hal-00476064>.
- [Kan06] L V Kantorovich. "On the translocation of masses." In: *Translated from Zapiski Nauchnykh Seminarov POMI* 133.4 (2006), pp. 11–14.
- [KM13] Trupti M Kodinariya and Prashant R Makwana. "Review on Determining of Cluster in K-means Clustering Review on determining number of Cluster in K-Means Clustering." In: *International Journal of Advance Research in Computer Science and Management Studies* 1.6 (2013). ISSN: 2321-7782. URL: <https://www.researchgate.net/publication/313554124>.
- [Kol+17a] Philip Kollmannsberger, Michael Kerschnitzki, Felix Repp, Wolfgang Wagermaier, Richard Weinkamer, and Peter Fratzl. "The small world of osteocytes: Connectomics of the lacuno-canalicular network in bone." In: *New Journal of Physics* 19.7 (July 2017). ISSN: 13672630. DOI: 10.1088/1367-2630/aa764b.
- [Kol+17b] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. "Optimal Mass Transport: Signal processing and machine-learning applications." In: *IEEE Signal Processing Magazine* 34.4 (July 2017), pp. 43–59. ISSN: 10535888. DOI: 10.1109/MSP.2017.2695801.
- [KV00] Bernhard Korte and Jens Vygen. "Network Flows." In: *Combinatorial Optimization: Theory and Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 153–184. ISBN: 978-3-662-21708-5. DOI: 10.1007/978-3-662-21708-5_8. URL: https://doi.org/10.1007/978-3-662-21708-5_8.
- [KPG02] P Kruizinga, N Petkov, and S E Grigorescu. "Comparison of texture features based on Gabor filters." In: *IEEE Transactions on Image Processing* (2002), pp. 142–147.
- [Laf+07] F Lafarge, X Descombes, J Zerubia, and S Mathieu. *Détection de feux de forêt par analyse statistique d'évènements rares à partir d'images infrarouges thermiques*. Tech. rep. 2007. URL: <http://documents.irevues.inist.fr/handle/2042/8905>.
- [Law63] Eugene L. Lawler. "The Quadratic Assignment Problem." In: *Management Science* 9.4 (July 1963), pp. 586–599. ISSN: 0025-1909. DOI: 10.1287/mnsc.9.4.586.
- [Lia+09] Chung Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. "IsoRankN: Spectral methods for global alignment of multiple protein networks." In: *Bioinformatics*. Vol. 25. 12. 2009. DOI: 10.1093/bioinformatics/btp203.
- [Lin] *Linear programming*. https://en.wikipedia.org/wiki/Linear_programming. Accessed: 2019-11-29.

- [Loi+07] Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. "A survey for the quadratic assignment problem." In: *European Journal of Operational Research* 176.2 (Jan. 2007), pp. 657–690. ISSN: 03772217. DOI: 10.1016/j.ejor.2005.09.032.
- [LWH03] Bin Luo, Richard C. Wilson, and Edwin R. Hancock. "Spectral embedding of graphs." In: *Pattern Recognition* 36.10 (2003), pp. 2213–2230. ISSN: 00313203. DOI: 10.1016/S0031-3203(03)00084-0.
- [MP11] Jianwei Ma and Gerlind Plonka. *A review of curvelets and recent applications*. Tech. rep. 2011. URL: <https://www.researchgate.net/publication/228684442>.
- [Mém07] Facundo Mémoli. "On the use of Gromov-Hausdorff Distances for Shape Comparison." In: *Eurographics Symposium on Point-Based Graphics* (2007).
- [Mém11] Facundo Mémoli. "Gromov-Wasserstein Distances and the Metric Approach to Object Matching." In: *Foundations of Computational Mathematics* 11.4 (Aug. 2011), pp. 417–487. ISSN: 16153375. DOI: 10.1007/s10208-011-9093-5.
- [Mon81] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. Tech. rep. 1781.
- [Ske] *Morphological skeletonization*. www.homepages.inf.ed.ac.uk/rbf/HIPR2/skeleton.html. Accessed: 2019-11-27.
- [MK13] Lisa D. Muiznieks and Fred W. Keeley. "Molecular assembly and mechanical properties of the extracellular matrix: A fibrous protein perspective." In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1832.7 (July 2013), pp. 866–875. ISSN: 0925-4439. DOI: 10.1016/J.BBADIS.2012.11.022. URL: <https://www.sciencedirect.com/science/article/pii/S0925443912002839>.
- [Muk+07] Lopamudra Mukherjee, Vikas Singh, Jiming Peng, Jinhui Xu, Michael J Zeitz, and Ronald Berezney. "Generalized Median Graphs: Theory and Applications *." In: *IEEE 11th International Conference on Computer Vision* (2007).
- [NRB06] Michel Neuhaus, Kaspar Riesen, and Horst Bunke. "Fast Suboptimal Algorithms for the Computation of Graph Edit Distance." In: *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (2006).
- [OS+11] Ellen van Obberghen-Schilling, Richard P. Tucker, Falk Saupe, Isabelle Gasser, Botond Cseh, and Gertraud Orend. "Fibronectin and tenascin-C: Accomplices in vascular morphogenesis during development and tumor growth." In: *International Journal of Developmental Biology* 55.4-5 (2011), pp. 511–525. ISSN: 02146282. DOI: 10.1387/ijdb.103243eo.
- [Min] *Patent Confocal Microscopy - M. Minsky*. <https://worldwide.espacenet.com/patent/search/family/024791595/publication/US3013467A?q=pn3DUS3013467>. Accessed: 2019-16-09.

- [PK97] N Petkov and P Kruizinga. “Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells.” In: *Biological Cybernetics* (1997).
- [Pet95] Nikolay Petkov. “Biologically motivated computationally intensive approaches to image pattern recognition.” In: *Future Generation Computer Systems* 11.4-5 (Aug. 1995), pp. 451–465. ISSN: 0167-739X. DOI: 10.1016/0167-739X(95)00015-K. URL: <https://www.sciencedirect.com/science/article/pii/0167739X9500015K>.
- [PC19] Gabriel Peyré and Marco Cuturi. “Computational Optimal Transport.” In: *Foundations and Trends in Machine Learning, vol. 11* (2019). URL: <http://arxiv.org/abs/1803.00567>.
- [PCST00] John C Platt, Nello Cristianini, and John Shawe-Taylor. “Large Margin DAGs for Multiclass Classification.” In: *Advances in Neural Information Processing Systems 12* (2000).
- [Poi] *Poisson Point Process*. https://en.wikipedia.org/wiki/Poisson_point_process_Spatial_Poisson_point_process. Accessed: 2019-12-03.
- [Pol+97] J-B Poline, K J Worsley, A C Evans, and K J Friston. “Combining Spatial Extent and Peak Intensity to Test for Activations in Functional Imaging.” In: *Neuroimage* (1997). URL: doi:10.1006/nimg.1996.0248.
- [Pro] *Procrustes analysis*. https://en.wikipedia.org/wiki/Procrustes_analysis. Accessed: 2019-12-09.
- [RDG11] Julien Rabin, Julie Delon, and Yann Gousseau. “Removing Artefacts from Color and Contrast Modification.” In: *IEEE Transactions on image processing* (2011).
- [Rab+14] Julien Rabin, Sira Ferradans, Nicolas Papadakis, J Rabin, Univ Caen, and S Ferradans. *Adaptive Color Transfer With Relaxed Optimal Transport*. Tech. rep. 2014. URL: <https://hal.archives-ouvertes.fr/hal-01002830>.
- [RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. “The Earth Mover’s Distance as a Metric for Image Retrieval.” In: *International Journal of Computer Vision* (2000).
- [SB09] E. A. Sander and V. H. Barocas. “Comparison of 2D fiber network orientation measurement methods.” In: *Journal of Biomedical Materials Research - Part A* 88.2 (Feb. 2009), pp. 322–331. ISSN: 15493296. DOI: 10.1002/jbm.a.31847.
- [SF83] Alberto Sanfeliu and King-Sun Fu. “A Distance Measure Between Attributed Relational Graphs for Pattern Recognition.” In: *IEEE Transactions on Systems, Man, and Cybernetics* (1983).
- [San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians-Calculus of Variations, PDEs and Modeling*. 2015. URL: <https://www.imo.universite-paris-saclay.fr/~filippo/OTAM-cvgmt.pdf>.

- [San18] Filippo Santambrogio. *A short story on optimal transport and its many applications*. 2018. URL: <https://imaginary.org/snapshot/a-short-story-on-optimal-transport-and-its-many-applications>.
- [SW] Denis Semwogerere and Eric R Weeks. *Confocal Microscopy*. DOI: 10.1081/E-EBBE-120024153. URL: www.dekker.com.
- [SDL08] Yan Shang, Yan-Hua Diao, and Chun-Ming Li. "Rotation invariant texture classification algorithm based on Curvelet transform and SVM." In: *2008 International Conference on Machine Learning and Cybernetics*. Vol. 5. 2008, pp. 3032–3036.
- [SK99] Peter Shoubridge and Miro Kraetzl. "Detection of abnormal change in dynamic networks." In: *Information, Decision and Control. Data and Information Fusion Symposium, Signal Processing and Communications Symposium and Decision and Control Symposium. Proceedings (Cat. No.99EX251)* (1999).
- [Sin] Sinkhorn Knopp algorithm. https://en.wikipedia.org/wiki/Sinkhorn's_theorem. Accessed: 2019-11-29.
- [SK67] Richard Sinkhorn and Paul Knopp. "Concerning nonnegative matrices and doubly stochastic matrices." In: *Pacific Journal of Mathematics* 21.2 (1967).
- [Sol+] Justin Solomon, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. *Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains*. URL: https://people.csail.mit.edu/jsolomon/assets/convolutional_w2.compressed.pdf.
- [SGP09] Vitomir Struc, Rok Gajšek, and Nikola Paveši Pavešić. "Principal Gabor Filters for Face Recognition." In: *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems* (2009).
- [Sze+14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. *Going Deeper with Convolutions*. 2014. arXiv: 1409.4842 [cs.CV].
- [Ta-94] Kashyap Chong-Nam Chu Ta-Chih Lee Rangasami L. "Building skeleton models via 3-D medial surface/axis thinning algorithms." In: *Computer Vision, Graphics, and Image Processing*, 56(6):462–478 (1994).
- [TF79] Wen-hsiang Tsai and King-Sun Fu. "Error-Correcting Isomorphisms of Attributed Relational Graphs for Pattern Analysis." In: *IEEE Transactions on Systems, Man, and Cybernetics* (1979).
- [Tur86] M R Turner. "Texture Discrimination by Gabor Functions." In: *Biol. Cybern* (1986).
- [Mic] *Types of microscopes used in biology*. <https://sciencing.com/advantages-studying-cells-under-light-microscope-9058.html>. Accessed: 2019-16-09.
- [Ull76] J R Ullmann. *An Algorithm for Subgraph Isomorphism*. 1976. URL: https://www.cs.bgu.ac.il/~dinitz/Course/SS-12/Ullman_Algorithm.pdf.

- [Ume88] Shinji Umeyama. "An Eigendecomposition Approach to Weighted Graph Matching Problems." In: *IEEE Transactions on pattern analysis and machine intelligence* 10 (1988).
- [V.P69] V.P.Nosko. "Local structure of Gaussian random fields in the neighborhood of high-level shines." In: *Dokl. Akad. Nauk SSSR*, 189:4 714–717 (1969).
- [Vap95] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag, 1995. ISBN: 0-387-94559-8.
- [Vay+18] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. *Optimal Transport for structured data with application on graphs*. May 2018. URL: <http://arxiv.org/abs/1805.09114>.
- [Ven15] Mario Vento. "A long trip in the charming world of graphs for Pattern Recognition." In: *Pattern Recognition* (2015). ISSN: 00313203. DOI: 10.1016/j.patcog.2014.01.002.
- [Ven+10] Elisa Ventura, Francesca Sassi, Arianna Parodi, Enrica Balza, Laura Borsi, Patrizia Castellani, Barbara Carnemolla, and Luciano Zardi. "Alternative Splicing of the Angiogenesis Associated Extra-Domain B of Fibronectin Regulates the Accessibility of the B-C Loop of the Type III Repeat 8." In: *PLOS ONE* (2010). DOI: 10.1371/journal.pone.0009145.
- [Vil03] C Villani. *Topics in Optimal Transportation*. Tech. rep. 2003.
- [Vil08] C Villani. *Optimal transport: old and new*. Tech. rep. 2008.
- [Yan+] Junchi Yan, Yu Tian, Shanghai Jiao, Hongyuan Zha, Xiaokang Yang, Ya Zhang, and Stephen M Chu. *Joint optimization for consistent multiple graph matching*. Tech. rep.
- [Yan+16] Junchi Yan, Xu-Cheng Yin, Weiyao Lin, Cheng Deng, Hongyuan Zha, and Xiaokang Yang. "A Short Survey of Recent Advances in Graph Matching." In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (2016). DOI: 10.1145/2911996.2912035.
- [YS07] Qingwu Yang and Sing Hoi Sze. "Path matching and graph matching in biological networks." In: *Journal of Computational Biology* 14.1 (Jan. 2007), pp. 56–67. ISSN: 10665277. DOI: 10.1089/cmb.2006.0076.
- [ZBV10] Mikhail Zaslavskiy, Francis Bach, and Jean Philippe Vert. "Many-to-many graph matching: A continuous relaxation approach." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2010. ISBN: 3642159389. DOI: 10.1007/978-3-642-15939-8_{_}33.
- [Zas+10] Mikhail Zaslavskiy, Directeurs De, Francis Bach, Jean-Philippe Vert, Colin De, and L A Higuera. *Graph matching and its application in computer vision and bioinformatics Jury*. Tech. rep. 2010.
- [ZDLT] Feng Zhou and Fernando De La Torre. *Factorized Graph Matching*. Tech. rep. URL: <http://humansensing.cs.cmu.edu/fgm>.
- [al99] Sohei Satoi et. al. "Different responses to surgical stress between extra domani A+ and plasma fibronectin." In: *Clinical and Experimental Pharmacology and Physiology* (1999).