Multi-view learning and deep learning for heterogeneous biological data to maintain oral health

Imangaliyev, Sultan

Citation for published version (APA):
Imangaliyev, S. K. (2016). Multi-view learning and deep learning for heterogeneous biological data to maintain oral health

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Appendix A

Summary

This is a brief summary in English language of the thesis entitled "Multi-view Learning and Deep Learning for Heterogeneous Biological Data to Maintain Oral Health" written by Sultan K. Imangaliyev.

In Chapter 1, a general introduction to data-driven medical research and computational challenges in multi-omics research are given. Then motivation for multi-omics research in oral health and a brief introduction to machine learning techniques are provided. The chapter is concluded by a brief description of the datasets used in the thesis and the outline of the thesis.

In Chapter 2, a novel multi-output co-regularized learning algorithm was proposed and its application on a real world problem in oral health domain was demonstrated. More precisely, we addressed the task of prediction of Porphyromonas species which are implicated in certain forms of dental diseases. In our semi-supervised modeling approach, we used the fact that interactions among different microbial species in various niches of human body are known to exist. Such biological interactions suggest possibility to use additional information in a form of "unlabeled data" from different niches to improve prediction of Porphyromonas. We have made use of a dataset available from the Human Microbiome Project. We derived a training set in which we used the abundance of Porphyromonas as output data utilized for multi-view semi-supervised regression problem. The algorithm can be extended and used in future simulations for identification of bacterial species related to the Porphyromonas abundance as predictive and/or prognostic biomarkers. Such biomarkers may allow identification of novel intervention strategies to lower abundance of Porphyromonas and diagnose or prevent dental disease which has an important value for practicing dental clinicians.

The results described in Chapter 3 suggest that microbial interactions on personalized network level differ in several ways. Co-regularized spectral clustering showed two groups of individuals with a different topology of their microbial interaction network; the microbial network suggested that niche-wise interactions differed between them. According to their oral microbiota, healthy individuals belong to one of two different microbial clusters. We showed that application of the unsupervised machine learning method to complex -omics data can lead to a personalized approach in oral health. Based on our results, we demonstrated that co-regularized spectral clustering applied to oral metagenomics data could discover personalized microbial networks. Such networks provide important insights into microbial composition and into understanding of complex microbial interactions, which can be
In Chapter 4, an Unsupervised Multi-View Feature Selection (UMVFS) algorithm was proposed; a novel technique that simultaneously performs clustering and identifies the most relevant subset of features across different views. Unlike existing unsupervised feature selection methods, our algorithm simultaneously uses information contained in different views and leads to more accurate results compared to the ones obtained via single-view techniques. Our algorithm is naturally suitable for application in the computational biology domain where datasets with multi-view representations are frequently available. We have demonstrated the efficacy of the UMVFS algorithm on the Human Microbiome Project study as well as on synthetic datasets. In our simulations the proposed algorithm successfully identified a relevant subset of features and led to increased clustering performance.

In Chapter 5, the efficacy of the UMVFS algorithm on the NCI-60 multi-view cancer dataset as well as on synthetic multi-view datasets was demonstrated. In principle, UMVFS can be applied in any biological problem where (a) labels are unknown, (b) the number of useful features is unknown and (c) phenomenon could be reflected with multiple complementary representations. Therefore, we used NCI-60 cell panel data as a proof of principle study to demonstrate the power of UMVFS application in computational cancer biology. Features extracted by UMVFS can help specialists involved in biomedical field to perform more efficient diagnosis based on discovered biomarkers and provide guidance in patient stratification in a more realistic unsupervised way. Moreover, the proposed model can be used by other bioinformatics scientists who may be interested in adapting UMVFS to different scenarios, e.g. parallelization, implementation as a part of standard bioinformatics toolboxes, publishing on-line as a web-based tool etc.

In Chapter 6, the ecobiological relations between the salivary microbiome, salivary metabolome and salivary biochemistry at health were quantified. Unstimulated overnight fasting saliva of 268 healthy adults was collected. The heterogeneity of the microbiome, metabolome, and biochemistry as well as the mutual interrelation between these datasets and host-related biological parameters were assessed. Individuals clustered into five microbiome and four metabolome-based clusters that significantly related to biochemical parameters of saliva. Low salivary pH and high lysozyme activity related to high proportions of streptococcal phylotypes and increased membrane lipid degradation products. The relation with specific host parameters suggests presence of highly specialized ecotypes of a healthy oral ecosystem. An over-specialization either toward proteolytic or saccharolytic ecotype may indicate a shift toward a dysbiotic state. This study shows that a healthy oral ecosystem is based on concerted interactions between its heterogeneous and multifactorial components and should be studied as such.

In Chapter 7, a Convolutional Neural Network (CNN) model was applied to automatically classify red fluorescent dental plaque images. A comparison with several other state-of-the-art classification methods clearly showed the advantage of the CNN model in achieving a higher prediction performance. Such a result was possible because the CNN model explicitly learns invariant feature representations from raw pixel intensity values. Such representations were visualized using the convolutional filters learnt during the training of the model. We expect that deep learning of red fluorescent dental plaque images can help dental practitioners to perform efficient fluorescent plaque assessments and provide guidance in patient diagnosis. The proposed model can be implemented as user-friendly software for dental
practitioners in daily practice and thus contribute to the improvement of patients’ oral health.

In Chapter 8, the associations between red fluorescent plaque and microbial data were evaluated. Stability selection with the elastic net regression was applied to oral microbial data which resulted in selecting smaller subsets of the most important features per each oral niche. Results suggested that it is possible to predict an individual’s plaque accumulation level 14 days before the actual plaque accumulation. These predictions can be achieved by using a small subset of original features. Simulations applied to data collected from different niches provided distinct microbial profiles per each niche. Usually, in dental care the patient population is treated as if patients are all at risk for developing disease, hence patients are asked to brush their teeth twice a day, to limit carbohydrate consumption frequency etc. This is caused due to the lack of good prediction models. Current risk prediction models rely on symptoms of current or previous disease, e.g. caries, gingivitis, periodontitis etc. Elastic net model used in this study predicts the response to a disease challenge based on biomarkers in the healthy situation.

The thesis is concluded by Chapter 9, which contains overview of the machine learning techniques applied in the thesis, complexity of the problems in oral health domain and a general discussion about future perspectives of machine learning applications in life sciences.
Appendix B

Samenvatting

Dit is een korte samenvatting in de Nederlandse taal van het proefschrift getiteld "Multi-view Learning and Deep Learning for Heterogeneous Biological Data to Maintain Oral Health", geschreven door Sultan K. Imangaliyev. Wanneer betekenend niet duidelijk wordt, moet de oorspronkelijke Engels versie als een referentie gebruikt worden.

In hoofdstuk 1 wordt een algemene inleiding tot data-driven medisch onderzoek en computationele uitdagingen in multi-omics onderzoek gegeven. Dan motivatie voor multi-omics onderzoek in mondgezondheid en een korte introductie van machine learning technieken zijn voorzien. Het hoofdstuk wordt afgesloten met een korte beschrijving van de datasets die in het proefschrift gebruikt worden en de schets van het proefschrift.

In hoofdstuk 2 wordt een nieuwe Multi-Output Co-regularized Learning algoritme genoemt en aangetoond dat het toe is te passen op echte wereld problemen. Preciezer gesproken onderzochten we de voorspelling van *Porphyromonas* soorten die betrokken zijn bij de bepaalde vormen van tandziekten. In het semi-supervised modelleringsbenadering, gebruikten we het feit dat interacties tussen verschillende microbiële soorten in verschillende niches van het menselijk lichaam bekend zijn. Dergelijke biologische interacties suggereren een mogelijkheid om extra informatie te gebruiken in de vorm van "unlabeled data" uit verschillende niches om voorspelling van *Porphyromonas* te verbeteren. We hebben gebruik gemaakt van een dataset die beschikbaar is vanuit het Human Microbiome Project. We hebben een training set aangemaakt waarin de overvloed een abundantie van *Porphyromonas* als output data wordt gebruikt in een multi-view semi-supervised regression zetting. Het algoritme kan worden uitgebreid in de volgende simulaties voor de identificatie van predictieve en/of prognostische biomarkers in verband met de *Porphyromonas* overvloed abundantie. Dergelijke biomarkers kunnen er voor zorgen dat er nieuwe interventiestrategien ontwikkeld kunnen worden om de hoeveelheid *Porphyromonas* te verlagen om vervolgens tandheelkundige ziekte te diagnosticeren of voorkomen van tandheelkundige ziekte die een belangrijke waarde voor het beoefenen van tandheelkundige clinici heeft.

De in hoofdstuk 3 beschreven resultaten suggereren dat microbiële interacties op persoonlijk netwerk niveau op diverse manieren verschillen. Co-regularized spectral clustering toonde twee groepen van personen met een verschillende topologie van hun microbiële interactie netwerk; het microbiële netwerk gesuggereerd dat niche-wise interacties verschilden tussen hen. Volgens hun orale microbiota behoren gezonde individuen tot een van twee verschillende microbiële clusters. We toonden aan dat de toepassing van de unsupervised
machine learning modellen om complexe -omics gegevens in een gepersonaliseerde manier in mondgezondheid gebruikt kan worden. Op basis van onze resultaten hebben we aange- toond dat spectral clustering (wanneer toegepast op orale metagenomica gegevens) gepersonaliseerde microbiële netwerken kan ontdekken. Dergelijke netwerken geven belangrijke inzichten in microbiële samenstelling en in begrip van complexe microbiële interacties, die met succes in een klinische praktijk kan worden toegepast.

In hoofdstuk 4 hebben we een zogenaamde Unsupervised Multi-View Feature Selection (UMVFS) algoritme gentroduceerd. UMVFS is een nieuwe techniek die tegelijkertijd clustert en de meest relevante subset van functies in verschillende views identificeert. In tegenstelling tot andere unsupervised feature selectie methoden, kan ons algoritme gelijktijdig informatie vanuit verschillende views gebruiken en leidt dit tot nauwkeurigere resultaten in vergelijking met degenen die via single-view technieken verkregen worden. Ons algoritme is vooral geschikt voor toepassing in de computatiele biologie domein waarin datasets in een vorm van multi-view representaties vaak beschikbaar zijn. We hebben de werkzaamheid van het UMVFS algoritme op de Human Microbiome Project dataset onderzocht, alsmede zijn doeltreffendheid op synthetische datasets gedemonstreerd. In onze simulaties ontdekt het gentroduceerde algoritme succesvol een relevante subset van features en het geleid tot betere clustering prestatie.

In hoofdstuk 5 worden de werkzaamheid van het UMVFS algoritme op de NCI-60 multi-view kanker dataset en op synthetische multi-view datasets aangetoond. In principe kan UMVFS worden toegepast in een biologisch probleem indien (a) labels niet bekend zijn, (b) het aantal nuttige features onbekend is en (c) het fenomeen door meerdere complementaire representaties kan worden voorgelegd. Daarom gebruikten we NCI-60 data cel panel als een proof of principle studie om de kracht van UMVFS toepassing in de computatiële kanker biologie aan te tonen. Features gevonden door UMVFS kunnen biomedisch specialisten helpen om een efficiëntere diagnose op basis van ontdekte biomarkers uit te voeren enende patiënt stratificatie op een meer realistische ‘unsupervised’ manier te begeleiden. Bovendien kan het voorgestelde model worden gebruikt door andere genteresseerde bioinformatie wetenschappers die in het aanpassen van UMVFS aan verschillende scenario’s, denk bijvoorbeeld aan de parallelisatie, implementatie als een onderdeel van de standaard bioinformatica gereedschapskisten, publishing on-line als een web-based tool, etc.

In hoofdstuk 6 worden de eco biologische betrekkingen tussen de speekselklieren microbiome, speeksel metaboloom en speeks biochemie aan de mondgezondheid gekwantificeerd. Gestimuleerde speeksel van 268 gezonde volwassenen werd verzameld nadat deze een nacht gevast hadden. De heterogeniteit van het microbioom, metaboloom en biochemie en de onderlinge relatie tussen deze datasets en gerelateerde aan de gastheer gerelateerde biologische parameters werden beoordeeld. Individuen worden in vijf microbioom en viermetaboloom gebaseerde clusters geclusterd die significant gerelateerd zijn aan de biochemische parameters in het speeksel. We hebben gevonden dat speeksel met een laag pH en een hoog lysozym activiteit een hoog streptokokken phylotypes bevat en een verhoogde concentratie membraan lipide-afbraakproducten aanwezig zijn. De relatie met specifieke gastheer parameters suggereert de aanwezigheid van zeer gespecialiseerde ecotypes van een gezond oraal ecosysteem. Een over-specialisatie, hetzij in de richting van proteolytische of saccharolytische ecotype kan een verschuiving in de richting van een dysbiotic toestand aangeven. Deze studie toont aan dat een gezonde oraal ecosysteem gebaseerd is op on-
derlinge interacties tussen hun heterogene en multifactorile onderdelen en moeten dus als zodanig bestudeerd worden.

In hoofdstuk 7 wordt een Convolutional Neural Network (CNN) model toegepast om rode fluorescerende tandplak beelden automatisch te classificeren. Een vergelijking met een aantal andere state-of-the-art classificatiemethoden toonde duidelijk het voordeel van het CNN model, een verbetering in de voorspelling prestaties. Een dergelijk resultaat was mogelijk omdat het CNN model invariant feature representations van rauwe pixelintensiteitswaarden expliciet heeft geleerd. Deze feature representations werden gevisualiseerd met behulp van de convolutionele filters die tijdens de training van het model worden geleerd. We verwachten dat het toepassen van Deep Learning op de rode fluorescerende tandplak beelden tandartsen kan helpen om rode fluorescerende plaque evaluaties efficiënt uit te voeren en om de patiënt te begeleiden bij de diagnose. Het gentroduceerde model kan worden gemplementeerd als gebruiksvriendelijke software die door tandartsen in de dagelijkse praktijk kan worden gebruikt en zo bijdraagt aan de verbetering van de orale gezondheid van de patiënten.

In hoofdstuk 8 werden de associaties tussen de rode fluorescerende plaque en microbiële gegevens gevalueerd. Stability selection en werd er Elastic Net regression toegepast op orale microbiële data waardoor een kleinere selectie aan subsets van de belangrijkste features per elke orale niche overbleef. Resultaten suggereerden dat het mogelijk is om een plaque level van een individu 14 dagen van de feitelijke plaqueaccumulatie te voorspellen. Deze voorspellingen kunnen bereikt worden met slechts een klein deel van originele features. Simulaties toegepast op de gegevens die zijn verzameld uit verschillende niches hebben onderscheidende microbiële profielen per niche. In tandverzorging wordt de patiëntenpopulatie meestal behandeld alsof alle patiënten een verhoogd risico op ziekte hebben, en wordt patiënten om die reden gevraagd om hun tanden twee keer per dag te poetsen of om hun koolhydraten consumptiefrequentie te beperken etc. Dit wordt veroorzaakt door het ontbreken van goede voorspelling modellen. Huidige risicopredictiemodellen zijn afhankelijk van symptomen van actuele of eerdere ziekte, zoals caris, gingivitis, periodontitis etc. Het model dat in deze studie gebruikt wordt, voorspelt een ongezonde toestand op basis van de biomarkers in gezonde toestand.

Het proefschrift wordt afgesloten met hoofdstuk 9. Dit hoofdstuk bevat een algemeen overzicht van de machine learning technieken die in het proefschrift toegepast worden. Dit hoofdstuk legt ook de complexiteit van de problemen in mondgezondheid domein uit en het bevat een algemene discussie over de toekomstperspectieven van machine learning toepassingen in de levenswetenschappen.
Appendix C

Publications list

C.1 Publications included in the thesis


C.2 Other publications

1. A Prodan, HS Brand, AJM Ligtenberg, SK Imangaliyev, E Levin, GA van der Weijden, W Crielaard, BJF Keijser, ECI Veerman. Interindividual variation, correlations, and


Appendix D

Co-authorship list

D.1 Chapter 2

SKI, BJFK and EL conceived the idea of the paper, designed the experimental setup and interpreted the results. EL derived the modelling methodology. SKI preprocessed the datasets and performed the simulations. SKI, EL, WC and BJFK drafted the manuscript. All authors have read and approved the final manuscript.

D.2 Chapter 3

SKI, BJFK and EL conceived the idea of the paper, designed the experimental setup and interpreted the results. EL derived the modelling methodology. SKI preprocessed the datasets and performed the simulations. SKI, EL, WC and BJFK drafted the manuscript. All authors have read and approved the final manuscript.

D.3 Chapter 4

SKI and EL conceived the idea of the paper, derived the modelling methodology, designed the experimental setup and interpreted the results. SKI preprocessed the datasets and performed the simulations. SKI, EL and BJFK drafted the manuscript. All authors have read and approved the final manuscript.
D.4 Chapter 5


SKI and EL conceived the idea of the paper, derived the modelling methodology, designed the experimental setup and interpreted the results. SKI preprocessed the datasets and performed the simulations. SKI, EL, BJFK, and WC drafted the manuscript. All authors have read and approved the final manuscript.

D.5 Chapter 6


EZ, NLH-H, DES, EAN, BPK, HSB, ECIV, MK, BGL, GAvdW, WC, BJFK developed study idea and protocol. NLH-H, DES, EAN, GAvdW managed the clinical part of the study. MJB, HSB, BJFK managed the laboratory analysis. EZ, BWB, SKI, BJFK managed the data analysis. AP, NLH-H, EAN, MDL, MMJ provided the sample collection and the clinical measurements. AP, JK, MJB EAN preformed the laboratory analysis. JK sequenced the samples. EZ, BWB, AP, SKI analyzed data. SKI provided the results of the spectral clustering analysis. FLPW developed and managed the data warehouse. SM designed managed food frequency questionnaires data analysis (non included in this thesis). EZ, BWB, AP, MJTdM, SKI, EAN, MDL, MMJ, MMF-G, BPK, HSB, ECIV, MK, BGL, WC, BJFK discussed and integrated the results. EZ, BWB, AP, MJTdM, SM, BJFK drafting the manuscript. All authors have read and approved the final manuscript.

D.6 Chapter 7


SKI and EL conceived the idea of the paper, designed the experimental setup and chose the modelling methodology. SKI performed the simulations. MHvdV and CMCV collected the dataset, calculated corresponding labels, and provided domain-specific knowledge. SKI, EL, MHvdV and CMCV interpreted the results and drafted the manuscript. WC, BJFK and BGL provided valuable feedback on the final manuscript’s content and they are also principal co-investigators of the clinical study. All authors have read and approved the final manuscript. MHvdV is co-inventor on several patents related to QLF. The authors declare that otherwise, there are no other conflicts of interest pertaining to the data presented in this chapter.
D.7 Chapter 8


SKI, EL, BJFK and WC conceived the idea of the paper. SKI and EL designed the experimental setup and chose the modelling methodology. SKI performed the simulations. MHvdV and CMCV collected the red fluorescent plaque dataset, calculated corresponding labels, and provided domain-specific knowledge. JK and BWB collected, processed the metagenomics dataset and provided domain-specific knowledge. SKI, EL, MHvdV, CMCV, JK, BWB, EZ interpreted the results and drafted the manuscript. WC and BJFK provided valuable feedback on the final manuscript’s content and they are also principal co-investigators of the clinical study. All authors have read and approved the final manuscript. MHvdV is co-inventor on several patents related to QLF. The authors declare that otherwise, there are no other conflicts of interest pertaining to the data presented in this chapter.
Appendix E

Financial justification of the research

The study presented in this thesis was funded by Top Institute Food and Nutrition (project number OH001), a public-private partnership in pre-competitive research on food and nutrition. The collection of fluorescent images used in chapters 7 and 8 was supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO) and partly funded by the Ministry of Economic Affairs (project number 10948). Organizations supporting this project had no role in study design, data collection and analysis, decision to publish, or preparation of the thesis.
Appendix F

Word of thanks

If I have seen further, it is only by standing on the shoulders of giants.

Sir Isaac Newton (1643 - 1727)

My journey to the PhD degree is almost finished and now I realize that it was a great journey. It was full of challenges and I learnt how to overcome them all to reach my destination on time. Writing a PhD thesis was a new process for me accompanied with time pressure and I had to learn many new skills to become an independent scientist. Fortunately, I was not alone and I met many great people who supported me on my way to the point where I am now. I would like to express my thanks to those people who influenced me and helped me to reach this point in my life.

First and foremost I wish to thank my promotor prof. dr. Wim Crielaard, a head of the Preventive Dentistry department at ACTA. He has been supportive since the first days I began working in this project. Ever since, Wim has supported me academically and emotionally through the rough road to finish this thesis. He has been actively interested in my work and has always been available to advise me. I am very grateful for his patience, motivation, enthusiasm, and immense knowledge in his field that, taken together, make him a great mentor. I would also like to express many thanks to my co-promotor prof. dr. Bart J. F. Keijser, who has been a tremendous mentor for me. I would like to thank him for encouraging my research and for allowing me to grow as an independent research scientist. His advice on both research as well as on my career have been priceless.

During my PhD fellowship years I had a privilege to work with my daily supervisor dr. Evgeni Levin. This dissertation would not be possible without his guidance and help. From the start Evgeni directed my research by sharing his perspectives on variety of machine learning problems that interested me a lot. I would like to thank Evgeni for his support throughout my research and thesis-writing period as well as for his sound advice, teaching, and great company. He has always had a good advice for me, would it be scientific or practical question. I would like to express my special appreciation to Evgeni’s personality as a mentor, because even during the most difficult times, he gave me the moral support and the freedom I needed to move on.

dr. D. Bogaert. I want to thank them for letting my defense be an enjoyable moment, for reviewing this dissertation, for taking part in my promotion and for their sharp comments and suggestions.

This thesis represents a milestone in more than four years of work at Department of Microbiology and Systems Biology (MSB) in TNO Earth, Life and Social Sciences as well as my PhD programme at ACTA Graduate School of Dentistry (AGSD). Since my first day as a PhD student I have felt at home at both TNO and ACTA. I will always be grateful to dr. Roy Montijn, dr. Jildau Bouwman and Geeske Reijer-Vissers for a chance to learn from and to collaborate with many researchers at MSB, as well as to AGSD dean dr. Martijn van Steenbergen for opportunities to learn more about dentistry.

The list of people who indirectly contributed to this thesis would not be complete without mentioning professors who supervised my BSc and MSc studies, prof. dr. Bulat Khissarov in Almaty and, later on, prof. dr. Ton Backx and dr. Jogchem van den Berg in Eindhoven. They have influenced me with their ideas, working style, and personal charisma.

This project would have never been possible without the support and guidance of various people involved in TIFN OH001 project. I praise the enormous amount of help and teaching from all of them throughout these years. I would like to thank those who helped to run this project: prof. dr. Enno Veerman, dr. Bastiaan Krom, dr. Henk Brand, dr. Elena Nicu, dr. Martijn Rosema, dr. Marja Laine, prof. dr. Michiel Kleerebezem. I also recognize the help from Margreet Heerikhuisen and Rianne Hermus for their support in managing formal requirements of TIFN regulations.

My particular gratitude goes to the administrative staff of MSB: Hennie Overeem, Petra van der Wielen-Albers, Karin Kempenaers and others. Travelling arrangements, post formalities, organizing various events, etc. would have been impossible to accomplish without their help. There is their large contribution for making MSB such a great place for conducting research.

For some chapters of this thesis, I was fortunate to collaborate with different inspiring people. All these collaborations have resulted in joint publications and new ideas. I would like to thank dr. Egija Zaura and dr. Bernd Brandt for their patience, critical eye and willingness to bring closer our fields of work. I am grateful to dr. Monique van der Veen for inspiring discussions and for giving me insight on many potential examples of machine learning applications in dentistry.

A good support system is important for surviving and staying sane during PhD fellowship time. I was lucky to be a part of the one we used to call “TIFN get together team” which consists of Marleen Janus, Andrei Prodan, Marcela Fernandez-Gutierrez, and myself. These three friends formed the core of my research team in this project. We’ve all been there for one another and have taught each other many tools and issues of oral biology. I know that I could always ask them for advice and opinions on biology related issues. Marleen is a wonderful and generous friend who has been through many challenges and I admire her positive outlook and her ability to smile despite the situation. I would also like to thank Andrei and Marclea who have been always prolific with new, cool ideas during our work on joint publications.

My deep appreciation is dedicated to my closest TIFN colleagues: Jolanda Kool and Ferry Jagers. Their excellent work during data collection has made an invaluable contribution towards my PhD. I am grateful to them for their friendship and the humour they extended to me during my time in the TNO and for always making me feel so welcome for any dis-
cussions and questions. I am also especially grateful to Jordy, Carlijn, Heleen, Tom, Paula, and Tim who shared my data analysis hobby with me. These members of the MSB bioinformatics group have contributed immensely to my personal and professional time at TNO, because they have been a source of friendly jokes as well as good advice and collaboration.

I express my gratitude to the colleagues from MSB department at TNO Zeist: Paulo, Martien, Volmar, Koen, Victor, Eddy, Anita, Daniel, Michel, Jos, Wim, Hakim. I thank them for our discussions, joint work and for having a good time in TNO. I also want to thank present and past members of the Preventive Dentistry department at ACTA: Rob, Mark, Marjoleine, Dongmei, Michel, Alexa, Nina for giving me an introduction into the department and for sharing their knowledge. I also thank my colleagues from the Dutch DNA team: Peter, Karin, Machtelt, Trifa, Lars, Marian, Vivi, Art, Cornelis, Alexander for the interesting debates during coffee time and lunch-breaks. People at TNO, DDNA and ACTA are genuinely nice and I’m glad to have interacted with all of them.

I thank the anonymous peers, referees of the peer-reviewed journals whose critical and mostly constructive remarks have improved some manuscripts. I would also like to thank the industrial partners: Wrigley, Cargill, GlaxoSmithKline, and Philips for the financial support of the project, and Wageningen University and Research Centre for the academic collaboration within TIFN OH001 project. I would like to thank the contributors of GitHub code repository and the members of the StackOverflow and CrossValidated communities for their invaluable assistance in saving my time during many of the late-night Python coding sessions. I would like to thank Chad Meserve and his project www.dayoftherobot.com for permission to include the copyrighted images of the dentist robots to the thesis cover.

I am greatly indebted to Patrick, Dulce, Tim, Paula, Abeer, Wouter, Ileana, and Marieke for proofreading prologue and epilogue chapters of this thesis, because they helped me to make my long chapters much shorter with their sharp comments. I also thank Ferry for his help in translating the English summary to the Dutch language. For their day by day support, enjoyable moments and for helping me understand a new kind of normality, I thank my PhD fellows from ACTA: Ali, Charifa, Jessica, Mercedes, Alejandra, Heinze, Ying, Sabrina, Ana, Hessaam, Francis, Dongyun for all the good times we had and their unconditional support.

I owe my sincere gratitude to Preventive Dentistry fellow PhD students Ilona, Marleen and Suzette for their invaluable advice and for being so supportive of my work. I am also very grateful to Catherine who helped me in going through various formal stages of my PhD thesis preparation. Moreover, I am very grateful to Patrick, Abeer, Paula, Marisol and Marieke for helping me enormously during the last years of my PhD to find the solutions on problems when I often got stuck.

I have been fortunate to come across many cheerful trainees at TNO, without whom life would be bleak. Special thanks go to Anja, Kimberley, Corjanne, Kevin, Sander, Pol, Yvonne, Remon, Noeska, Noortje, Jos, Laura, Elodie, Fion, Dorine, Leo, Amber, Elisa for their ever-present support. I also wish to thank Anna for the enjoyable dinners and sushi workshops in Zeist. Also as a part of the awesome TNO trainees community, I thank Jade for drinks and parties in Leiden and Utrecht. Most of all, I would like to thank Leonie for the movies, dinners, concerts and opera performances we enjoyed together. Leonie distracted me completely and joyously, because roller skating to De Uithof was sometimes exactly what I needed after a day spent fighting with data. I want to thank her parents for their hospitality in Osnabruck and her friends for my awesome vacation in Berlin.
My time at TNO has been enjoyable in large part due to the many friends and groups that became a part of my life. I am grateful for the time I spent with fellow PhD students Marisol, Carlota, Dina, Ilse, Lenka, Juri, George, Mariel for dinners, borrels, coffee-breaks, parties and for many other joyful events and memories. I will always be grateful to my office neighbours and my friends Yu and Dulce for their help and valuable suggestions during ups and downs of my journey.

I am happy to especially thank my friends and paranymphs, Abeer and Patrick. Their help has been valuable all these years. I owe my gratitude to them for the mental support, for the open doors, for the conversations around a coffee corner, for the tailored tips in hard times, for the comprehension, for the company and friendship.

For my friends, those people that I can call at any time, I cannot thank them enough for the brilliant moments we spent together. In each place where, be that in Utrecht, Nijmegen, Rotterdam, Delft, Eindhoven or elsewhere, friends have been always there for me. I thank Ileana, Coen, Tanya, Katya for their friendship. Special thanks to Emilia and Andrei for their hospitality and helping to find my way in my career and my life as a foreigner in the Netherlands. Madina deserves a special mention, because she is the only Kazakh friend I have now and I thank her very much for helping me to keep in touch with my native language and culture.

Despite the large distance, the most basic source of my life energy, my family, still resides in Kazakhstan. I have an amazing family, unique in many ways, and the stereotype of a perfect family in many others. Their support has been unconditional all these years; they have given up many things for me to be here in the Netherlands; they have cherished with me every great moment and supported me whenever I needed it. Many thanks to my father Kairzhan and my mother Svetlana for raising me to who I am today and to my brother Islam for always being by my side.

I would also like to express my gratitude to my grandmother Zulpat and my grandfather Khassim (may he rest in peace) for the unconditional love and endless support. Their openness and perspective were essential to shape my character. My grandfather raised me with a love for wisdom and supported me in all my pursuits. I am grateful to him for being an inspiration and an example of how to do the things right. I hope he would have been proud of me if he could see me now.
Appendix G

Curriculum vitae

Sultan K. Imangaliyev was born on 1986 in Gurev, the Union of Soviet Socialist Republics (present day Atyrau, the Republic of Kazakhstan)

2008-2010  Database software developer in the private financial sector, Almaty, Kazakhstan
2010-2012  Recipient of the Amandus H. Lundqvist Scholarship sponsored by the Eindhoven University of Technology, Eindhoven, the Netherlands
2011-2012  Technical Trainee, Cargill BV, Starches and Sweeteners Europe, Bergen op Zoom, the Netherlands
2012-2016  PhD candidate in Computational Systems Biology, University of Amsterdam and Free University Amsterdam, Academic Centre for Dentistry Amsterdam, Amsterdam, the Netherlands. Thesis title: "Multi-view Learning and Deep Learning for Heterogeneous Biological Data to Maintain Oral Health"
2012-2016  Guest Junior Researcher, TNO Earth, Life and Social Sciences, Zeist, the Netherlands