Cost-sensitive feature selection - information theory approach

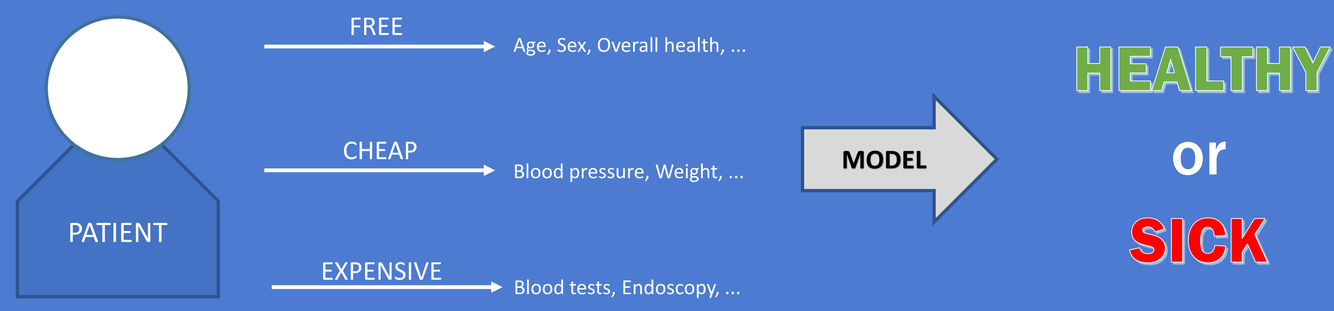
Tomasz Klonecki - Institute of Computer Science, Polish Academy of Sciences

Research Objective

Feature selection is a crucial problem in many bioinformatics tasks. Usually the considered variables are cheap to collect and store but in some situations the acquisition of feature values can be problematic. For example, when predicting the occurrence of the disease we may consider the results of some diagnostic tests which can be very expensive.

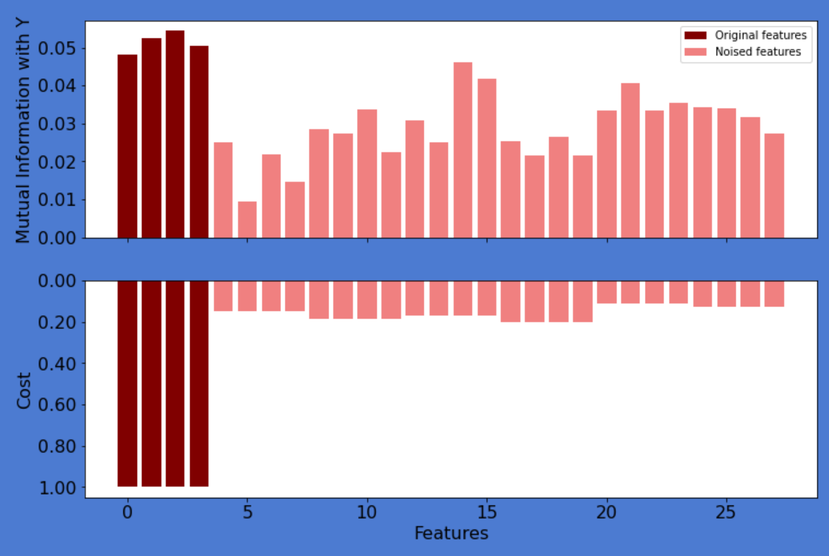
The existing feature selection methods usually ignore costs associated with the considered features. The goal of cost- sensitive feature selection is to select a subset of features which allow to predict the target variable (e.g. occurrence of the diseases) successfully within the assumed budget.

The main purpose of this research is to review filter methods of feature selection based on information theory and to propose new variants of these methods considering feature costs.

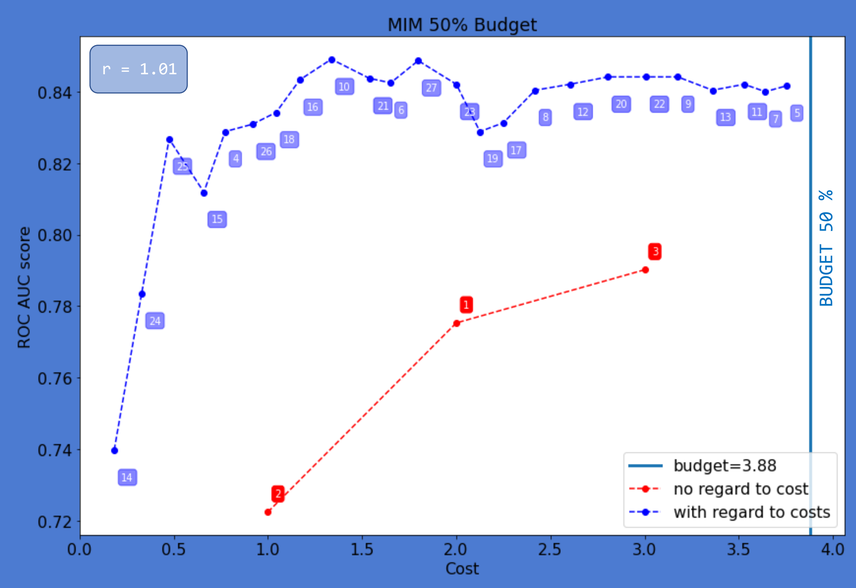


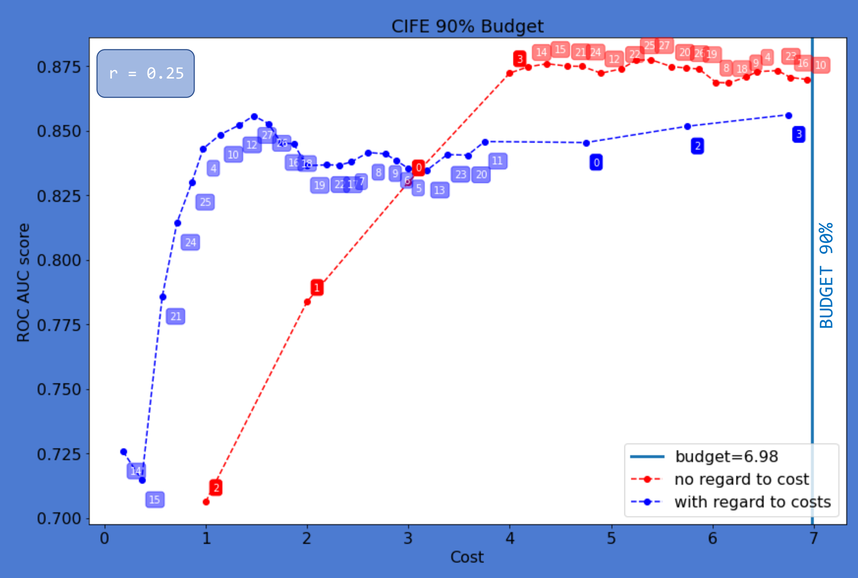
Artificial Dataset

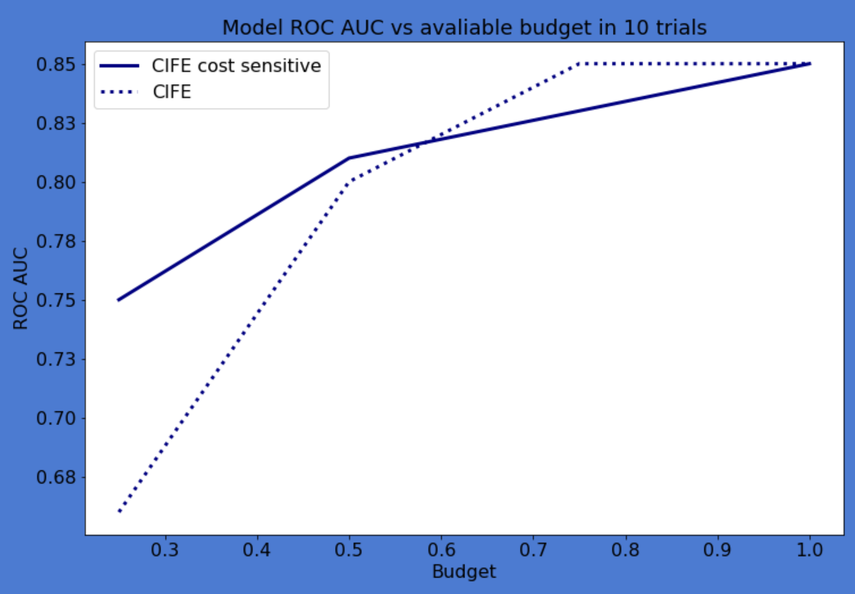
1. Generate original features from normal distribution $X_1, X_2, \ldots, X_p \sim N(0,1)$
2. Generate target variable Y based on $X_1, X_2, \ldots, X_p$ with binomial distribution.
3. Generate noised features $X_i' = X_i + E_j$ where $E_j \sim N(0, \sigma_j)$.
4. Assign cost to each feature $c_i = 1$ and $c_{i(j)}' = \frac{1}{1+\sigma_j}$.
5. Discretize data with uniform method (each bucket range is equal length) for 20 buckets.



Experiments







Feature Selection Procedure

Problem statement

$$S^* = \arg\max_{S:C(S)<B} I(Y,S)$$

Iterative greedy algorithm

$$X_k = \arg\max_{X_k:C(S+X_k)<B} F(I(Y,X_k|S), c_{X_k})$$

Specific form of greedy algorithm

$$X_k = \arg\max_{X_k:C(S+X_k)<B} \frac{I(Y,X_k|S)}{(c_k)^r}$$

SOLVE

F FUNCTION EXAMPLE

Approximations of the conditional mutual information

$$I(Y,X_k|S) = I(Y,S\cup X_K) - I(Y,S) = \begin{cases} J_{MIM}(Y,X_k) = I(Y,X_k) \\ J_{MIFS}(Y,X_k|S) = I(Y,X_k) - \beta \sum_{X_j \in S} I(X_k,X_j) \\ J_{CIFE}(Y,X_k|S) = I(Y,X_k) - \beta \sum_{X_j \in S} [I(X_k,X_j) - I(X_k,X_j|Y)] \end{cases}$$

MIMIC3 Dataset

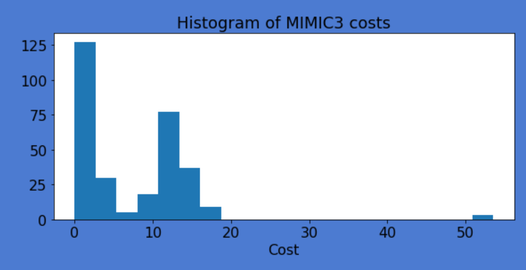
MIMIC III is one of the most popular medical datasets in the world. For experiments we use data of 6500 patients.

Types of features:

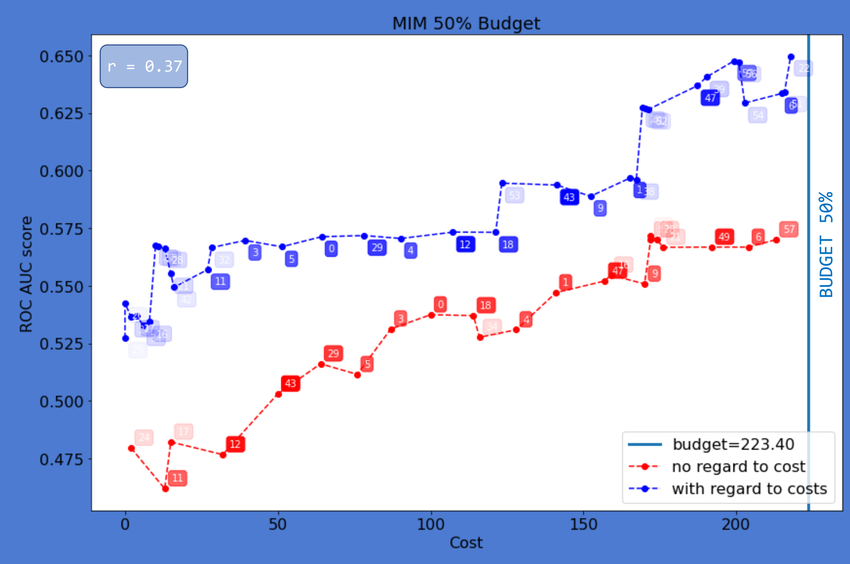
- basic patient information (Age, gender, ...)
- basic medical tests (HR, Blood pressure, ...)
- advanced medical tests (Blood tests, Urine tests, ...)

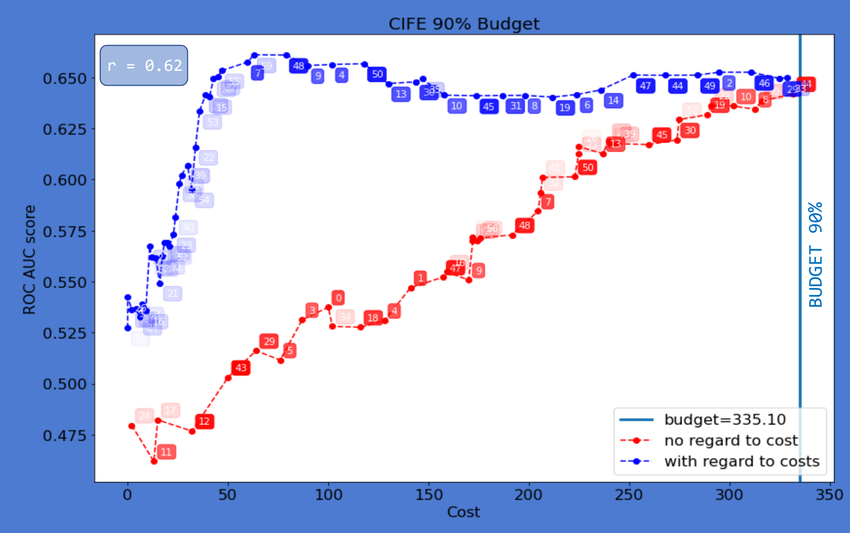
Target variable:

We can choose one of many target variables, each represents a positive or negative diagnosis of the specific disease. For experiments on this poster we will focus on **hypertension disease**, which almost 4500 patients were diagnosed with.



ID	Feature	Cost	ID	Feature	Cost
0	Anion Gap Blood STD	13.0	30	Not clear urine CNT	14.0
1	Anion Gap Blood RNG	13.0	31	Bilirubin in urne NEG	13.0
3	Calcium in blood STD	11.0	32	Color of urine OTHER	1.3
4	Creatinine AVG	12.0	34	Leukocyter in Urine	2.0
5	Creatinine MED	12.0	35	PH of Urnie AVG	3.0
6	Creatinine STD	12.0	36	PH of Urine MED	3.0
8	Phosphate AVG	11.0	37	Gravity of urine AVG	2.0
9	Phosphate MED	11.0	38	Gravity of urine RNG	2.0
10	Potassium AVG	11.0	39	Urobilinogen in urine MEDLeve	3.0
11	Potassium MED	11.0	40	Age	1.0
12	Sodium RNG	17.0	41	Activity tolerance GOOD	1.0
15	Hermatocrit AVG	2.0	42	Activity tolerance POOR	1.0
16	Hermatocrit MED	2.0	43	Body surface at admission	18.0
17	Hemoglobin Blood AVG	2.0	45	Braden moisture	16.0
18	INR in blood MED	14.0	47	Braden Nutrition POOR	16.0
20	Erythrocyte MED	2.0	48	Braden Sensory Percep NO IMPAIR	16.0
21	Erythrocyte volume AVG	2.0	49	Braden Sensory Percep LIMIT	16.0
22	Erythrocyte volume MED	2.0	51	Ectopy Frequency PRESENT	1.0
23	Erythrocyte volume STD	2.0	52	Ectopy type NONE	1.0
24	Platelets in blood RNG	2.0	53	Eye opening SPONTAN	2.0
25	APTT in blood STD	0.0	54	Eye opening STIMUL	2.0
26	APTT in blood RNG	0.0	55	Eye opening NO	2.0
27	Erythrocyte dist MED	2.0	56	Heart Rate AVG	1.5
28	Leukocytes MED	2.0	57	Lung Sound NOT CLEAR	9.0
29	Clear urine CNT	14.0	58	Level of conscious ALERT	1.0





GitHub

For the purposes of this research, I created an open-source library in Python, the library includes:

- Feature selection using information theory.
- Cost sensitive feature selection.
- Generating artificial data sets.

BC selector

https://github.com/Kaketo/bcselector

Conclusions

- Cost sensitive feature selection methods choose variables much more cost efficient than traditional methods.
- We experimented with various F functions, but division function is the most natural way of scaling two completely different numbers (costs and information increase).
- We are currently experimenting with r parameter selection, to obtain the best possible results. Method is based on maximization of J criterion increases.
- In future we will try to extend our selection method to consider features with shared cost. For example various blood results can be obtained during one test, for which we pay only once.

Hierarchical clustering in search for the most relevant variables in small-n-large-p datasets

Radosław Piliszek and Witold Rudnicki
Computational Centre, University of Bialystok

Introduction

Gene expression and genomic datasets from biomedical studies belong to the so-called small-n-large-p class. Such datasets describe a relatively small number of objects (records), counted in tens, hundreds and thousands, using a large number of variables (features), counted in tens, hundreds and thousands of thousands. Many machine learning algorithms suffer performance penalties in such a case. Moreover, human analysis of the studied phenomenon is severely hampered.

Various feature selection algorithms have been proposed to tackle this problem. However, there might still exist many relevant features. A naive approach of top-N ranking will usually discard relevant information and still keep sets of variables carrying the exact same information. Eliminating correlations upfront is of no use because correlation does not map exactly to information about the decision variable.

Datasets under scrutiny

The presented results have been obtained on datasets from the CAMDA 2017 Neuroblastoma Data Integration Challenge. There are 3 datasets in total, all describing the same set of **145** patients:

- **CNV** – **39 115** array comparative genomic hybridization (aCGH) copy number variation (CNV) profiles,
- **MA** – **43 349** GE profiles analysed with Agilent 44K microarrays,
- **G** – **60 778** RNA-seq GE profiles at gene level.

Proposed algorithms

We reuse the concept of hierarchical clustering applied in a bottom-up fashion (i.e. starting from one-feature clusters) but modify its linkage properties. The most common linkage – single (also known as minimum linkage) does not suit the problem well because of its tendency to merge early. There is also no clear notion of the cluster representative in the basic hierarchical clustering. Average linkage does not apply either because it is not known what an average feature would mean. Hence, we propose representative-based linkage with 3 ways to establish the representative:

- **HCN** – hierarchical clustering with native (natural) ordering – using the ordering from all tuples of potentially relevant variables,
- **HCO** – hierarchical clustering with original ordering – using the ordering from initial MDFS-2D output,
- **HCS** – hierarchical clustering with subset ordering – using the ordering from MDFS-2D applied only on potentially relevant variables.

Results

	CNV						MA						G					
IG	HCN		HCO		HCS		HCN		HCO		HCS		HCN		HCO		HCS	
1	150	0.24	142	0.22	150	0.21	978	0.12	991	0.12	974	0.16	1194	0.12	1195	0.12	1184	0.13
2	98	0.21	100	0.22	96	0.21	447	0.13	460	0.11	450	0.13	547	0.10	574	0.10	544	0.12
3	59	0.21	57	0.17	63	0.22	218	0.10	227	0.12	216	0.13	271	0.09	291	**0.07**	276	0.13
4	40	0.17	39	0.19	36	0.19	106	0.09	120	0.10	107	0.12	137	**0.07**	152	0.08	142	**0.11**
5	26	0.19	23	0.17	26	0.19	53	**0.08**	71	0.08	60	0.14	67	0.08	72	0.08	76	0.15
6	13	0.15	13	**0.15**	17	**0.17**	28	0.10	43	**0.07**	37	0.12	36	0.08	40	**0.07**	45	0.12
7	11	0.15	10	0.17	10	0.20	15	0.09	26	0.08	19	0.12	20	0.08	20	0.08	25	0.13
8	8	**0.13**	8	0.16	6	0.20	9	0.11	18	**0.07**	13	0.13	15	0.10	15	0.08	18	0.14
9	6	**0.13**	6	0.17	4	0.22	7	0.12	11	0.08	9	**0.10**	9	0.12	8	0.12	9	0.13
10	5	0.17	5	**0.15**	4	0.22	5	0.12	6	0.10	7	**0.10**	5	0.15	5	0.15	5	0.15
11	2	0.14	2	0.24	1	-	3	0.13	6	0.10	3	0.12	3	0.14	4	0.10	3	0.19
12	2	0.14	1	-	-	-	2	0.15	3	0.15	2	0.13	1	-	3	0.12	3	0.19
13	2	0.14	-	-	-	-	1	-	3	0.15	2	0.13	-	-	1	-	3	0.19
14	1	-	-	-	-	-	-	-	3	0.15	-	-	-	-	-	-	2	0.18
15	-	-	-	-	-	-	-	-	2	0.19	-	-	-	-	-	-	2	0.18
16	-	-	-	-	-	-	-	-	2	0.19	-	-	-	-	-	-	2	0.18
17	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1	-

The very first column (IG) shows the threshold at which the result is obtained. First subcolumn of the following columns shows the number of clusters (representing features). Second shows the OOB score (the less, the better; the best score in **bold**).

Our proposal

We propose an approach to limit the number of variables further by clustering variables using an existing measure of relevant variable discovery and scoring – the MultiDimensional Feature Selection (MDFS). We searched for clusters of variables having relatively negligible information gain between themselves. Each cluster is then replaced by the cluster representative variable. There are, however, several ways to build such clusters, even when constrained to hierarchical methods. There are also different ways to choose the representative.

Methodology

The basis for our research is the information gain (IG) metric as obtainable from MDFS. In particular, the interesting one is the two-dimensional MDFS variant, also called MDFS-2D. Such a metric can be computed two-way, once to obtain the potential relevant variables list (along with their tentative ranking). Secondly, to compute all pairwise IG values for selected features. These both serve as the input to further, clustering algorithms which are meant to remove redundancy from the selection.

It is unknown upfront what threshold of IG is relevant for a particular case. Hence, we compute classification score using random forest OOB score from features selected at integer levels of IG threshold (since they map to integer increases in explainability).

The potentially relevant features are discovered using MDFS-2D with 30 random discretisations and Benjamini-Yekutieli p-value adjustment. The cutoff threshold is set to 0.10.

Discussion

The different variants of the algorithm behave differently and may give varying results even with the same threshold and/or number of clusters.

The subset variant (HCS) performs noticeably worse. This might be due to losing the information about really relevant variables.

Further research is required, including different datasets, especially artificial ones with a known structure, and cross-validation.

Furthermore, it can be argued that reapplying clustering algorithms designed for object clustering may give suboptimal results for feature clustering as they disregard important properties not present in object relations, e.g. correlations and synergies. For such cases a more dedicated approach might be needed.

Bibliography

- Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, Wang J, Furlanello C, Devanarayan V, Cheng J, Deng Y. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome biology. 2015 Dec;16(1):1-2.
- Polewko-Klim A, Lesiński W, Mnich K, Piliszek R, Rudnicki WR. Integration of multiple types of genetic markers for neuroblastoma may contribute to improved prediction of the overall survival. Biology Direct. 2018 Jan 1;13(1):17.
- Piliszek R, Mnich K, Migacz S, Tabaszewski P, Sulecki A, Polewko-Klim A, Rudnicki WR. MDFS: MultiDimensional Feature Selection in R. R J.. 2019 Jun 1;11(1):198.
- Mnich K, Rudnicki WR. All-relevant feature selection using multidimensional filters with exhaustive search. Information Sciences. 2020 Mar 12.

All computations have been carried out on the computer cluster of the Computational Centre of University of Bialystok.

Exploring the microbiome protein structure space using simulations and deep learning

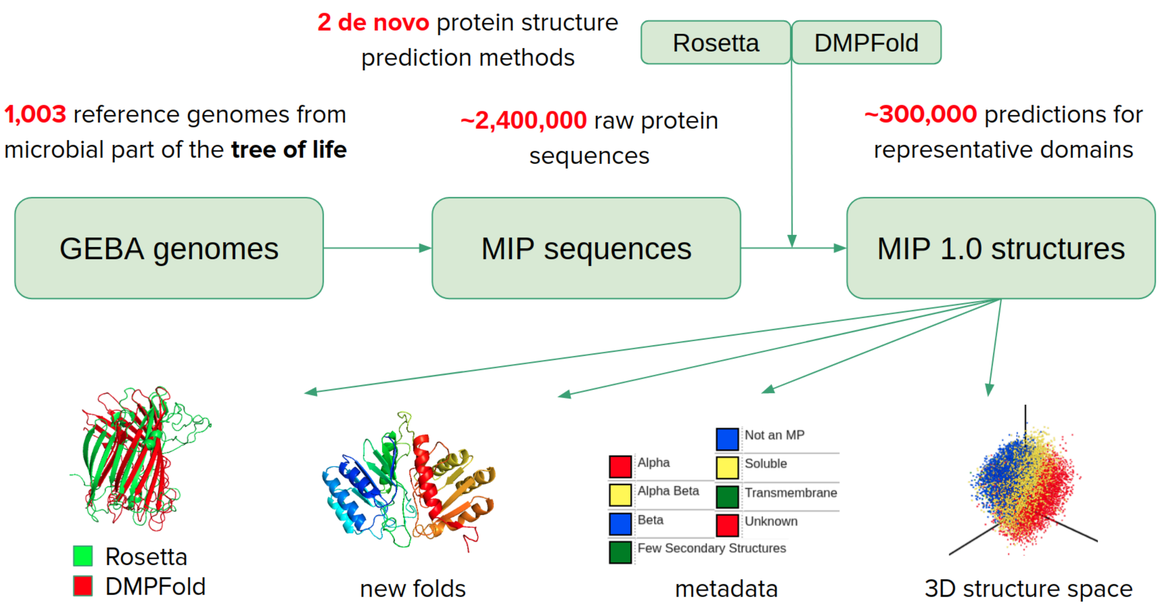
Paweł Szczerbiak[1], Douglas Renfrew[2], Julia Koehler Leman[2], Daniel Berenberg[2],
Chris Chandler[2], Vladimir Gligorijević[2], Richard Bonneau[2], Tomasz Kościółek[1]

[1]Małopolska Centre of Biotechnology, Jagiellonian University
[2]Flatiron Institute, Simons Foundation, New York, USA

Microbiome Immunity Project (MIP)

- Human gut microbiome comprises about 3 million unique bacterial genes
- Main goal of the MIP [1] is to understand the role played by microbiome bacteria
- Exploring them would give us a possibility to treat diseases that originate in our microbiome

In the first stage of the project we want to map all proteins produced by those bacteria. For this purpose we prepared a dataset consisting of ∼300,000 unique newly predicted structures which we call MIP 1.0. We used two methods: Rosetta [2] and DMPFold [3] which utilize different approaches to the protein structure prediction problem.



In the poster we are showing differences between both methods with special emphasis on new folds identification and structure space visualization. We also plan to create an open access database that anyone can use in their own analysis.

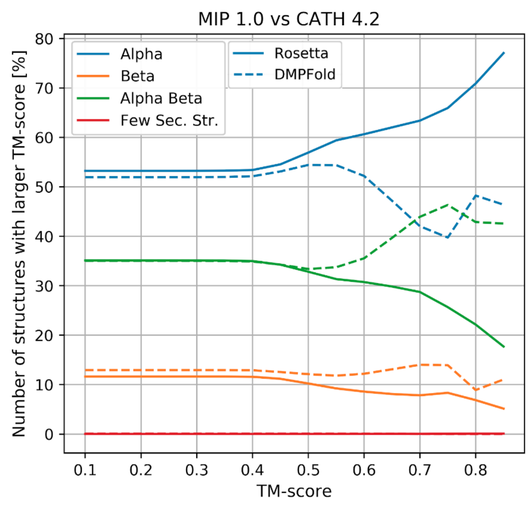
Rosetta vs DMPFold

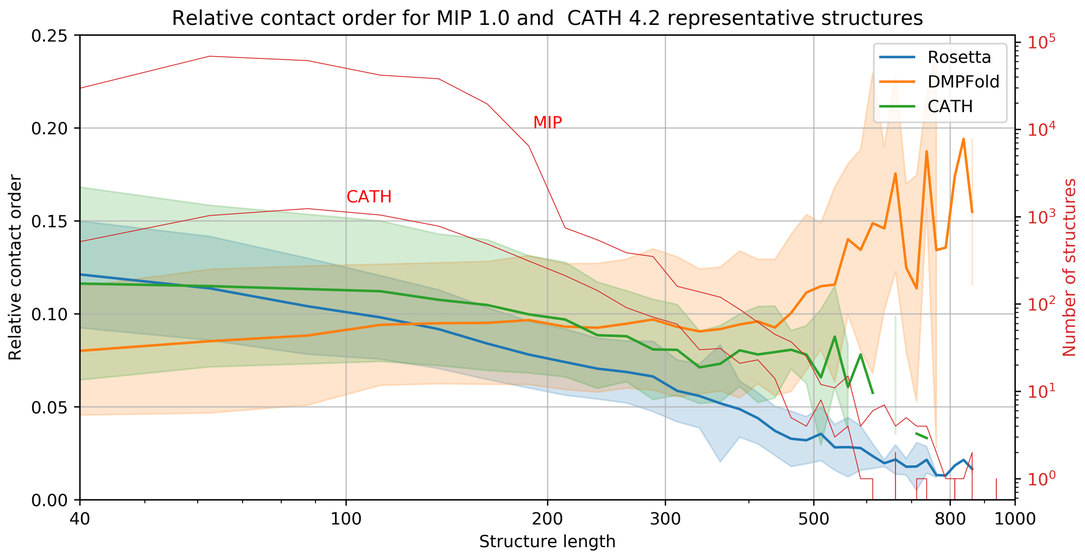
Rosetta

- Developed in 2002 but constantly improved
- Monte Carlo search through space of conformations to find minimal energy fold

DMPFold

- Developed in 2018 deep learning based procedure of inter-atomic distances, torsion angles and hydrogen bonds prediction
- Faster than Rosetta; predicts less α (more α/β and β) structures





References

1. www.worldcommunitygrid.org/research/mip1/overview.do
2. C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein structure prediction using rosetta Methods Enzymol., 383 (2004), pp. 66-93.
3. J.G. Greener, S.M. Kandathil, D.T. Jones, Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat Commun 10, 3977 (2019).
4. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations, PNAS, 117: 1496-1503 (2020).

New folds

Working definition: structures with TM-score below some predefined threshold (usually 0.5) with respect to the known fold space.

Non-redundant databases (our choice):

- CATH superfamilies (6119) – **done**
- PDB90 (∼60k) – **to be done**

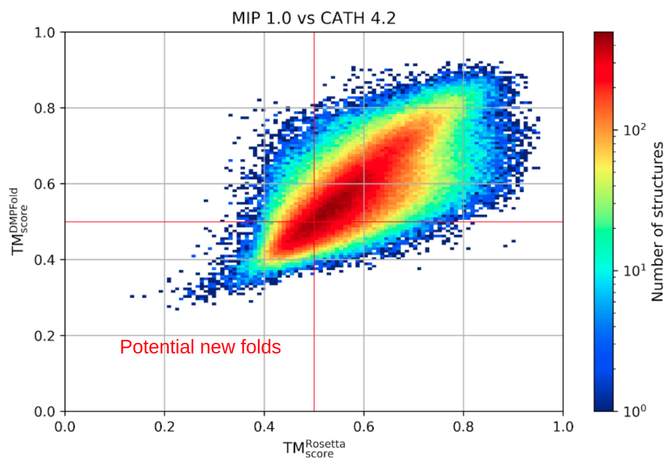
TM-score:

]0 – 0.5[different folds

[0.5 – 0.1] same folds

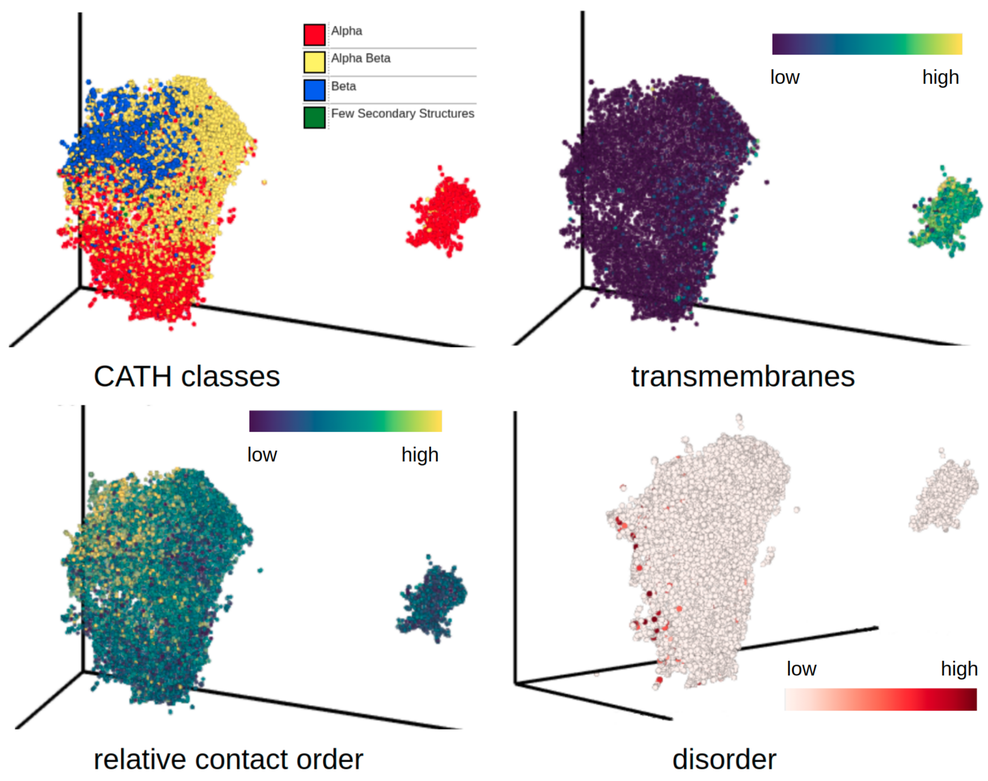
TM-score = **0.3**

TM-score = **0.7**



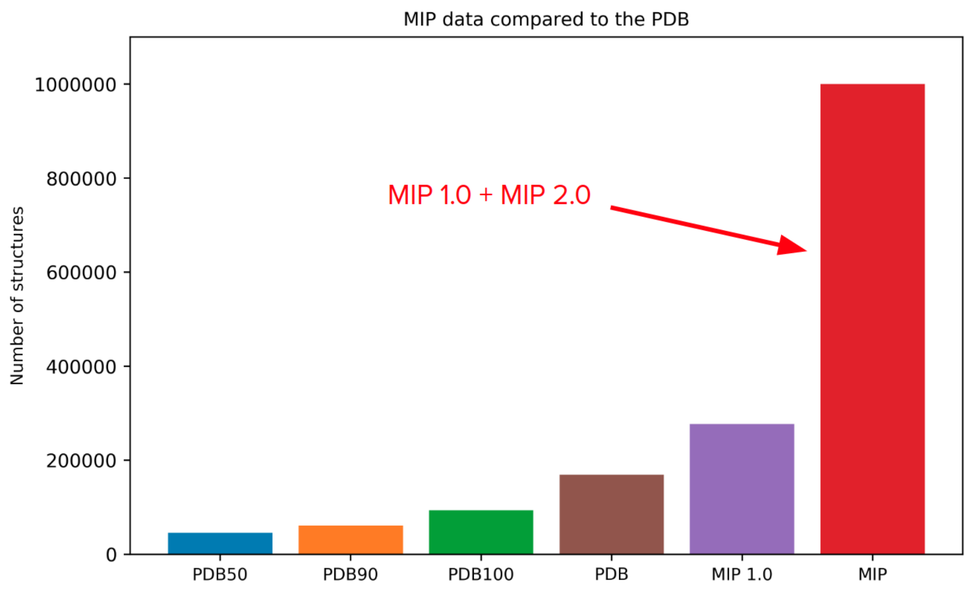
Structure space visualization

- Structure models were encoded using pretrained autoencoders
- Number of dimensions was further reduced using UMAP
- Visualizations show ∼9,000 Rosetta and DMPFold models



What's next?

- Our ultimate goal is to reach ∼1,000,000 annotated protein models
- MIP 2.0 will gather structures from the Unified Human Gastrointestinal Genome catalogue
- For structure prediction we will use trRosetta [4] – improved, deep learning inspired Rosetta



Comprehensive functional annotation of metagenomes and microbial genomes using deep learning-based method

Mary Maranga[1], Paweł P. Łabaj[1], Richard Bonneau[2], Tommi Vatanen[3,4], Tomasz Kościółek[1]

[1]Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland
[2] Flatiron Institute, New York, NY, USA
[3] Liggins Institute, University of Auckland, New Zealand
[4] Broad Institute, Cambridge, MA, USA

Introduction

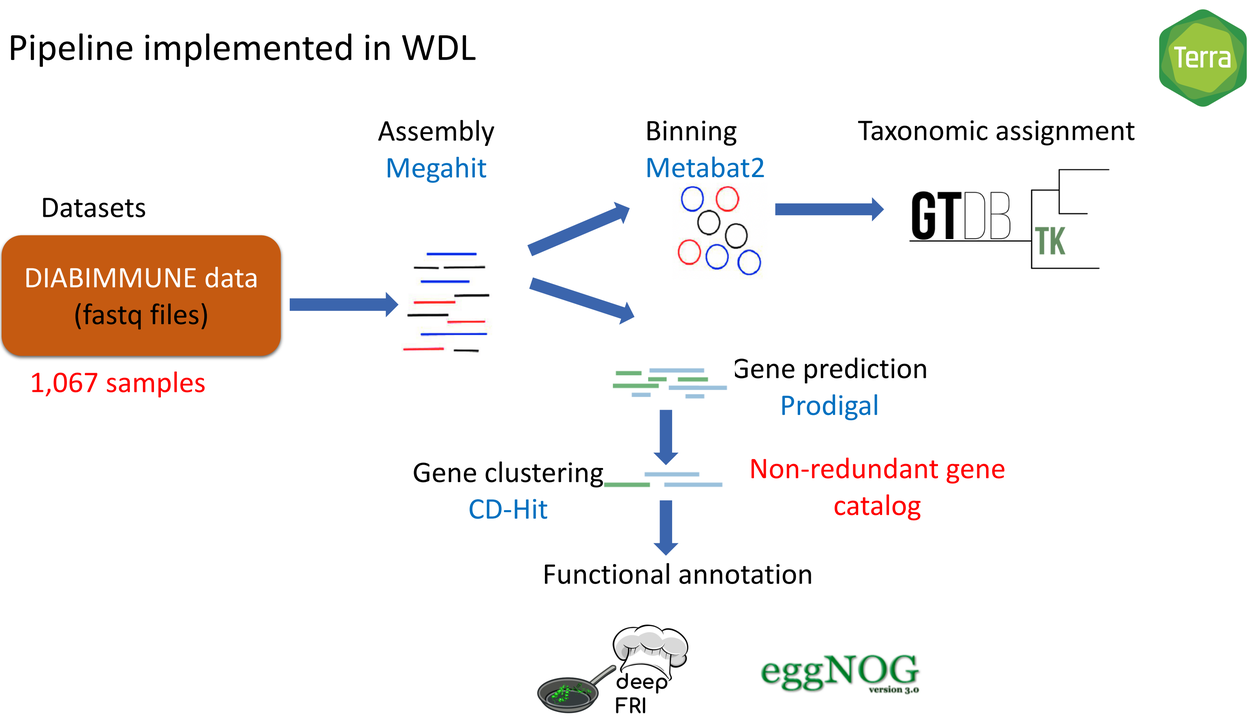
- The human gut microbiome contributes to the development and persistence of diseases such as type-1 diabetes (T1D), ulcerative colitis, obesity and many others.
- Exact mechanisms of how gut microbiota influences health remains poorly understood.
- Only 50% of microbial protein-coding genes may be functionally annotated.
- Low functional annotation coverage poses a major challenge in understanding of how the microbiome contributes to certain disease phenotypes.
- We aim to characterize the functional potential of the human gut microbiome in type-1 diabetes.

Methods overview

- Diabimmune infant gut microbiome cohort data previously collected in Finland, Estonia and Russian Karelia as case study
- Shotgun metagenome sequencing (1067 samples)
- A custom metagenomics annotation pipeline based on DeepFRI machine learning protein function annotation method
- Our method integrates *de novo* genome reconstruction, taxonomic profiling and functional annotation

Taxonomy aware function annotation pipeline

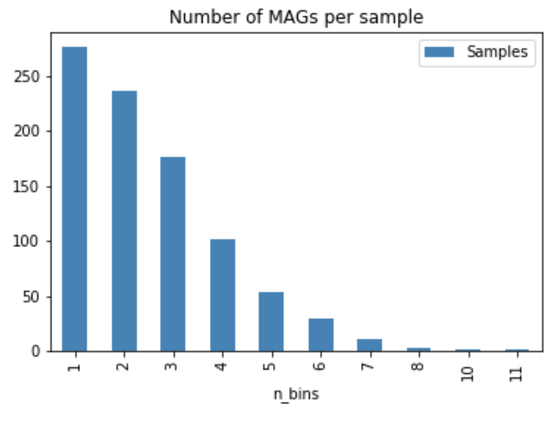
Pipeline implemented in WDL

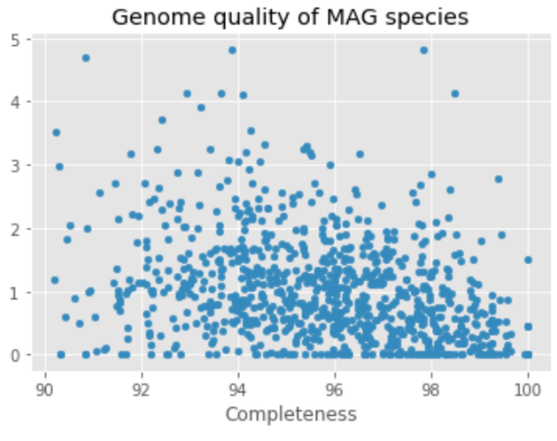


Predicted Gene ontology terms

Annotation method	Gene ontology terms predicted
DeepFRI (CNN-MF model)	13,896,275
EggNOG	280,959

Quality and completeness of metagenome assembled genomes





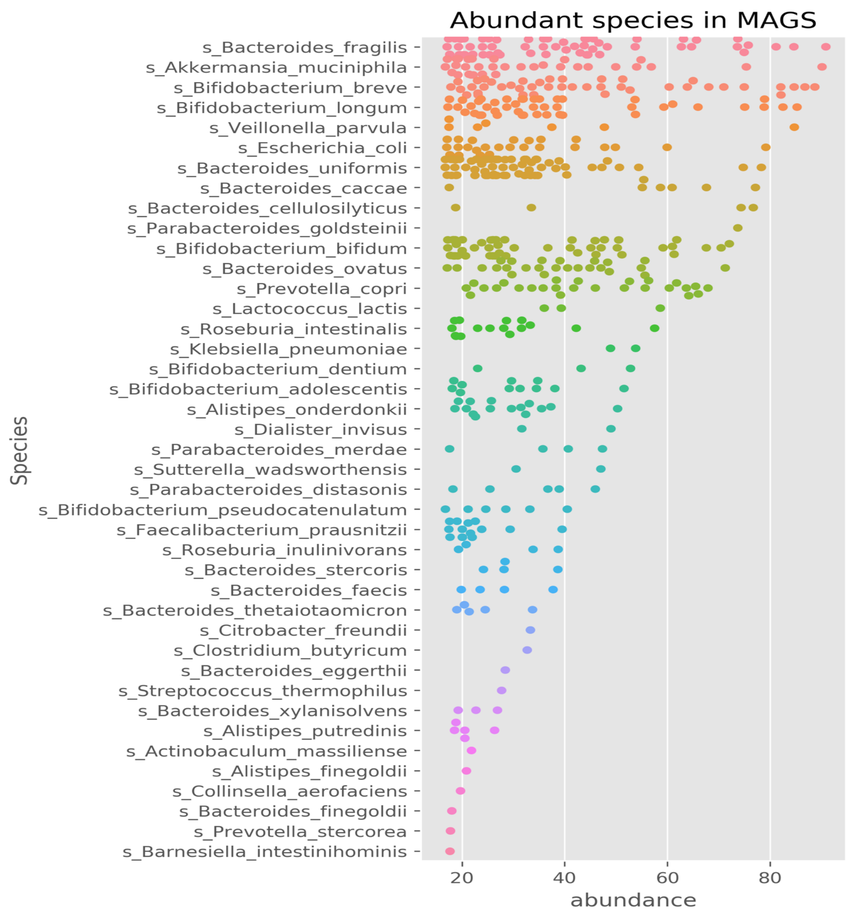
Genome quality threshold of >90% genome completeness and <5% contamination, the final genomes matching these criteria were 2,256

Results

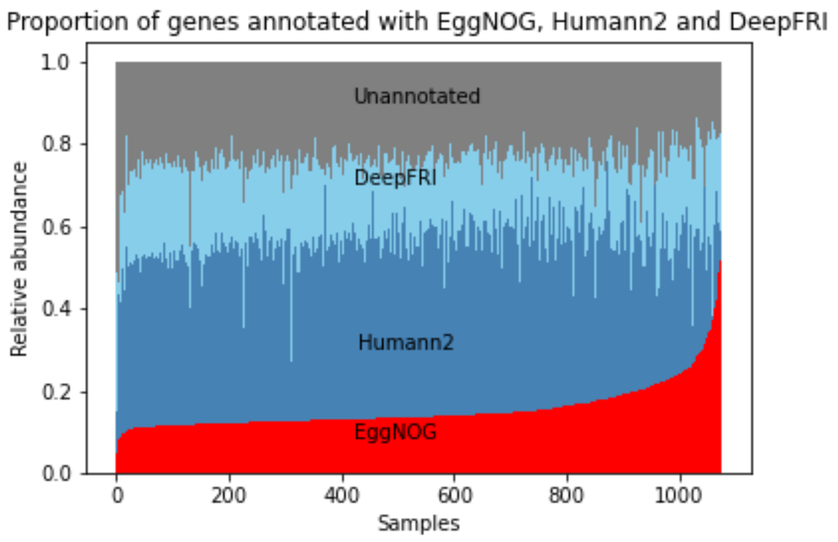
Assembly and gene prediction statistics

Assembly	Count
Contigs	17 M
MAG genes	1.7 M
NR- gene catalogue	1.9 M

Abundant species in the Diabimmune datasets



Proportion of genes annotated with EggNOG, Humann2 and DeepFRI methods



We observed an increase in annotation coverage with DeepFRI compared to Humann2 and EggNOG

Conclusions

- Result shows that DeepFRI method increases the annotation coverage
- Next step is to expand the annotations to incorporate 3D structure DeepFRI predictions

References

1) https://beta.deepfri.flatironinstitute.org/
2) Vatanen, T., Plichta, D. R., Somani, J., Münch, P. C., Timothy, D., Hall, A. B., Rudolf, S., Oakeley, E. J., Ke, X., Young, A., Haiser, H. J., Kolde, R., Yassour, M., & Luopajärvi, K. (2019). Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat Microbiol.*, *4*(3), 470–479.

Standardizing 16S rRNA gene sequencing downstream analysis for Oxford Nanopore and Ion Torrent technologies

Katarzyna Kopera [1], Dedan Githae [1], Maria Kulecka [2], Jerzy Ostrowski [2], Paweł Łabaj [1], Tomasz Kościółek [1]

[1] Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

[2] Department of Gastroenterology, Hepatology and Clinical Oncology, Centre of Postgraduate Medical Education, Warsaw, Poland

Abstract

16S rRNA marker gene sequencing is a staple technique for microbiome analyses that provides rapid and cheap bacterial identification. The most popular and well-standardized experimental technique is based on Illumina short-read sequencing. Alternative techniques are long-read Oxford Nanopore (ONT) and short-read IonTorrent platform (PGM). While both producers provide complete 16S analysis workflows, they are often not fully transparent, unadaptable, and limited to the basic methodology implemented within a given workflow. This produces a community-wide need for more in-depth workflows which at the same time will validate the applicability of the two sequencing methods in the area of 16S experiments.

We describe the powers and limitations of the two methods (PGM and ONT) by comparing them with our alternative downstream analysis created in QIIME2. The workflow was tested on 16S data generated on the Oxford Nanopore's and Thermo Fisher's sequencing machines and their 16S metagenomics kits. 16S sequencing data from 126 fecal samples from mice humanized with human stool were analysed. Different diversity metrics, taxonomy classification, and differential abundance methods were performed. For 21 common samples, Mentel test and Procrustes were made to compare the correlation of beta diversity between the two platforms.

We have managed to achieve powerful results using the approach we created, despite the limitation of information imposed by manufacturers' policies. Mentel test and Procrustes suggest good correspondence of the results from the two platforms. However, we would like to stress the further need for the entire community to cross-validate results and develop new standardized approaches for the data produced from PGM and ONT 16s sequencing solutions.

Introduction

16S rRNA sequencing on Ion Torrent and Oxford Nanopore

16S rRNA gene has been universally used for taxonomic studies of prokaryotic species. Table 1 presents these approaches as proposed by the technology provider [1, 2].

	Ion Torrent ThermoFisher SCIENTIFIC		Oxford NANOPORE Technologies	
SEQENCING METHOD	Detection of hydrogen ion release during incorporation of new nucleotides	Fast; cheap; high-quality reads	The magnitude of the electric current density across a nanopore surface	Long sequence read lengths; relatively high sequencing error rate; high throughput; portability; fast; low price
16S SEQUENCING KIT; REGION SEQUENCED	Ion 16S™ Metagenomics Kit	Hypervariable regions V2-4-8 and V3-6,7-9; forward and reverse reads; bidirectional; proprietary primer sequences	16S Barcoding Kit	full-length 16S rRNA gene
SOFTWARE	Ion 16S™ metagenomics analyses module within the Ion Reporter™ software	BLAST to either the premium curated MicroSEQ® ID or curated Greengenes or a two-step alignment	EPI2ME 16S analysis workflow	BLAST basecalled sequence against the NCBI 16S bacterial database,

Table 1

Powers and challenges of the two methods

The scarcity of tools specifically designed to work with Nanopore, and Ion Torrent sequences make it challenging to carry out a specialized microbiome analysis.

Ion Torrent [3, 4]

- studies available showed significant correlation of genera identified in Illumina and PGM
- hypervariable regions and unknown primer sequences have a big effect on a lot of aspects of data, larger than a lot of biological effects:
  1) Mixed-orientation reads will inflate diversity estimates.
  2) Reads from the same bacterium but different variable regions may be interpreted as different bacteria
  3) Some OTUs may be underrepresented and some may be counted multiple times.
  4) The data becomes impossible to use/reuse when looking for a specific ASV.

Nanopore [5]

- capturing the entire 16S rRNA gene improved classification at the genus and family levels,
- bacterial species identification is highly error-prone,
- outside of EPI2ME analysis:
  1) applying ONT to microbial diversity uses a similar approach to previous studies, mostly Illumina-based
  2) Limited quality sequences should sometimes be a constraint to apply existing tools designed for other technologies.
  3) The final output from EPI2ME is usually not compatible with tools for analyses such as diversity and taxonomic differential abundance.

Humanization experiment

16S rRNA marker gene sequenced on PGM platform (123 samples) and ONT platform (23 samples) was done in experiment in which NUDE and NSG mice were humanized with a single human stool sample over the course of three months. 123 samples were sequenced using PGM and 23 using ONT devices and chemistry.

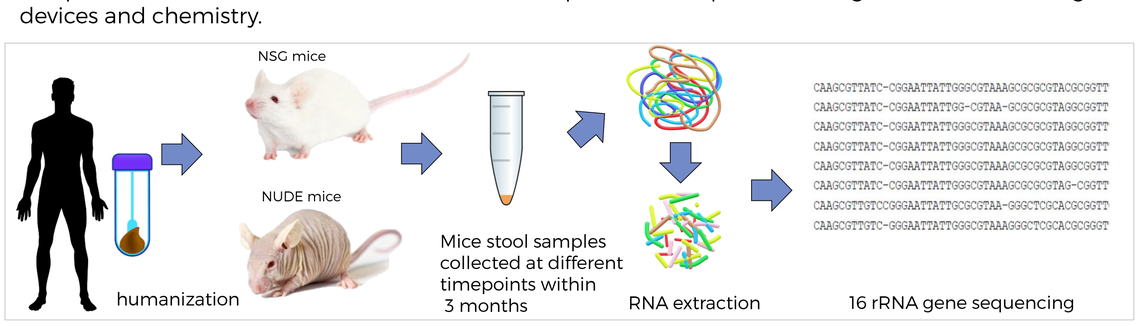


Figure 1

QIIME2 downstream analysis worflow

We have created an alternative downstream analysis workflow in Qiime2 [6] tailored to PGM and ONT prerequisites. Some of the adjustments and settings are presented in the Figure 2.

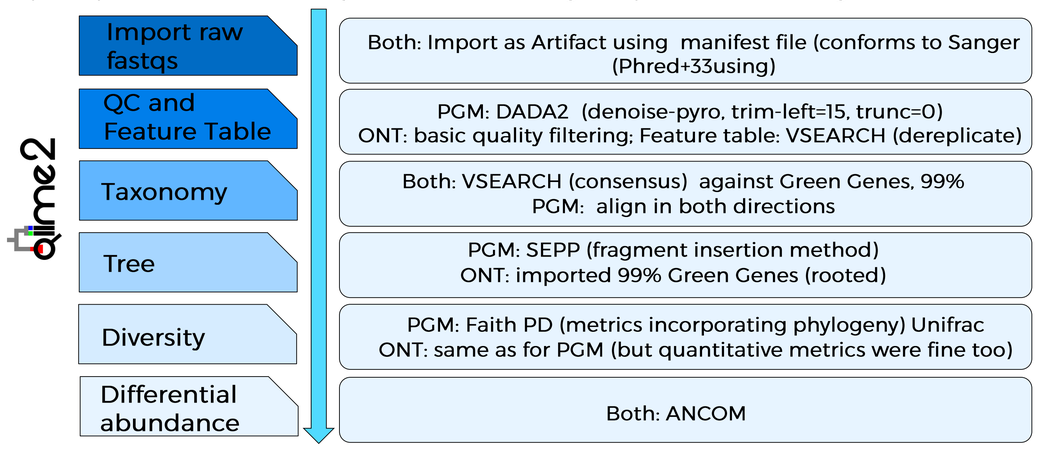


Figure 2

Results

Quality control

The higher number of sequenced samples on the PGM platform (126 vs. 23 in ONT) translates directly into the number of detected features in the two sample sets. However, the alpha-difference curves indicate that increasing the depth above the values in the table does not cause new biodiversity to appear (alpha-diversity curves are saturated with the values indicated in the table). At the same time, such sampling depths make it possible to preserve all collected samples.

FEATURE TABLE SUMMARIES FOR ONT AND PGM

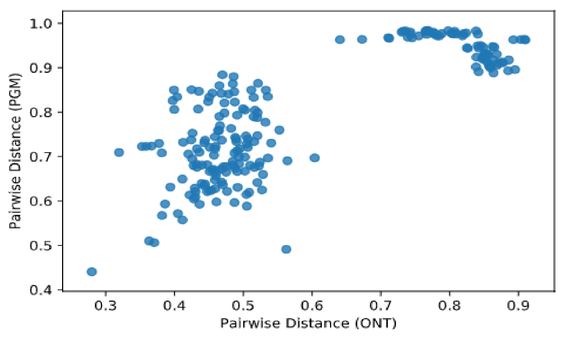
	Ion Torrent	Oxford Nanopore
# Samples	126 (123 mice, 1 human, 2 mock)	23 (22 mice, 1 human, 2 mock)
Unique features	9,877	3,543
Total features	17,908,604	1,130,914
Features per sample (median)	129,097	41,628
Reads per feature (median)	83	11
Features per sample at even sampling depth	45,000	9,000
Features retained at even sampling	5,850,000 (32.67%)	212,727 (18,8%)

Table 2

Ion Torrent and Oxford Nanopore performance comparison

CORRELATION OF BETA DIVERSITY BETWEEN THE TWO PLATFORMS

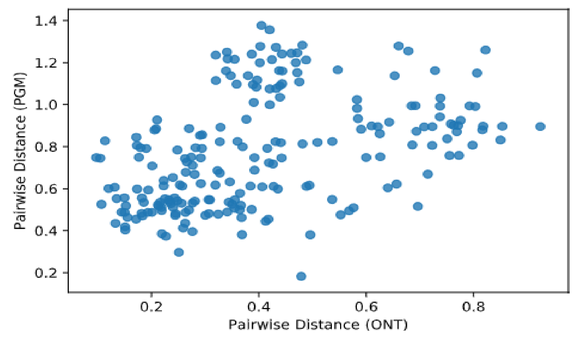
Mantel Test for unweighted Unifrac



Method	Spearman
Sample size	21
Permutations	999
Alternative hypothesis	two-sided
Spearman rho	0.70408
p-value	0.001

Figure 3

Mantel Test for weighted Unifrac



Method	Spearman
Sample size	21
Permutations	999
Alternative hypothesis	two-sided
Spearman rho	0.517567
p-value	0.003

Figure 4

The Mantel test is a statistical test of the correlation between two matrices. The matrices must be of the same dimension, so the sample size was subsampled to 21 for the samples present in both sequencing methods. Test showed good positive correlation (rho = 0.70408, p-value 0.001) of beta diversity measured by unweighted UniFrac and moderate positive correlation (rho = 0.517567, p-value 0.003) for weighted UniFrac.

Procrustes Visualization

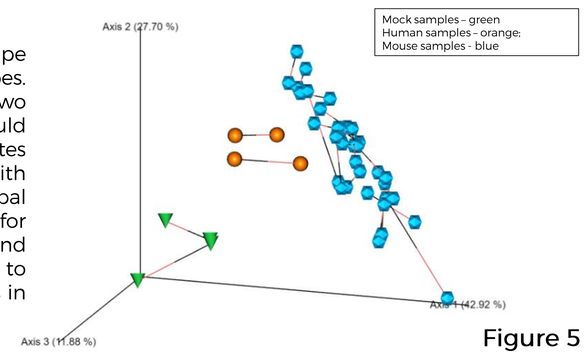


Figure 5

Procrustes analysis is a form of statistical shape analysis used to analyse the distribution of a set of shapes. It can be used in microbial biloogy to compare two matrices for example to determine whether we would derive the same beta diversity conclusions. Procrustes analysis takes as input two coordinate matrices with corresponding points (generated by running principal coordinate analysis on a distance generated from for example weighted UniFrac) and transforming the second coordinate set by rotating, scaling, and translating it to minimize the distances between corresponding points in the two shapes.

Conclusions

1. There is a shortage of sophisticated bioinformatic tools for ONT and PGM at the current level of methodological advancement.
2. In the case of the Ion Torrent, efforts have focused on strategies to combine results from multiple variable regions and mixed orientations while for Nanopore it is designing tools for base-calling, demultiplexing and taxonomic assignment.
3. Qiime2, can be adapted to facilitate the methodological implications specific to PGM and ONT with a robust alternative to alignment, taxonomic analysis and phylogenetic analysis, such as diversity indicators, has been developed.
4. Analyzing sequencing data using a unified QIIME 2 framework, we show that Ion Torrent and Nanopore results are comparable with each other

References

1. https://www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/Ion-16S-Metagenomics-Kit-Software-Application-Note.pdf
2. https://nanoporetech.com/nanopore-sequencing-data-analysis
3. F. Fouhy, , A.G. Clooney, C. Stanton et al. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol. 2016;16:123 16.
4. J. Barb, A. Oler, HS. Kim, et al. Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. PLoS One. 2016;11(2):e0148047.
5. A. Santos, R. van Aerle, L. Barrientos, J. Martinez-Urtaza, Computational methods for 16S metabarcoding studies using Nanopore sequencing data, Comput. Struct. Biotechnol. J. 2020;18:296-305.
6. M. Estaki et al. QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. Curr Protoc Bioinformatics. 2020;70:e100.