The First Mining Science Data for Medicine Challenge (MiSDaM01) Handbook

Miriam Barry[1], Mike Merchant[2], Alfred Oliver[3], Ken Raj[4], Marina Romanchikova[1], Stephen Watts[5,*]

[1]National Physical Laboratory, UK, [2]Division of Cancer Sciences, The University of Manchester, [3]University of Hull, [4]Radiation Effects Department, Centre for Radiation, Chemical and Environmental Hazards, Public Health England, [5]School of Physics and Astronomy, The University of Manchester.

Draft Version March 25th 2019

Abstract

This paper describes the science and data for the first Mining Science Data for Medicine Challenge. Participating teams are encouraged to register for this challenge which has an end date of 26th August 2019. The challenge has two parts. First, to use machine learning (ML) to identify a set of DNA CpG methylations that are specific to senescent cells. Second, to make recommendations on how ML results can be explained to patients, doctors and the public. Results from participants will be made available in early 2020 following a "Results Workshop" in September 2019.

1. Background

This project resulted from a Sandpit Event organized by the STFC Global Challenges Network+ in Advanced Radiography [1]. The MisDaM project has three aims,

i) To obtain interesting medical science results with potential to apply to individual patients.
ii) Create a community of data miners to support the analysis of big data associated with medical science.
iii) Identify the algorithms and visualisations that are useful in this science area.

To achieve these aims, the project team will release challenges to the global community and invite anyone to solve a specific medical science problem using data mining and machine learning.

This handbook provides details on how anyone can get involved in the first challenge, called MiSDaM01.

2. The first challenge – MiSDaM01

We have followed a similar format to the GREAT08 Challenge in astrophysics [2]. However, there is no leader board or competition. It is a collaborative challenge with teams using different methods to encourage friendly competition. The teams need to register, agree to the rule conditions, obtain the data, and then attend a workshop in September 2019 to present their results. A summary of the results will then be published with all teams involved and methods compared.

The first challenge is in two parts, and participants are welcome to join in either or both parts. The first and second part of the challenge is described in Sections 2.1 and 2.2 respectively.

2.1 Identification of DNA methylation–based markers of cellular senescence.

[*] Corresponding Author: Stephen.Watts@manchester.ac.uk

This is the first part of the challenge.

Human cells have a finite capacity to proliferate. After a certain number of divisions, they retire from proliferation and enter an irreversible state that is called senescence. Cells enter senescence due to a number of reasons; one of which is when they are unable to repair damaged DNA. Many things such as radiation can damage DNA. When cells try and fail to repair the damage, they enter the senescent state. Importantly, senescent cells accumulate in increasing numbers as a function of age.

Senescent cells were initially thought to be benign – that although they do not contribute to the functioning of the tissue in which they reside, they are harmless. This view is clearly wrong as we now know that senescent cells synthesise proteins that they should not and secrete them inappropriately to neighbouring cells and the blood stream, imposing a detrimental effect on their function. This is one of the major causes of tissue and organ dysfunction associated with old age.

In spite of their importance, the current methods for detecting senescent cells are insensitive. Biologists have for decades known that cellular DNA undergoes changes with age. These are not changes to the DNA sequence (A,C,G,T). Instead, one of the four DNA bases exists in two possible forms – cytosine (C) or methyl-cytosine (mC). The methyl modification of cytosine is almost exclusively the preserve of cytosines that precedes a guanine (G). These cytosines are referred to as CpGs, and there are 28 million of them in the human genome and approximately 70-80% of these are methylated. With age however, some of the unmodified CpGs acquire the methyl modification while some modified CpGs lose their modifications. About 5 years ago, mathematicians took an interest in the vast amount of DNA methylation data that were deposited in open access platforms. These data consisted of methylation states of thousands of specific CpGs of human genomes from across a wide range of ages (birth to 100years old). They employed machine learning techniques, to correlate the DNA methylation states of specific CpGs to age. While the majority of CpG methylations was found to be unrelated to age, a small proportion of them were (either increasing or decreasing with age). Machine learning analyses were able to consolidate methylation changes at these specific age-related CpGs and discern a clear correlation between them and age. The correlation was so clear and strong that it lent itself to the generation of age-predictive algorithms that we refer to as epigenetic clocks.

A review of the science background can be found in ref. [3].

2.1.1 The Aim

We have since ascertained that while epigenetic clocks predict age to a very high degree of accuracy, they do not measure senescent cells, which also increase with age. Instead epigenetic clocks measure a route of ageing that is distinct from senescent cells. Since senescent cells undoubtedly contribute to ageing and its associated diseases, it is hoped that machine learning can once again be used to identify a set of CpG methylations that is specific to senescent cells. If possible, it would firstly provide a new and sensitive assay to detect senescent cells and secondly, it would enable the consolidation of senescent cells with epigenetic ageing into a unified and single age-determining algorithm.

2.1.2 The Data

Towards this end we have, as a start, generated 24 cell populations (fibroblasts) from human donors. One set of these 24 were grown as control, while the second corresponding set was irradiated with X-rays (20Gy) and after 2 weeks, they became senescent (still alive but no longer actively dividing). DNA from the two sets of cells (24 control and 24 irradiated) were isolated, purified and the methylation states of 850,000 specific CpGs of their DNA was ascertained. The data, which is now available for analyses, is tabulated in Excel format. Appendix 1 describes how the data can be obtained.

2.1.3 Information pertaining to the data

As mentioned above, the data is a measurement of the degree of methylation of a CpG at a specific location in the human genome. Since methylation is a binary event, i.e. a CpG is either methylated or not, it is intuitive to expect the 850,000 CpGs to have the numerical value of "0", for un-methylated or "1", for methylated. This however is not the case. Instead, the value of each of the 850,000 CpGs lies between 0 and 1. This is because the values obtained are not from a single cell (which would be binary), but a population of thousands of cells. Cell populations are heterogenous, in that a particular CpG may be methylated in some cells, but unmethylated in others. As a case in point, a value of 0.6667 for a particular CpG means that it is methylated in 67% of cells in that population. These measurements are called "Beta-values". The beta values change as a function of many things including age (as was demonstrated) and cell state (the aim of this challenge). The DNA methylation heterogeneity of cells is a feature that is inherent in all cell populations. Despite this, or perhaps because of this, methylation of CpGs have been very successfully used to predict to a high degree of accuracy, biological age of individuals, time-to-pathology and even time-to-death. These successes attest to the power of machine learning in identifying patterns that are defined by very small changes of a very large number of data points. It is highly likely that similar approaches will be able to successfully identify DNA methylation patterns that are specific to senescent cells (this challenge).

The data consist of beta values from 24 un-irradiated cell populations and 24 irradiated cell populations. These are "paired samples", in that the 24 samples from each group are cells derived from the same 24 human donors. For example, "Sample 6" and "Sample 6X" are cells isolated from human donor 6, whereby a portion of the cells are not irradiated (the former) and another portion is X-irradiated (6X). This explanation does not suggest that there is a need to pair up samples for analyses. Samples can be analysed as two collective entities of un-irradiated vs irradiated. There are no restrictions to how they are analysed.

2.2 Explaining Machine Learning (ML) results to patients and doctors

This is the second part of the challenge.

The ability of patients to understand, and doctors to understand and explain ML predictions, especially for complicated situations with major healthcare consequences, is an important and topical issue. A recent paper on this subject is given in ref. [4].

Teams are invited to suggest how ML results can be explained to patients, doctors and the public using;

Either a) their analysis of the cellular senescence data described in Section 2.1

or b) apply ML to the much smaller and well known CORIS dataset on coronary heart disease, which can be downloaded at [5].

or both.

Teams are also encouraged to illustrate their ideas by referencing to other publically available datasets.

References

[1] https://www.advanced-radiotherapy.ac.uk/

[2] "Handbook for the GREAT08 Challenge: An image analysis competition for cosmological lensing", Sarah Bridle et al., Ann. Appl. Stat., Volume 3, Number 1 (2009), 6-37.

https://projecteuclid.org/euclid.aoas/1239888361, doi:10.1214/08-AOAS222

[3] "DNA methylation-based biomarkers and the epigenetic clock theory of ageing", Steve Horvath and Kenneth Raj. Nat Rev Genet. 2018 Jun;19(6):371-384. doi: 10.1038/s41576-018-0004-3

[4] " Clinical applications of machine learning algorithms: beyond the black box", David Watson et al. BMJ 2019;364:l886 doi: 10.1136/bmj.l886 (Published 12 March 2019)
https://www.bmj.com/content/364/bmj.l886

[5] CORIS dataset. Original data, J. E. Rossouw et al., " The prevalence of ischaemic heart disease in three rural South African communities", South African Medical Journal, 1983. Reduced dataset used by Hastie and Tibshirani, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 36, No. 3(1987), pp. 260-276. Data available at,

http://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.datahttp://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.data

Appendix 1 Challenge Rules and How to Register

A1.1 Challenge Rules

We are grateful to the GREAT08 Challenge Handbook,[2], from which these rules derive.

1) The data will be released publically at the end of the challenge.
2) The challenge end date will be shortly before the challenge " Results Workshop" in September 2019. To be specific, the end date is, 26 August 2019.
3) All registered teams are expected to present their results at the "Results Workshop" in September 2019. The date for the results workshop will be agreed at the kick-off workshop on 16 April 2019.
4) Teams signing up to the project will have access to the "DNA methylation–based marker of cellular senescence" data under condition that they abide by these rules.
5) Teams will be encouraged to publish if they wish, but commit not to do so until after the final challenge report has been issued in pre-print form to the arXiv or January 2020, whichever date is earliest.
6) Publications from individual teams should acknowledge the source of the data and the MiSDaM01 challenge. The fact that the data was provided as the result of UKRI/STFC funding should be acknowledged.
7) Participants may use a pseudonym or team name on the participants list, however real names (as used in publications) must be provided when requested during the result submission process.
8) Participants must provide a report detailing the results and methods used, at the challenge deadline. We would prefer that any code developed for this challenge is made public. We suggest the use of GitHuB.
9) We expect all participants to allow their results to be included in the final Challenge Report. We will however be flexible in cases where methods performed badly compared to other methods or if participants are strongly against publicising them.
10) Clarification concerning the data will be provided (if possible) when requested and will be made available to all participants.

Some additional competition rules apply to members of the Local Project Team who submit entries:

A) For the purpose of these rules, "Local Project Team" includes the authors of this document and staff or students associated with them.

B) Only information available to non-team participants may be used in carrying out the analysis.

A1.2 How to register for the challenge ?

A link to the data will be available on the project website, http://www.hep.manchester.ac.uk/MiSDaM/ and will be delivered via the workshop site at

http://indico.hep.manchester.ac.uk/categoryDisplay.py?categId=59. Instructions on obtaining a password can be found at the site. Downloading the data will automatically imply that the participating team has agreed to abide by the challenge rules.