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.
Chapter 1

Prologue

You know nothing, Jon Snow.

A Song of Ice and Fire

GEORGE R. R. MARTIN

1.1 Data-driven research in life sciences

1.1.1 A historical introduction to data-driven research in life sciences

The history of data analysis is as old as the history of science itself. Scientists make observations, collect them into datasets and generate new knowledge by analysing them. For example, John Snow, the father of the modern epidemiology, studied the Broad Street cholera outbreak case in London in 1854 and he suggested that the source of the outbreak was the contaminated water from the public water pump. Snow used a dot map to demonstrate how cases of cholera were centred around the pump and he used data analysis to illustrate the connection between the quality of the water from the pump and the numbers of cholera cases. It turned out that the Broad Street pump had been dug next to an old cesspit, which had begun to leak. In the cesspit the diapers of babies, who had contracted cholera from another source, had been washed [218, 28]. Although the basic principles of generating knowledge by analysing available data have not changed since then, the volume of datasets changed dramatically. For example, sequencing of one entire human genome was achieved for the first time by The Human Genome Project in 2003, resulting in identification of more than 1.4 million single nucleotide polymorphisms [113]. Storing and analysing such a large amount of biological data is a great challenge and this is true for many branches of modern life sciences due to progress made in molecular biology.

The discovery of the molecular structure of DNA by James Watson and Francis Crick in 1953 revolutionized biology. Francis Crick formulated the central dogma of molecular biology [47] as "the detailed residue-by-residue transfer of sequential genetic information cannot be transferred back from protein to either protein or nucleic acid." Since then, molecular biology data became a useful source of discovery for new biological hypotheses, because of a technological progress made in -omics technologies, such as metagenomics, proteomics, transcriptomics, metabolomics etc. Multi-omics research originates from the idea that every
process within a living organism can be measured and quantified. These processes can be split into several -omics layers and there is a flow of genetic information from one layer to another one, which keeps the organism in a state of homeostasis with the environment. Each of the -omics technologies measures a part of the genetic information flow allowing to understand biological processes in a very detailed way. For example, studies which used metagenomics data revealed new discoveries about microbiota of the gut [149, 223, 57, 147], airways [209, 172, 56, 104], oral cavity [16, 48, 18], or urogenital tract [21]. Transcriptomics data can help to find cancer biomarkers [166], peptidomics data analysis revealed oral health biomarkers [158], and studies analysing metabolomics data allow to discover biomarkers of pulmonary injury and infection [22].

1.1.2 Data-driven research and hypothesis-driven research

Wide implementation of -omics technologies into daily practice has split life sciences to pre-omics and post-omics era. An illustrative example of such a split can be found in the history of genomics studies. Pre-genomic molecular biology tended to follow function first attitude and sought genes that were involved in providing that function. Post-genomics starts with all the genes, for many of which there is no corresponding biochemical activity or function known, and thus followed gene first attitude. At the present days, it is known that the function first mode of reasoning failed to find approximately 40% of the genes that were uncovered, even in well-worked model organisms, after whole-genome sequencing methods were applied [99]. This function first approach is an example of a traditional hypothesis-driven research, where the word hypothesis originates from the ancient Greek word which means to suppose. Thus, in pre-omics era, scientists supposed that there is a factor X causing phenomena Y and then they tested this hypothesis by performing the experiments which either proved or disproved the hypothesis. Such a method is risky, because the researcher first tries to predict or guess the biological hypothesis and then he starts to collect data pertaining to it. The problem here is that the scientist approaches the experiments with his inherent limitations of knowledge and experiential bias, merely by choosing one possible cause out of perhaps hundreds of possibilities [214].

The Broad Street cholera outbreak case is an excellent example of the data-driven research, because John Snow analyzed data to find the environmental or other characteristics that are differentially prevalent in those who were sick compared to those who were healthy. Although Snow had no access to metagenomics data in 1854, his research was essentially data-driven, because his water pump contamination hypothesis was the result of the data analysis, not its starting point. Furthermore, the analysis of large datasets allows to perform follow-up hypothesis-driven experiments, based on the results obtained from the data-driven ones. Such a combination of data-driven and hypothesis-driven research might lead to a better understanding of biological systems in health and disease [121, 2, 102].
1.2 Data-driven research in oral health

1.2.1 A novel approach in defining oral health

Oral health is an integral component of overall health and well-being. The human oral cavity is the entry point for all food, and the first active step of the human digestive system. Also, the mouth plays an essential role in verbal as well as non-verbal communication between people [130]. The oral cavity is an entry point for many pathogens, often enforcing systemic diseases [70, 105, 132, 133]. Whereas most of the previous studies on the oral health condition have mainly focused on progression and treatment of oral disease, less is known about the physiological and biological processes involved in the maintenance of good oral health.

Experimental data and some isolated evidences in the scientific literature [103, 85, 45, 151, 154, 48, 239] suggest that biological interactions between salivary components, the oral microbiota and the host defense system actively promote the oral health status. This may be an indication that oral health is the ability of the oral ecosystem to counteract perturbing stresses. These perturbing conditions could include the metabolic, inflammatory as well as oxidative stresses induced by diet, lifestyle and environment. The elucidation of the interplay between host, bacteria and stress requires application of an integrative approach using large amounts of data. Therefore, a good data integration model is crucial to define what oral health is.

1.2.2 Oral systems biology

Modern biology is far from understanding the full complexity of the human organism, because thousands of variables are involved in its proper functioning. To understand the interplay between these variables, an integration of metagenomic, transcriptomic, proteomic, and metabolomic data is required. Data-driven biology can help to generate new hypotheses from -omics data by discovering subgroups of individuals that have similar -omics profiles. These subgroups can subsequently be associated with ecosystem shifts and help to diagnose health problems at the early stage. Hence, data-driven biology can help to distinguish various healthy states, without requiring detailed knowledge of the exact mechanisms driving the changes in the underlying ecosystem. This approach is different from the traditional view on a health as an absence of symptoms. Data-driven biology provides an opportunity to progress medicine by providing treatment focused on restoring individual’s normal ecosystem, rather than merely suppressing the symptoms [209].

The human oral microbiome is a good example of a highly complex ecosystem within the human organism with around 1,000 species present [60] and it is considered to be the second most complex in the human body, after the gut ecosystem [43]. The human mouth is a host for over 600 prevalent taxa at the species level and it is estimated that 20 billion resident microbes exist in the oral cavity [60]. Thanks to progress in -omics techniques, it becomes clearer that the traditional “one microbe - one disease” approach is too simplistic and it does not reflect the complexity of the biological systems [71]. Instead of that, it is now established that many potentially dangerous microbial species live and interact with other microorganisms in vast communities [38, 164] and diseases associated with the commensal microbiota are not caused by a single species, but they result from the concerted actions of a
number of microbes [227, 165]. In a normal state, commensal microbes and their hosts are in a symbiotic equilibrium, but external or internal disturbances may lead to a condition where microbial imbalance exerts adverse effects on the host known as dysbiosis [7]. A healthy oral cavity is also a result of a well-tuned balance between the host and the oral microbiome, thus dysbiosis state may disrupt the normal functioning of the human oral cavity [137, 74]. To reduce the burden of oral chronic infections on general health and to keep the costs for oral healthcare at an acceptable level, more effective strategies for the prevention of oral diseases are needed [112]. An optimal approach in preventing oral diseases is to identify novel strategies for actual improvement of oral health status. Oral health is much more than absence of disease, but how we exactly need to define oral health is currently not clear. To face this challenge, future research in oral health should be directed towards the next level, using a data-driven systems biology approach [239].

Systems biology can be defined broadly as the integration of large amounts of biological data from various sources to create one or more comprehensive models of a system to enable: visualization of the changes in the various working parts within a particular system (e.g. data from changes in genes, proteins, or metabolites in response to different biological conditions); visualization of the known and/or predicted interactions between those parts; and creation of a mathematical model of interaction paths from which testable predictions about the system can be made [61]. Although provided definition makes emphasis on visualization part only, oral systems biology approach should also include biomarker selection, data clustering, multi-omics data integration, and predictive modeling. That is why application of the most appropriate computational algorithms to oral health data is crucial if one wants to understand what oral health is. This PhD thesis provides an overview of a diversity of problems in defining oral health as well as which computational challenges must be tackled to solve these problems.

1.2.3 Computational challenges in analysis of -omics data

The opportunities brought by introducing of new -omics technologies also bring new computational challenges, e.g. analysis of clinical proteomics data [187, 46], genetic data [224], and metabolomics data [216] required solving a few computational issues. The main challenge is a so called ”curse of dimensionality” [11], which means that the number of samples (examples) needed to accurately describe a problem increases exponentially with the number of dimensions (features). Biological data samples are expensive to collect, thus the number of examples n is usually much lower than the number of features p. In this a so called the “p >> n” case, many of the conventional statistical methods become numerically unstable and may lead to misleading conclusions. Another challenge is a multicollinearity of data features, i.e. many of the biological features are highly correlated to each other. These, if not properly handled, may lead to overfitting, i.e. when a model performs well on a training dataset, but makes poor predictions on a test dataset.

Although analysis of every heterogeneous -omics data layer separately may lead to useful discoveries, one can find even more useful biological knowledge using systems biology approach. Systems biology focuses on complex interactions within biological systems, using a holistic approach instead of the more traditional reductionism. This approach poses computational challenge of the most optimal -omics data integration strategy in many
fields of life sciences such as cancer research [228], immunology [100], or microbiology [177, 61, 203, 204]. This research area attracts the attention of many researchers who indicate various advantages of such an integrated approach [42, 201, 34, 40]. Biological data integration is an actively growing area and there are many different computational strategies for its successful implementation [163, 200, 15].

The next challenge, associated to the growth of data volume, is a lag between speed of data generation and speed of knowledge discovery. Since it is now possible to acquire large datasets quicker and cheaper than it was before, many studies focus on collecting the data without clear separation on treatment and control groups. A good example of one of such studies is the National Institute of Health’s Human Microbiome Project [153]. This project allowed to identify over 10,000 microbial species residing in 18 body parts collected from over 5,000 healthy individuals. Since all individuals were healthy, traditional data analysis methods focused merely on validating or rejecting statistical hypotheses are not directly applicable. Instead of finding statistically significant differences among known groups, -omics data-driven studies allow discovering previously unknown differences within a healthy population.

### 1.3 Machine learning in oral health: a brief overview

Systems biology approach applied to oral health domain may provide many opportunities to discover insights in functioning of oral ecosystem. Detailed insights in the processes which are involved in maintaining healthy oral ecosystem may enable the development of novel methods to evaluate oral health of an individual. Applying these insights in clinical practice would allow implementation of more effective diagnosis procedures and self-care instructions. Also, these insights would expand our fundamental knowledge about what oral health is. To identify the biological interactions that underlie maintenance of oral health, one should apply a data analysis strategy that integrates multiple heterogeneous -omics datasets using the best possible computational methodology.

Since conventional data analysis techniques cannot handle large amount of complex -omics data, one has to use appropriate ones. Results presented in this PhD thesis are obtained by using a set of data analysis techniques united under umbrella term ”Machine learning” [41]. This thesis, entitled ”Multi-view Learning and Deep Learning for Heterogeneous Biological Data to Maintain Oral Health”, is aimed to study the application of single-view machine learning, multi-view machine learning and deep learning techniques to address some of the important challenges in defining oral health. Machine learning is a growing area of science and it is impossible to cover all topics within a few pages of the prologue to this PhD thesis. Therefore, only a brief introduction to some of the relevant topics together with the references to major review papers are given in following subsections of this chapter, so that the reader would become familiar with the techniques used in the subsequent chapters.

#### 1.3.1 Supervised, unsupervised and semi-supervised machine learning

Machine learning is a set of techniques that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision
making \[143\]. Depending on the amount of \textit{a priori} information, machine learning can be divided into three types: supervised, unsupervised and semi-supervised machine learning. The goal of supervised machine learning is to learn a mapping from inputs \( x \) to outputs \( y \) sometimes called targets or labels, given a set of matching input-output pairs. Input training data \( x \) consists of \( n \) rows (samples, observations, examples) and \( p \) columns (features, variables, attributes). In this thesis, terms "example" and "feature" will be used to describe a training dataset.

In contrast to supervised learning, in unsupervised learning one has no access to training labels. The aim here is to find a natural division of the input data \( x \) into homogeneous groups sometimes called clusters. Therefore, unsupervised learning is often called data clustering \[183\]. Clustering techniques can be roughly divided into two categories: hierarchical and partitioning. In this thesis applications are demonstrated for a partitioning technique called spectral clustering \[222\]. This technique arises from concepts in spectral graph theory and the basic idea is to construct a weighted graph from the input dataset where each node represents a pattern and each weighted edge takes into account the similarity between patterns. Unlike classical partitioning clustering algorithms, spectral clustering is able to produce non-linear separating hypersurfaces among data, since it constructs an adjacency structure from the input dataset \[69\].

Semi-supervised learning is halfway between supervised and unsupervised learning. In addition to unlabeled data, the algorithm is provided with some supervision information. In this setting, this information will be the labels associated with only some part of the training examples \[32\].

\subsection{1.3.2 Feature selection and regularization}

The goal of supervised feature selection is to select a subset of input features, which can efficiently describe the input data while reducing effects from noise or irrelevant features and still provide good prediction results. Supervised feature selection techniques can be divided into three types: filter, wrapper and embedded \[170, 31, 80\]. In this thesis applications are demonstrated for supervised feature selection techniques called least absolute shrinkage and selection operator (LASSO) \[199\] and Elastic Net \[245\]. Both LASSO and Elastic Net belong to shrinkage techniques based on a regularization approach. The main idea of regularization is to fit a model to data containing all \( p \) features using a technique that constrains or regularizes the corresponding coefficient estimates, or equivalently, that shrinks the corresponding coefficient estimates towards zero \[97, 144\]. Besides feature selection, shrinkage techniques also efficiently reduce overfitting.

Unsupervised feature selection techniques aim to select subset of features that are relevant to data clustering. Many different unsupervised feature selection techniques were proposed to eliminate irrelevant and redundant features while keeping the relevant ones in order to improve clustering efficiency and quality \[3\]. Similar to feature selection for supervised learning, unsupervised feature selection techniques are categorized into filter \[52\], wrapper \[169\], and embedded techniques \[125\]. In this thesis applications are demonstrated for an unsupervised feature selection technique called Joint Embedding Learning and Sparse Regression \[92\], because it efficiently combines advantages of both spectral clustering and LASSO.
1.3.3 Multi-view learning and co-training

In many scientific domains data are collected from heterogeneous sources or obtained from diverse feature extractors. Therefore, features of each data example can be naturally partitioned into several heterogeneous groups. Each feature group is referred to as a particular view, and the multiple views for a particular problem can take different forms. Multi-view machine learning is a common name for a group of techniques which are specially designed to exploit all the advantages of datasets represented in a multi-view form. These techniques explicitly incorporate a multi-view training in an optimal way, which often results in an improvement of the algorithm’s learning performance [234, 193].

Multi-view machine learning algorithms can be categorized into three groups: co-training [19], multiple kernel learning [77], and subspace learning [84, 33, 231, 225]. Multi-view learning is based on consensus and complementary principles. The consensus principle aims to maximize the agreement on multiple distinct views. The complementary principle states that in a multi-view setting, each view of the data views may contain some knowledge that other views do not have [234].

In this thesis applications are demonstrated for the co-training style multi-view learning algorithms. The co-training style algorithms usually train separate optimization functions on distinct views, which are then regularized to be consistent on their predictions across all views. These algorithms are based on three main assumptions: sufficiency, compatibility, and conditional independence. Sufficiency assumptions states that each view is sufficient for classification on its own. Compatibility is based on the assumption that the target functions in all views predict the same labels for co-occurring features with high probability. Conditional independence assumption means that the views are assumed to be conditionally independent given the class label [234]. Theoretical analysis [19] and empirical verification [124] demonstrate that co-training works well if these assumptions are satisfied.

1.3.4 Deep learning

Deep learning refers to a class of machine learning techniques, where many hierarchical layers of information-processing stages are exploited for pattern recognition and representation learning. Deep learning architectures can be broadly categorized into three main architectures: generative, discriminative, and hybrid [58]. In this thesis applications are demonstrated for a discriminative deep architecture technique called Convolutional Neural Network (CNN) [51, 118, 107]. CNN architectures are designed to process data that come in the form of multiple structured arrays, for example in the form of a colour image, which is composed of three 2-dimensional arrays containing pixel intensities in the three colour channels.

There are four key ideas behind these architectures that take advantage of the properties of the structured arrays: local connections, shared weights, pooling and the use of many layers [116]. In general, the predictive performance of a model on previously unseen data, i.e. its generalization, can be improved if certain a priori information about the problem is added into the choice of the model architecture [117]. In case of images, a priori information of the problem can be implemented in a model if such a model is able to learn spatial information between the pixels of an image. This property is explicitly embedded into the CNN model via a discrete convolution operation [116].
1.4 A guideline for the reader

The study described in this thesis was performed within a framework of the project entitled “Novel strategies to promote oral health” funded by the Top Institute Food and Nutrition. Despite of oral health being the focus of this thesis, many of the techniques and approaches presented here are applicable to any branch of life sciences where the biological phenomena can be quantified by multiple heterogeneous -omics datasets. The chapters in this thesis represent a collection of articles, which are either published in or submitted to several scientific journals or presented on scientific conferences. Although the chapters are interlinked by the joint objective of applying machine learning algorithms to provide insights in oral health, they are generally self-contained and there is no need to read all chapters successively. The datasets used in this thesis were either downloaded from publicly available resources or were generated in ”Novel strategies to promote oral health” project.

1.4.1 The datasets used in the thesis

The machine learning algorithms described in the chapters of this thesis were applied to three different multi-view and single-view datasets. The studies described in Chapters 2, 3, and 4 used Human Microbiome Project single-view metagenomics dataset collected from about 5,000 healthy individuals [153]. Different strategies were used to adapt this dataset to a multi-view setting. Views in Chapter 2 were defined by randomly splitting the original dataset into two datasets with approximately equal number of features. Chapter 3 describes the application of spectral clustering algorithm to a dataset which was transformed from the original one, so that the examples in the transformed dataset would correspond to human individuals, not to oral niches. The multi-view dataset used in Chapter 4 was obtained from the original dataset by using various subsets of the oral niches as view indicators.

In Chapter 5, multi-view algorithm was applied to the National Cancer Institute’s NCI-60 dataset [161] [181]. This dataset consists of several multi-omics views collected from 60 different cancer patients. Hence, this dataset can be used directly without any prior transformations to a multi-view setting and it contains labels which, if necessary, can be used in a supervised machine learning setting.

The studies in Chapters 6, 7, 8 describe the application of single-view machine learning techniques to various views of multi-view datasets collected in ”Novel strategies to promote oral health” project. The study in Chapters 6 describes the application of unsupervised learning and supervised feature selection algorithms. The studies in Chapters 7 and 8 describe the application of deep learning and supervised feature selection algorithms respectively. Unlike the dataset used in Chapter 6, the dataset used in Chapters 7 and 8 has clinical labels in a form of dental plaque data, because the goal of the second clinical study within the project was to observe changes in dental plaque accumulation induced during clinical intervention.

1.4.2 Outline of the thesis

In Chapter 2 an online multi-view semi-supervised algorithm is applied to the Human Microbiome Project dataset to show that a large amount of unlabelled data helps to improve the
model’s predictive performance. In Chapter 3 a personalized microbial network is inferred via application of the co-regularized spectral clustering algorithm to the Human Microbiome Project dataset to visualize clusters of healthy individuals as a network of interacting microbial species. Chapter 4 presents a novel Unsupervised Multi-View Feature Selection algorithm applied to the Human Microbiome Project dataset to show that there are only a few important features driving differences among clusters, when taking into account the interaction between different views defined by oral niches. In Chapter 5, the same method is applied to the National Cancer Institute NCI-60 dataset panel, to model the interaction between different -omics datasets defined by clustering of tumour subtypes and to select the corresponding discriminative features. All of the algorithms used in Chapters 2, 4 and 5 are multi-view.

Chapter 6 demonstrates an application of both unsupervised and supervised machine learning models on salivary metagenomics, salivary metabolomics, and salivary biochemistry datasets to show the diversity in the range of healthy parameters of oral health. A deep multi-layer convolutional neural network is applied to the classification of the dental plaque accumulation images in Chapter 7, where the performance of the Deep learning model is compared with the performance of other shallow models. In Chapter 8 supervised feature selection is applied to predict plaque accumulation and to extract predictive biomarkers per each of the five oral microbial niches. All algorithms used in Chapters 3, 6, 7 and 8 are single-view. Chapter 9 is the epilogue which contains general discussion and overview of the thesis.