In silico exploration of quadruplex structures

Joanna Miśkiewicz (1), Mariusz Popenda (2), Joanna Sarzyńska (2), Tomasz Żok (1,3), Paulina Pielacińska (1), Natalia Kraszewska (1), Marta Szachniuk (1,2)

Corresponding authors: jmiskiewicz@cs.put.poznan.pl, mszachniuk@cs.put.poznan.pl

(1) Institute of Computing Science, Poznan University of Technology, Poland
(2) Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland
(3) Poznan Supercomputing and Networking Center, Poland

Introduction

G-quadruplexes (G4s) are structural motifs that appear in the DNA and RNA of many organisms. They are built of tetrads stacked upon one another. Four guanines in a pseudo-planar arrangement, connected by hydrogen bonds, form a tetrad. G4s are involved in many biological processes, for example, transcription regulation and genome stabilization. Thus, they constitute an interesting target of novel therapeutic designs.

Quadruplex structures

I GGGATGGGACACAGXGGACGGG

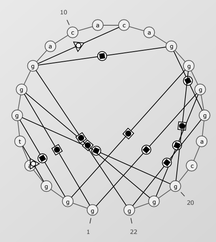
Motif: $G_{x1}N_{y1}G_{x2}N_{y2}G_{x3}N_{y3}G_{x4}$

$x \geq 2$, $y \geq 0$

Quadruplex structures

II

((...)[(.....)[])(..]])
([...{)(.....(][)..}])



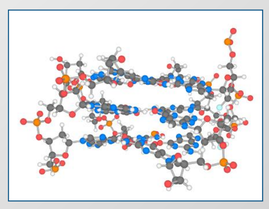
n4-helix with 3 tetrads
Mh* quadruplex with 3 tetrads
A.DG1 A.DG16 A.DG21 A.DG7 cWH-cWH-cWH-cWH N-
A.DG2 A.DG6 A.DG20 A.GF2/15 cWH-cWH-cWH-cWH N+
A.DG8 A.GFL14 A.DG17 A.DG22 cWH-cWH-cWH-cWH O+

This work was supported by the Polish National Science Centre, grant number 2019/35/B/ST6/03074.

Quadruplex structures

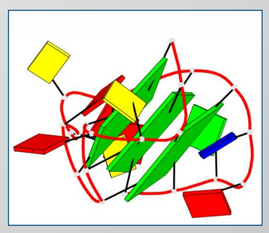
III

Tetrad	ONZ class	Planarity
1	N-	0.17
2	N+	0.34
3	O+	0.10



Base pair	Twist	Rise
1 - 2	23.43	3.45
2 - 3	56.18	6.79

Analysed structure: 6TC8



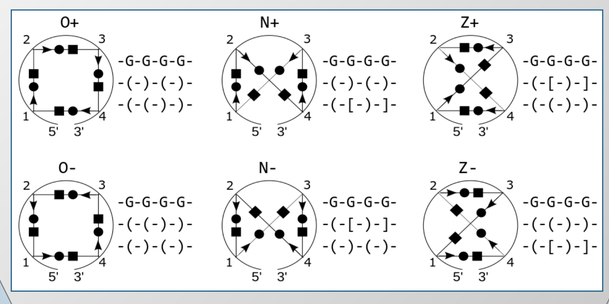
ONZ classification

The secondary structure of tetrad T can be represented as cyclic graph G = (V,E), where $|V|=|E|=4$, each $v \in V$ represents one nucleotide from the tetrad, every $e \in E$ corresponds to a hydrogen-bonding interaction between respective nucleotides. If we placed the Vertices of G at equal distances on a circle clockwise, in the order imposed by the sequence, we'd see that graph takes the shape of a square (O-shaped), a bow tie (N-shaped), or an hourglass (Z-shaped). This observation made us distinguish 3 groups of tetrads and define their ONZ taxonomy:
Let T denote a tetrad build of N_1, N_2, N_3, N_4 nucleotides.
We define ONZ classes:

- Class O if $T = \{(N_1,N_2), (N_2,N_3), (N_3,N_4), (N_4,N_1)\}$,
- Class N if $T = \{(N_1,N_2), (N_2,N_4), (N_4,N_3), (N_3,N_1)\}$,
- Class Z if $T = \{(N_1,N_3), (N_3,N_2), (N_2,N_4), (N_4,N_1)\}$.

ONZ classification

We can annotate the tetrad according to the interaction arrangement. If the first nucleobase binds with the next one along the Watson-Crick edge, the tetrad is tagged positive (+), otherwise, it is tagged negative (-).



Thus, every class in ONZ is divided into two subcategories: O+, O-, N+, N-, Z+, Z-.

ElTetrado

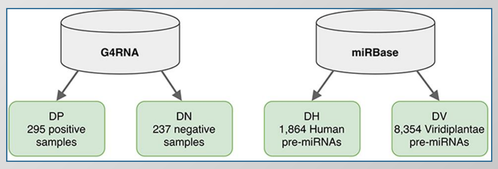
ElTetrado identifies and describes tetrads and quadruplexes in the 3D structures of nucleic acids, by searching for G-based and non-G-based motifs. It allocates tetrads and quadruplexes to ONZ classes according to their 2D structure topology, calculates strand direction, planarity deviation, rise and twist parameters. The program also outputs the graphical representation of the 2D structure (top-down arc diagram) and its dot-bracket encoding in a two-line format—both designed specially to handle tetrads and quadruplexes.

	Op	Oa	Oh	Na	Nh	Mp	Mh	Total
+	105	-	12	11	3	8	8	147
-	2	-	-	-	-	-	-	2
*	13	40	23	27	4	5	6	118
Total	120	40	35	38	7	13	14	267

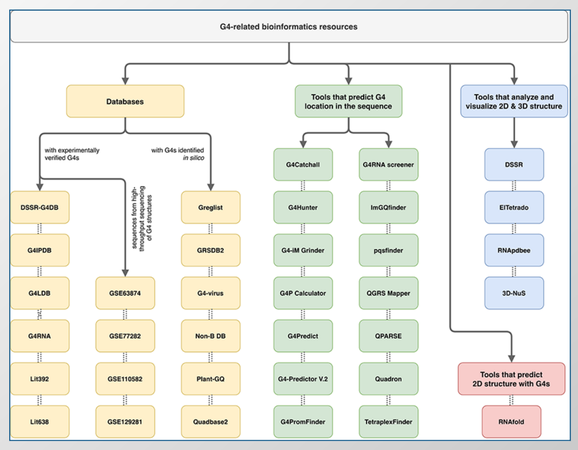
ONZM class coverage by unimolecular quadruplexes

Quadruplex resources

With the growing interest in quadruplexes, computer programs for their analysis began to appear. Most of them rely solely on a sequence and parse it to find a predefined G4 motif. This goes hand in hand with creation of G4-related databases that primarily collect information about sequences with the ability to form quadruplexes. We distinguished the following subsets of resources: databases, tools to predict putative quadruplex sequences, tools to predict secondary structure with quadruplex motifs, and tools to analyze and visualize quadruplex structures.

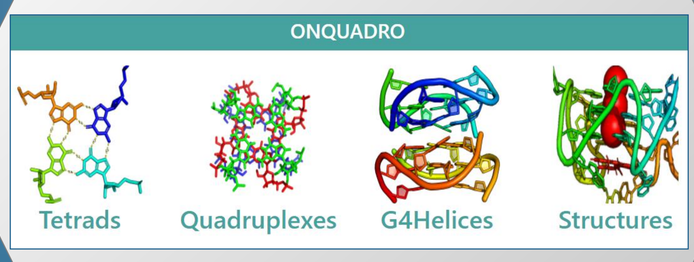


Quadruplex resources



ONQUADRO

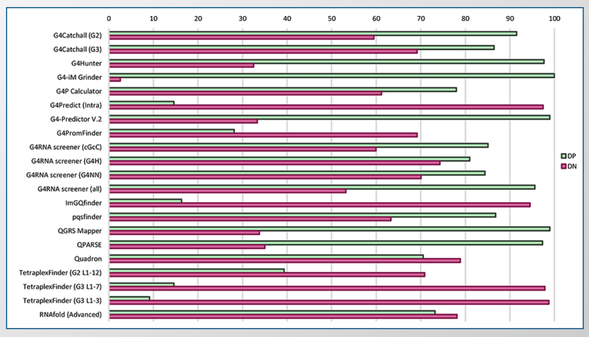
ONQUADRO database collects tetrads and quadruplexes found in PDB-deposited structures of nucleic acids.



It stores their sequences, 2D and 3D structures, and motif-specific description including planarity, rise and twist parameters, ONZ classification, dot-bracket encoding, arc diagrams, graphical views of 2D and 3D structure, etc.

Quadruplex resources

Analysis of G4-dedicated programs



Coverage of DP and DN datasets with correct predictions: positive in DP and negative in DN [%]. The best results: G4RNA screener, G4Catchall, RNAfold.

References

[**ONZ**] Popenda et al. Topology-based classification of tetrads and quadruplex structures. Bioinformatics 2020, 36:1129–34.

[**ElTetrado**] Zok et al. ElTetrado: a tool for identification and classification of tetrads and quadruplexes. BMC Bioinformatics 2020, 21:40.

[**Quadruplex resources**] Miskiewicz et al. How bioinformatics resources work with G4 RNAs. Briefings in Bioinformatics 2020, ahead of print.

[**ONQUADRO**] http://onquadro.cs.put.poznan.pl/

RNA junctions from a 3D structure perspective

J. Wiedemann*[1], M. Antczak*[1,2], J. Kaczor [1], M. Milostan[1,3], T. Zok[1], M. Szachniuk[1,2]

1. Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
2. Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland
3. Poznan Supercomputing and Networking Center, Jana Pawla II 10, 61-131 Poznan, Poland

Introduction

Computational methods for the 3D structure prediction allow for prototyping the shape of RNA, yet some of its fragments require more attention and manual or semi-automatic adjustment. Among them are multibranched loops (n-way junctions) - hard to predict structural motifs that significantly impact the structure of the whole molecule.

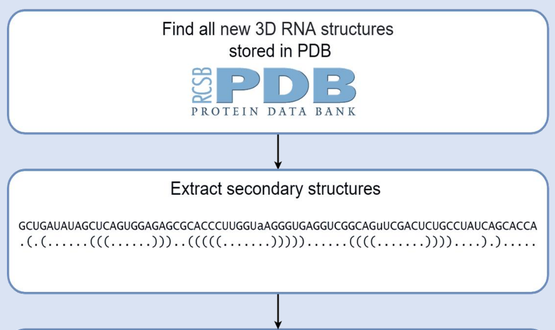
In this work, we created the RNAloops database that collects structural data of RNA n-way junctions. The novelty in our tool is the loop description that contains, i.a., a set of angles (Euler and planar) to determine spatial relationship between outgoing helices.

Data analysis performed with the RNAloops contents showed that every eight RNA contain n-way junctions; in some structures we found even >100 of them per molecule. We believe that data collected in RNAloops can be used to improve the accuracy of *in silico* modeling of RNA 3D structures.

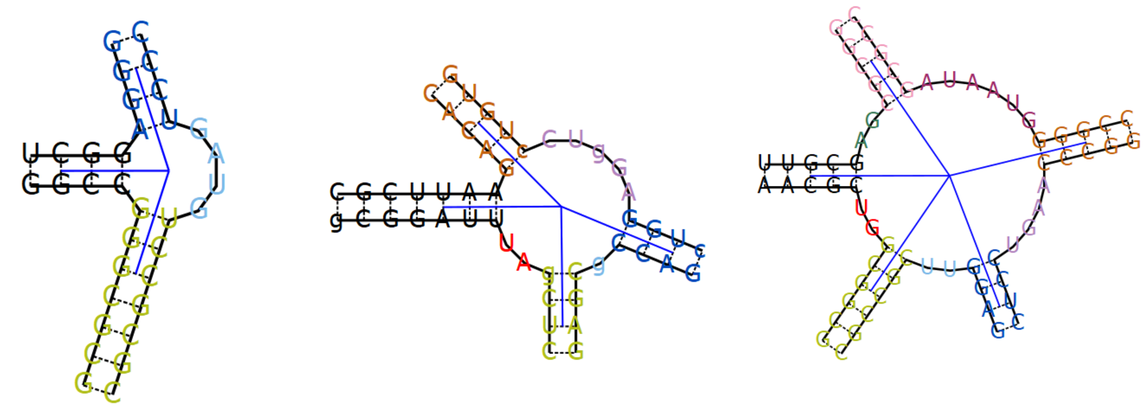
About RNAloops

RNAloops stores the information about n-way junctions found in experimentally determined RNA 3D structures deposited in the Protein Data Bank. The stored data include sequence, secondary structure, tertiary structure, one planar, and three Euler angles that describe the relationship between stems coming out of the loop.

The database is automatically updated every Sunday. Newly deposited RNA 3D structures are automatically downloaded from the RCSB PDB repository. For every downloaded RNA, its secondary structure is extracted using RNApdbee algorithm and encoded in extended dot-bracket notation. All n-way junctions are identified in each RNA, based on its secondary structure processing. Planar angle and Euler angles are computed for every pair of outgoing helices and attached to the description of the multiloop.



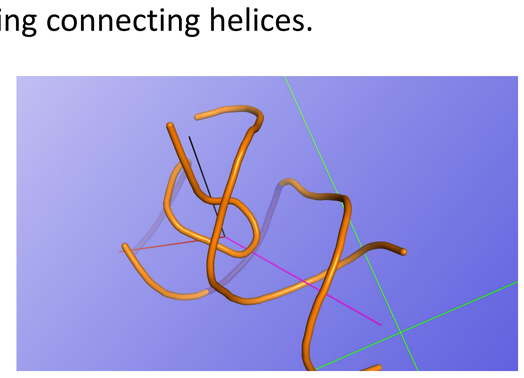
Example n-way junctions in RNAloops



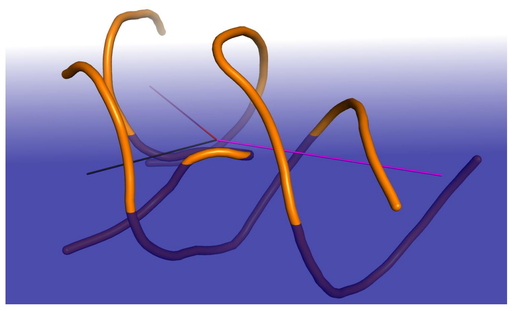
3-way junction | 4-way junction | 5-way junction

Angular representation

- **Euler angles** (X, Y, Z) – three angles describing the rotation around axes in 3D that are required to align two neighbouring connecting helices.

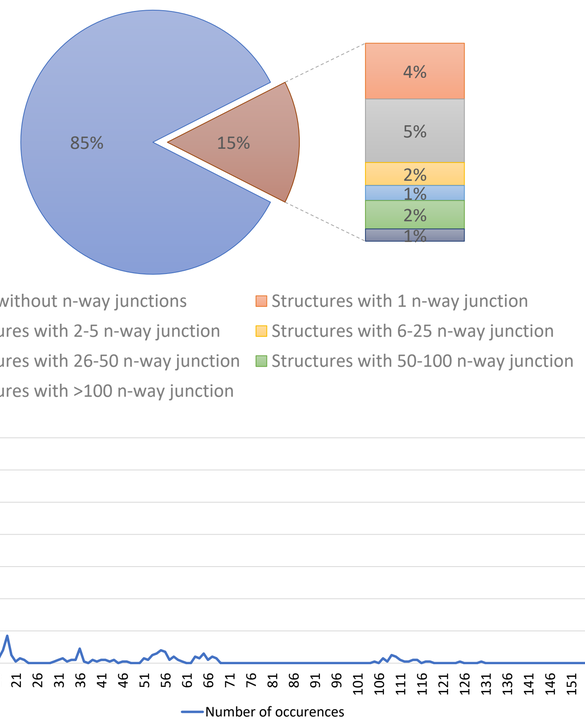


- **Planar angle** – a single angle describing angle between two neighbouring connecting helices. Figure below visualizes example of the planar angle.

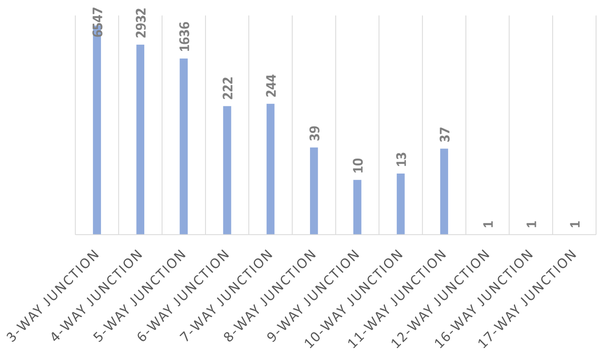


Results

The RNAloops database stores information of 11 984 junctions from 510 RNA structures, as of 3-Nov-2020. It is the result of processing 2 418 RNA structures from the RCSB PDB repository.



We identified junctions with 3-17 outgoing helices. Out of these 11 984 junctions, over 50% were 3-way junctions (6 547), about 30% were 4-way junctions, 13% - 5-way junctions.



The project has been supported by the National Science Centre, Poland [2016/23/B/ST6/03931 and 2019/35/B/ST6/03074], national grant for young researchers Mloda Kadra [0311/SBAD/0705], and the statutory funds of Poznan University of Technology, Poland.

POZNAN UNIVERSITY OF TECHNOLOGY

The influence of structure size on similarity metrics values.

Piotr Kłosowicz[2], Tomasz Żok[1]

[1] Poznan University of Technology, Poland

[2] Adam Mickiewicz University Poznan, Poland

Introduction

A comparison of tertiary structure between several molecules is very important in order to understand function of specific RNA molecules and interactions between them. With it we can find motifs, that are crucial to identify and recognize to the role of newly-found molecules or extend our knowledge about known RNA chains, which are still not fully studied. But in order to make a comparison, we need a measure, that is reliable and can be used to a wide variety of molecules, with as low number of limitations as possible.

Our inspiration and objectives

In order to overcome this problem, a group of scientists from University of North Carolina took a closer look at root-mean-square deviation (RMSD). Their work was described in article “On the significance of an RNA tertiary structure prediction”[1], written in 2010. They have discovered, that RMSD is highly dependant on the size of the molecule and in order to make it more reliable, they proposed usage of prediction significance (P-value). This value allows to evaluate, if the given prediction is better than one expected by chance for molecule of that size.

But there are more measures, that can be used to compare tertiary structures than just RMSD. In our research we are concentrating on mean of the circular quantities (MCQ), which is an approach used in the MCQ4Structures tool. It is described in “MCQ4Structures to compute similarity of molecule structures”[2] article, written in 2014 by the group of scientists from Poznan University of Technology. We have decided to take a closer look on this measurement in order to make it size-independant.

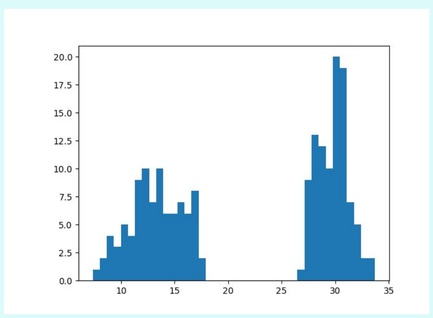
Methodology

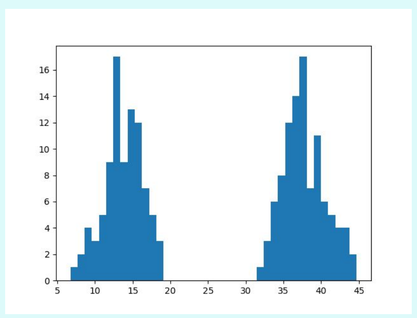
In order to achieve our objectives, we decided to use the same molecules, that were used in research on RMSD, so our work can be more comparable to one done by scientists from University of North Carolina. Than we had to create a set of decoy structures, which would later be used in comparison with the original molecule. To make those we have used RNAsubopt tool from ViennaRNA package, which gave us a set of 2D predictions. Then we took received structures and the original ones, retrieved from PDB site, and loaded them into RNAComposer tool, which is used in prediction of 3D RNA structures. As a result of this, we have received a set of 3D configuration, which we could compare with each other. In order to do this, we have used previously mentioned MCQ4Structures tool.

Results

At this moment we have created histograms for distances between our decoy structures. Then we tested our score with normality tests, but in opposition to our assumptions, the data appeared to be not normally distributed. What’s even more interesting, for some molecules distribution appeared to be bimodal. We will have to take a closer look at this case.

We still have much more to do in this matter. Our research is far from the end, but it’s going in the right direction and we hope to present final answer to presented problem.





Workflow

Obtaining 2D structures of original molecules

GGAGUUCACCGAGGCCACGCGGAG
UACGAUCGAGGGUACAGUGAAUU

RNApdbee:

..(((((((...((((.((((.....))..))..))).).)))))))

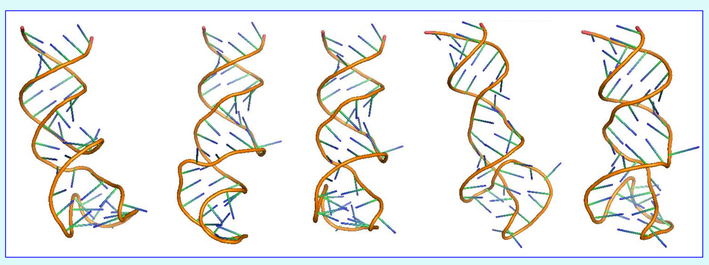
RNAsubopt:

..(((((((.(..(((.((((.....))..))..))).).)))))))
..(((((((....(((.((((.....))..))..)))...)))))))
...((((((.(..(((.((((.....))..))..))).).)))))).
..(((((((.(..(((.(.(((.......))).)))).).)))))))

↓

Creating a set of 2D decoy structures

↓



↓

Creating a 3D models

↓

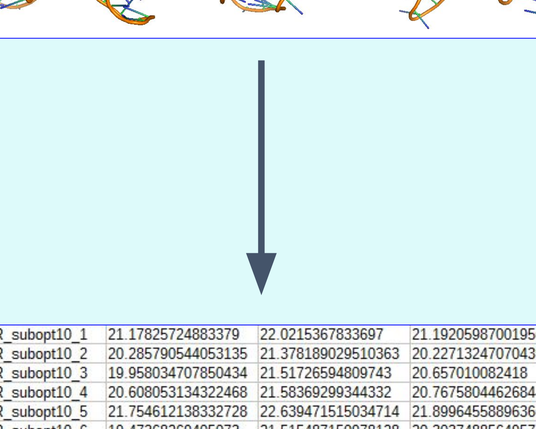
13	1XJR_subopt10_1	21.17825724883379	22.0215367833697	21.19205987001956	20.178
14	1XJR_subopt10_2	20.285790544053135	21.378189029510363	20.227132470704397	19.489
15	1XJR_subopt10_3	19.958034707850434	21.51726594809743	20.657010082418	19.951
16	1XJR_subopt10_4	20.608053134322468	21.58369299344332	20.76758044626844	20.544
17	1XJR_subopt10_5	21.754612138332728	22.639471515034714	21.899645588963665	21.955
18	1XJR_subopt10_6	19.47368269405073	21.515487150978128	20.303748856495712	19.765
19	1XJR_subopt10_7	20.562551584740863	21.285180993736148	21.170325317029157	20.917
20	1XJR_subopt10_8	20.90369783609336	22.06117832390136	20.920245866605626	20.986

↓

Measuring distances between structures

↓

Creating a histogram of given distances and comparing it to the normal distribution



References

[1] Hajdin C, Ding F, Dokholyan N, Weeks K. (2010). On the significance of an RNA tertiary structure prediction. RNA (New York, N.Y.). 16. 1340-9. 10.1261/rna.1837410.

[2] Zok T, Popenda M, Szachniuk M. (2013). MCQ4Structures to compute similarity of molecule structures. Central European Journal of Operations Research. 22. 10.1007/s10100-013-0296-5.

RNAlign2D – a rapid tool for combined RNA structure and sequence-based alignment using pseudo-amino acid substitution matrix

Tomasz Woźniak[1], Małgorzata Sajek[2], Jadwiga Jaruzelska[1], Marcin Sajek[1]

INSTITUTE OF HUMAN GENETICS
POLISH ACADEMY OF SCIENCES

1. Institute of Human Genetics, Polish Academy of Sciences, Poznań, Poland
2. Department of Human Molecular Genetics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

INTRODUCTION

The functions of RNA molecules are mainly determined by their secondary structures. These functions can also be predicted using bioinformatic tools that enable the alignment of multiple RNAs to determine functional domains and/or classify RNA molecules into RNA families. Here, we introduce an extremely fast Python-based tool called RNAlign2D. This tool is dedicated to multiple alignment of RNA molecules with known secondary structures. It converts RNA sequences to pseudo-amino acid sequences that incorporate structural information and uses a customizable scoring matrix to align these RNA molecules using the multiple protein sequence alignment tool MUSCLE. This approach can be customized for virtually all protein aligners. RNAlign2D is freely available from https://github.com/tomaszwozniakihg/rnalign2d .

MATERIALS AND METHODS

RNAlign2D is a command line tool written as a Python script that works in UNIX-based operation systems. To compare RNAlign2D with other tools that can use fixed 2D structure for multiple RNA alignment LocARNA and CARNA, we used 2 available benchmark datasets: BraliBase 2.1 and RNAStralign.In the next step, the sum-of-pairs-scores (SPSs) and positive predictive value (PPVs) were calculated for each alignment. Alignment time was also measured for subset of datasets from RNAStralign benchmark.

RESULTS

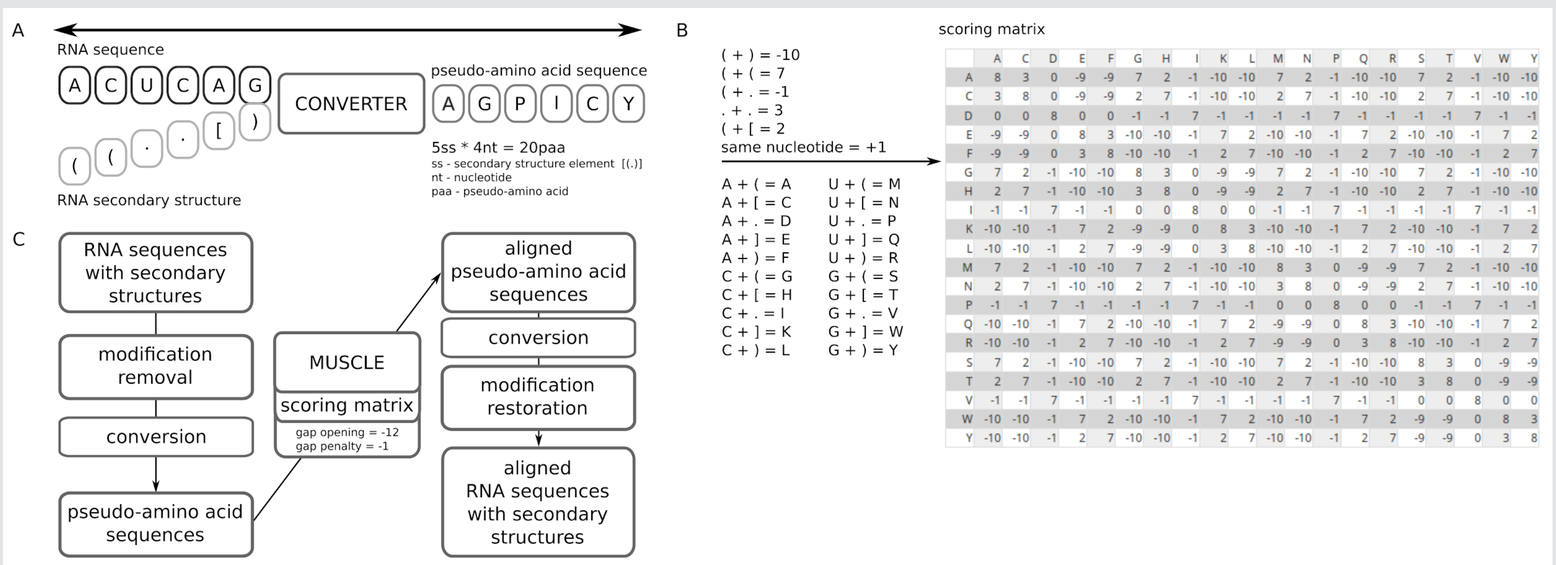


Figure 1. Schematic representation of the RNAlign2D workflow. **A**. Basic concept of RNA sequence-structure conversion to a pseudo-amino acid sequence. **B**. Conversion of 20 RNA sequence-structure elements to pseudo-amino acids and their scores (left) and the default scoring matrix (right). **C**. Block diagram of the RNAlign2D workflow.

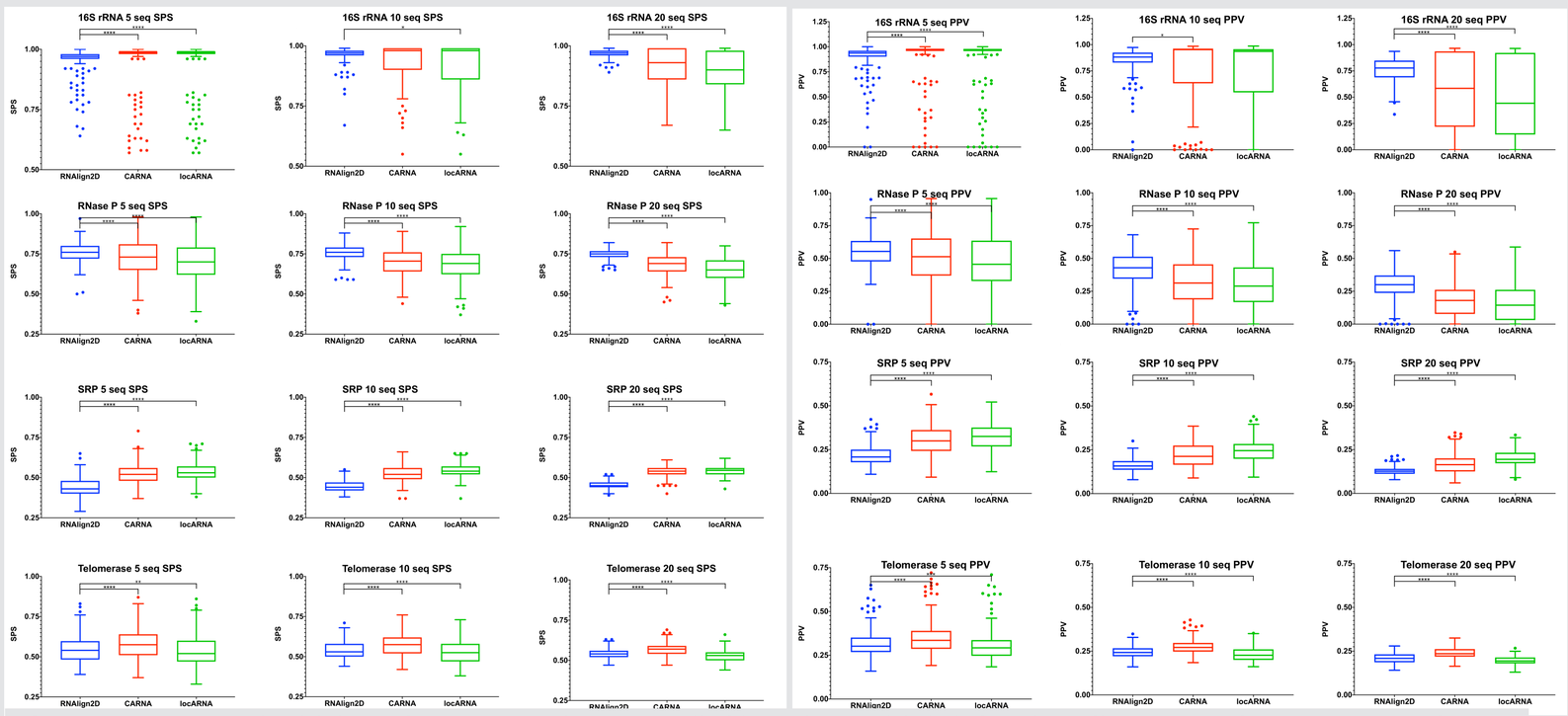


Figure 2. Box and whisker plots comparing SPSs (left) and PPV (right) for the alignment of 200 groups of 5, 10 and 20 homologous sequences from the entire RNAStralign benchmark dataset with RNAlign2D, CARNA and LocARNA. **** p-value < 0.0001; *** p-value < 0.001; ** p-value < 0.01; * p-value < 0.05.

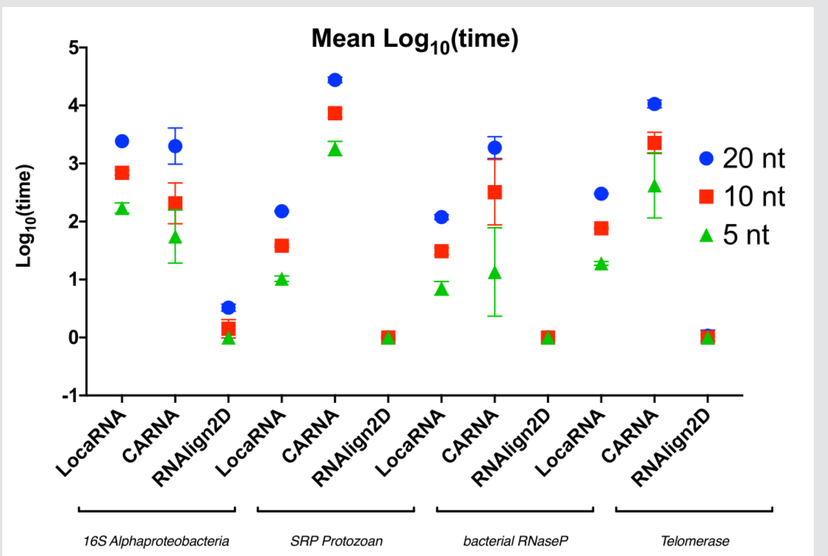


Figure 3. Comparison of alignment performance times for RNAlign2D, CARNA and LocARNA presented as a graph with standard errors indicated.

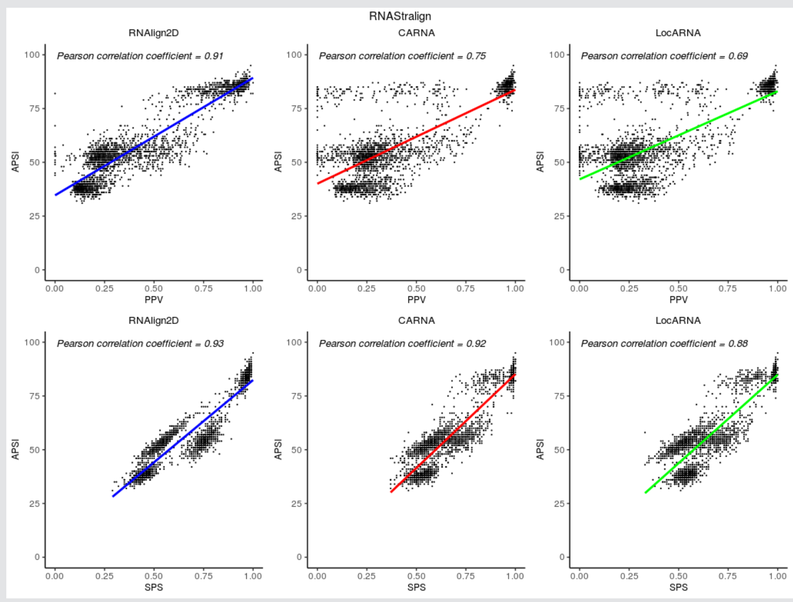
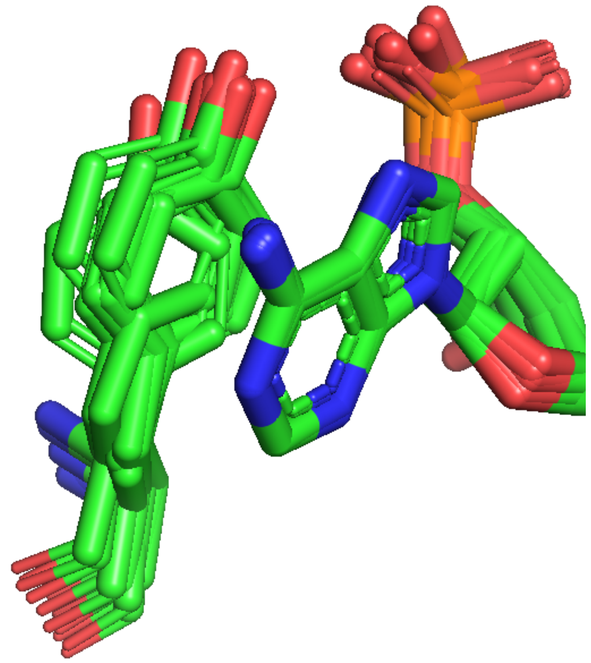


Figure 4. Pearson correlation plots of APSI vs. PPV (upper panel) and APSI vs. SPS (lower panel) for alignment of the RNAStralign benchmark dataset with RNAlign2D (left panel), CARNA (middle panel) and LocARNA (right panel). The correlation coefficients are shown at the top of each plot.

Mining biomacromolecular interactions with the BioShell package

Justyna Krys, Monika Pikuzinska, Dominik Gront
Faculty of Chemistry,University of Warsaw



INTRODUCTION

Molecular modelling is a technique commonly used in deciphering life on a molecular level. Outputs from these simulations usually comprise 3D structures and their energies, evaluated in a given force field. These tools however rarely explain how the biomolecules interact.

Thirteen close homologs from CYP family which bind NAP molecule were analysed as an example. Only interactions between ligand and a protein were taken into account.

STACKING INTERACTIONS

In the table there are all stacking interactions between NAP and a protein. It turns out that in all cases it is the same amino acid TYR604 (in 3QFC it is TYR 607). Geometry is presented on a figure. Thirteen proteins was superimposed to have NAP in the same place so the differences between homologs can be easily seen.

VAN DER WAALS INTERACTIONS

Multiple sequence alignment was created using MAFT program and coloured by Van der Waals interactions. Color scale was used to show the distances between interacting residues.

HYDROGEN BOND INTERACTIONS

Hydrogens was added to the structures with reduce[1] program. Residues that creates hydrogen bonds with NAP was colored orange on multiple sequence alignment.

PDB code	1st residue	2nd residue	r	angle	xy	z
1JA0	NAP852 B	Y604 B	3.751	166.402	1.519	3.430
1JA0	NAP752 A	Y604 A	3.698	167.154	0.997	3.561
1JA1	NAP1852 B	Y604 B	3.884	13.503	1.600	3.539
1JA1	NAP1752 A	Y604 A	3.715	7.488	1.235	3.503
3ES9	NAP753 A	Y604 A	4.269	55.074	1.139	4.114
3ES9	NAP753 B	Y604 B	4.196	32.883	0.896	4.100
3OJX	NAP753 A	Y604 A	3.930	12.828	1.702	3.542
1J9Z	NAP752 A	Y604 A	3.644	7.551	0.828	3.549
1J9Z	NAP852 B	Y604 B	3.711	6.020	1.296	3.477
6NJR	NAP703 A	Y604 A	3.751	162.636	0.884	3.645
6NJR	NAP703 B	Y604 B	3.961	164.295	1.076	3.812
3QFC	NAP753 A	Y607 A	3.865	10.405	1.604	3.516
3QFC	NAP753 B	Y607 B	3.725	12.987	1.099	3.559

Funding: This research was funded by the National Science Centre (Poland) Grant No. 2018/29/B/ST6/01989.

1. Word, et al.(1999) "Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation" J. Mol. Biol. 285, 1735-1747.

Macnar, J.M.; Szulc, N.A.; Kryś, J.D.; Badaczewska-Dawid, A.E.; Gront, D. BioShell 3.0: Library for Processing Structural Biology Data. Biomolecules 2020, 10, 461.

grahics prepared with VisuaLife Package https://visualife.readthedocs.io

fingeRNAt - a novel tool for high-throughput analysis of nucleic acid - ligand interactions

Natalia A. Szulc[1], Zuzanna Mackiewicz[1], Janusz M. Bujnicki[1,2,*], and Filip Stefaniak[1,*]

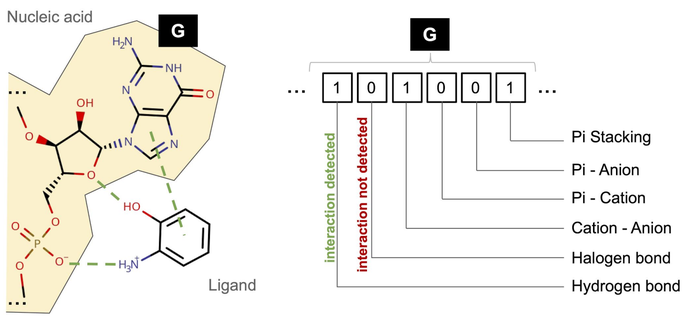
1. International Institute of Molecular and Cell Biology in Warsaw, Poland; 2. Adam Mickiewicz University, Poznań, Poland;
* - corresponding author's email: {iamb, fstefaniak}@genesilico.pl (J.M.B. and F.S.)

Abstract

Nucleic acids are becoming increasingly attractive targets for potential drugs. Since most targets of small molecule drugs are proteins, the portfolio of nucleic acids-oriented bioinformatics tools is limited. Here we present **fingeRNAt** - a novel and open-source software for **calculation of Structural Interactions Fingerprints (SIFs) for nucleic acid - ligand complexes**. SIFs translate information about 3D interactions in a target-ligand complex into a string, where the respective bit in the fingerprint is e.g. set to 1 in case of detecting particular interaction, and to 0 otherwise. By using SIFs, the interactions are represented in a unified fashion, thus allowing for easy analysis and comparison, as they provide a full picture of all interactions within the complex.

Background

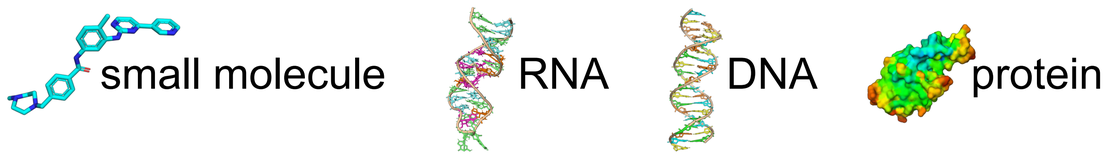
- Many nucleic acids are disease-associated with ability to adapt a tertiary structure hence constituting promising targets for drugs.
- Structural Interactions Fingerprints (SIFs) represent interactions within a complex in a form of a binary or a hologram string, a convenient input to computational analyses.



- No freely available tool to calculate SIFs for nucleic acid - ligand complexes.

Overview

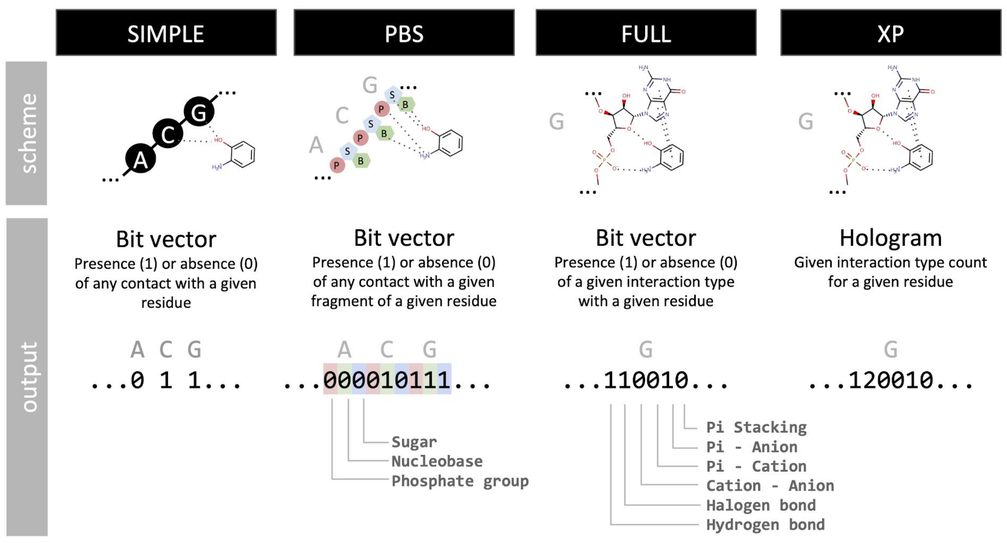
- fingeRNAt is a Python 3.x program which calculates SIFs in complexes of RNA/DNA and:



- **Input/Output**

  Requires (i) RNA/DNA structure in pdb/mol2 format and (ii) ligands' structures in sdf format.
  The output is a SIF calculated for each complex saved to separate row of a tab-separated file.

- **SIFs types**



fingeRNAt applications

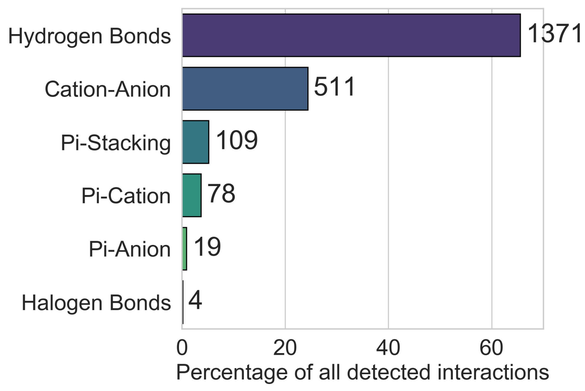
What are the non-covalent interactions statistics in RNA - ligands complexes?

Dataset

Non-redundant complexes of RNA with small molecule ligands.

Calculation of interactions

Non-covalent interactions in all the complexes from the dataset were detected and converted to **SIFs using fingeRNAt**. SIFs were used to calculate interactions statistics.



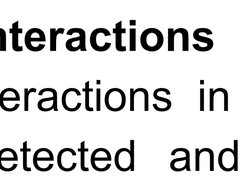
Hydrogen bonds are most frequent (over 65%), but ionic interactions play second most important role, constituting almost one quarter of all interactions.

fingeRNAt applications

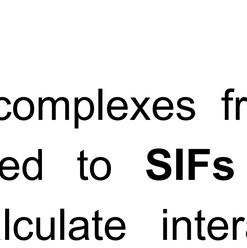
Can interaction patterns be used to discriminate between active and inactive compounds?

Dataset

Target: HIV TAR



Ligands:



Calculation of interactions

Docking was performed using rDock. Non-covalent interactions in all the complexes were detected and converted to **SIFs using fingeRNAt**. SIFs were used to calculate average number of contacts for each interaction.

residue	nucleotide	interaction type	active	inactive	difference	p-value
21	G	Pi-Cation	0.000	0.011	-0.011	0.00
21	G	Pi-Anion	0.000	0.004	-0.004	0.01
22	A	Halogen Bonds	0.000	0.003	-0.003	0.03
22	A	Pi-Anion	0.000	0.011	-0.011	0.00
23	U	Hydrogen Bonds	1.000	0.963	0.037	0.00
23	U	Halogen Bonds	0.000	0.009	-0.009	0.00
23	U	Cation-Anion	0.000	0.007	-0.007	0.00
23	U	Pi-Anion	0.000	0.003	-0.003	0.03
26	G	Hydrogen Bonds	1.000	0.952	0.048	0.00
26	G	Pi-Anion	0.000	0.012	-0.012	0.00
27	A	Halogen Bonds	0.000	0.009	-0.009	0.00
27	A	Pi-Anion	0.000	0.024	-0.024	0.00
39	C	Halogen Bonds	0.000	0.012	-0.012	0.00
39	C	Pi-Anion	0.000	0.004	-0.004	0.01
40	U	Pi-Anion	0.000	0.007	-0.007	0.00

Active and inactive ligands have different binding patterns and this variance may be utilized in rational drug design.

Code availability
github.com/n-szulc/fingeRNAt

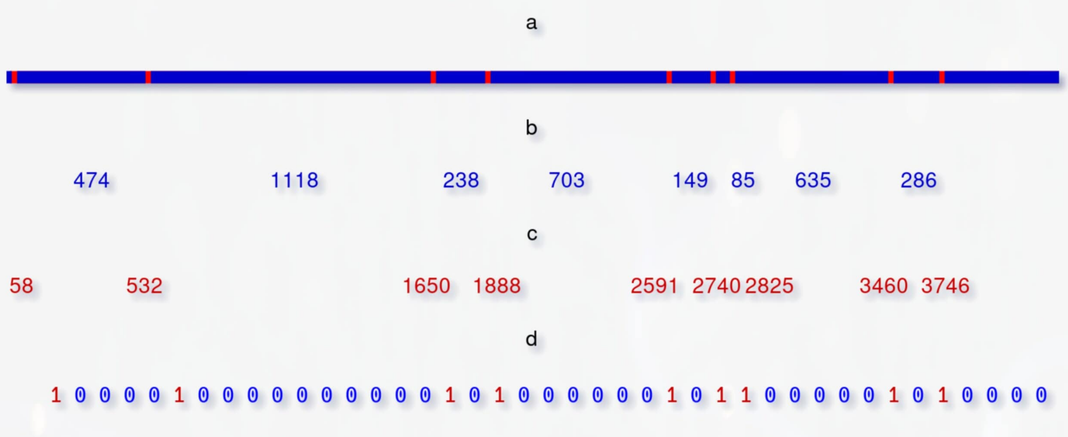
References
Deng, Chuaqui, Singh, *J. Med. Chem.* **2004**, 47(2), 337−344.
Salentin et al., *Nucleic Acids Res.* **2015**, 43(W1), W443-W447.
O'Boyle et al., *J Cheminform.* **2011**, 3(1), 33.
Philips et al., *RNA.* **2013**, 19(12),1605-1616.
Ruiz-Carmona et al., *PLoS Comput Biol.* **2014**, 10(4), e1003571.

Acknowledgments
The „Modeling of dynamic interactions between RNA and small molecules and its practical applications" project is carried out within the TEAM programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (Grant TEAM/2016-3/18 to J.M.B.).

European Funds Smart Growth
FNP Foundation for Polish Science
European Union European Regional Development Fund

Binary genome maps assembly

Binary representation



Representation of consensus genome maps and single restriction maps [rmaps] are similar. It is as ordered set of distances between markers or set of marker positions relative to beginning of genome fragment or chromosome. In our new algorithm we propose a new representation based on quantization and binary sequences. Each position in binary sequence represents constant length genome fragment called quant. 1 in the sequence indicates at least one marker present in quant, 0 indicates no markers. Different optical maps representations are visualised above, where:

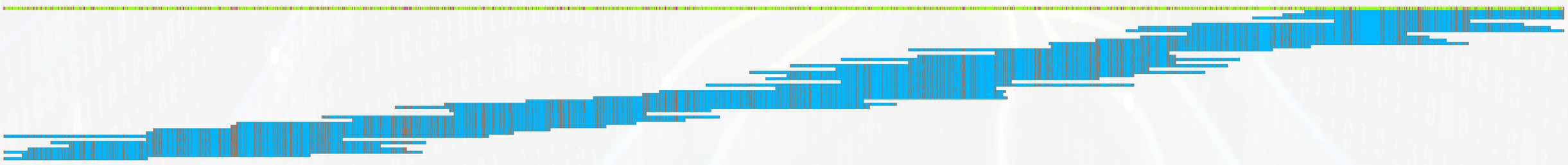
a. is restricted genome with red markers,
b. is distances between markers,
c. is set of positions,
d. is binary genome map

Overlap algorithm

```
1:  function findLeftAlignment(ref, aligned)
2:      maxShift ← minLen(ref, aligned)
                    ▷ could be adjusted e. g. to be half of smaller rmap
3:      bestAlign ← 1          ▷ indicates differnce, 1 means all diffrent bits
4:      bestShift ← 0
5:      for shift ∈ {1, ..., maxShift} do
6:          test ← aligned ≪ shift
7:          result ← test xor ref
8:          truncateLonger(test, ref)
9:          if count(result) / len(result) < bestAlign then
                                                  ▷ it's better alignment
10:             bestAlign ← count(result) / len(result)
11:             bestShift ← shift
        return bestShift
```

Aligning 2 diffrent maps is possible with diffrent estimated distances between map ends. For each combination of positioning of 2 maps only overlaping part of both maps is taken into analysis. For each position of overlaping part XOR operation is performed that is 1's means diffrence at given position, 0's indicates conformation. Lastly number of differences between maps is counted to determine maps similarity for given distance. Alborithm above optimizes differenting part (line 9) but this could be modified into any arbitrary quality function e.g. prefering very long overlaps with a bit more mistakes over smaller overlaps.

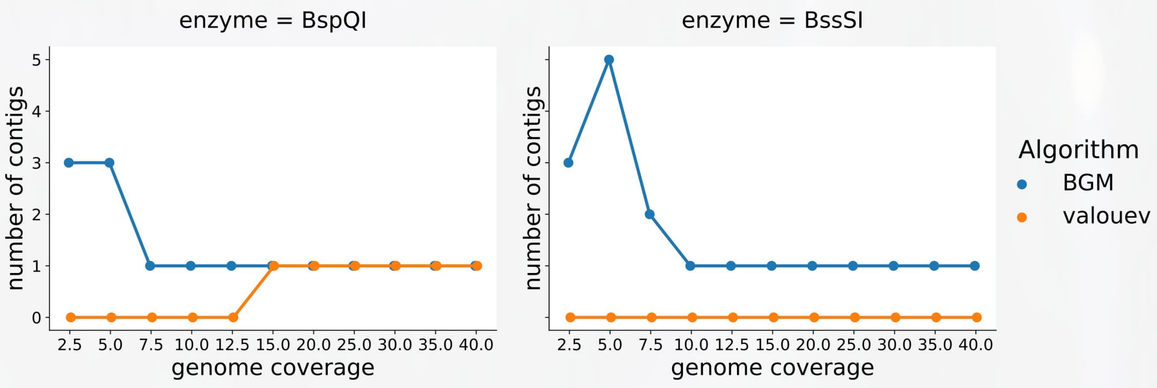
E.Coli maps visualisation



E.Coli maps with coverage x15 generated *in silico* from reference genome using *BspQI* simulated enzyme.

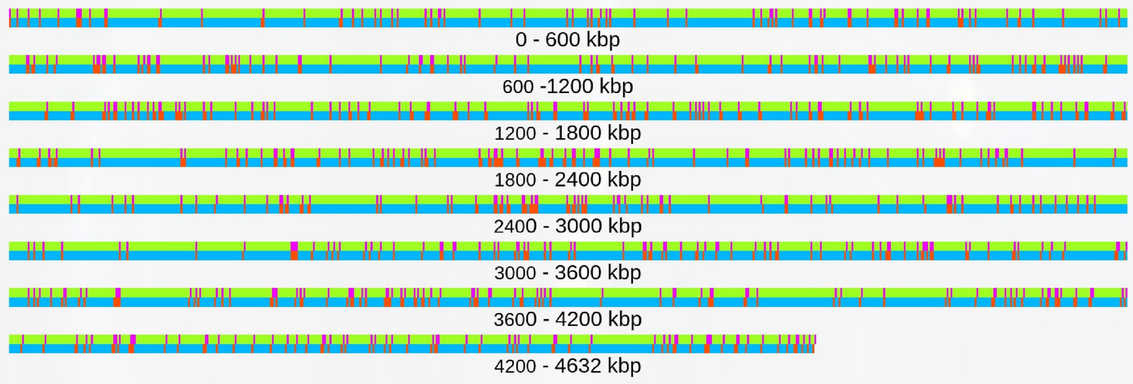
E.Coli experiments and comparison with *valouev et. al.* [1][2]

We performed experiments using simulated datasets from e.coli genome, using BspQI and BssSI enzymes. Both BGM and *valouev et. al.* algorithms used the same set of maps in appropriate format. We measuered time needed by algorithms to finish assembly.

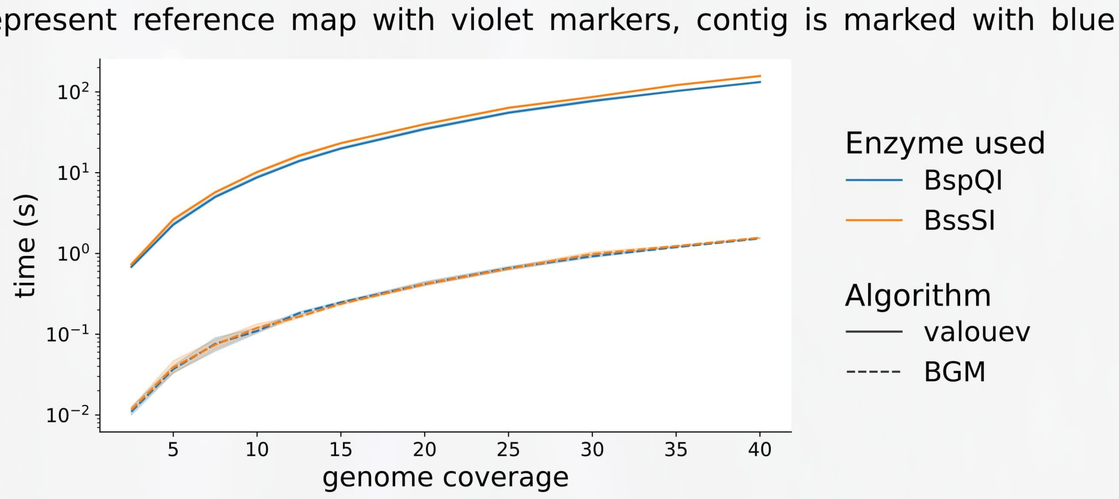


Using only 7.5x coverage of e.coli genome and *BspQI* enzyme we were able to obtain 1 contig. Larger contig was containing information about whole genome. The exact accuracy is not measurable due to nature of quantisation process as discussed above but very restriction site was restored with some artifacts in areas of high marker density and minor missplacement of single bit. In comparison *valouev et. al.* algorithms needed at least 15x coverage.

To obtain 1 contig from e.coli genome maps created with simulated BssSI enzyme we needed 10x coverage. In comparison valouev et. al. algorithms did not produce any contig even with x40 coverage.



Visualisation of 1 BGM BspQI contig where : green color is used to represent reference map with violet markers, contig is marked with blue



Comparison of running time. measurements were performed using single-threaded version of BGM algorithm. Each value vas measured 5 times.

[1] Valouev A, Schwartz DC, Zhou S, Waterman MS. "An algorithm for assembly of ordered restriction maps from single DNA molecules" Proc Natl Acad Sci U S A. 2006 Oct 24;103(43)
[2] Valouev A, Li L, Liu YC, Schwartz DC, Yang Y, Zhang Y, Waterman MS. "Alignment of optical maps" J Comput Biol. 2006 Mar;13(2):442-62.

Przemysław Stawczyk, Robert Nowak
Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland,
e-mail: przemyslaw.stawczyk.stud@pw.edu.pl

A new overlap graph method for DNA sequence assembly

Sylwester Swat[1], Artur Laskowski[1], Jan Badura[1], Wojciech Frohmberg[1], Pawel Wojciechowski[1,2], Aleksandra Swiercz[1,2], Marta Kasprzak[1], Jacek Blazewicz[1,2]

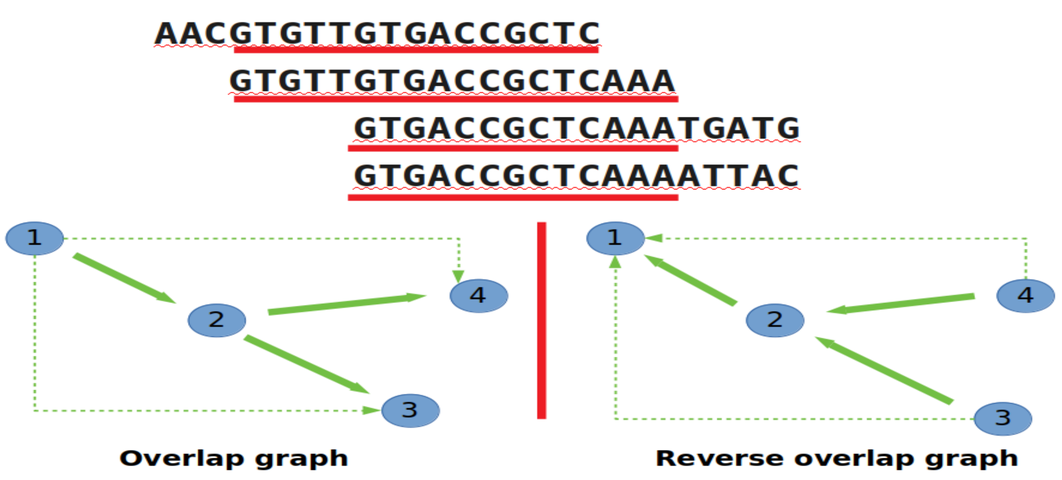
[1] Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[2] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

Introduction

Reconstruction *de novo* of a genome sequence is a great challenge, largely due to computational difficulties connected with processing millions of reads at once. ALGA (ALgorithm for Genome Assembly) is a new method realizing this process and is based on the overlap-layout-consensus approach. The approach consists of three phases: construction of the overlap graph, preparation of the graph for traversal and agreement of final sequences. It is generally viewed as more accurate than the so-called de Bruijn graph approach, but much more demanding in the sense of time and memory. Several new ideas were implemented in order to increase efficiency at each of the phases.

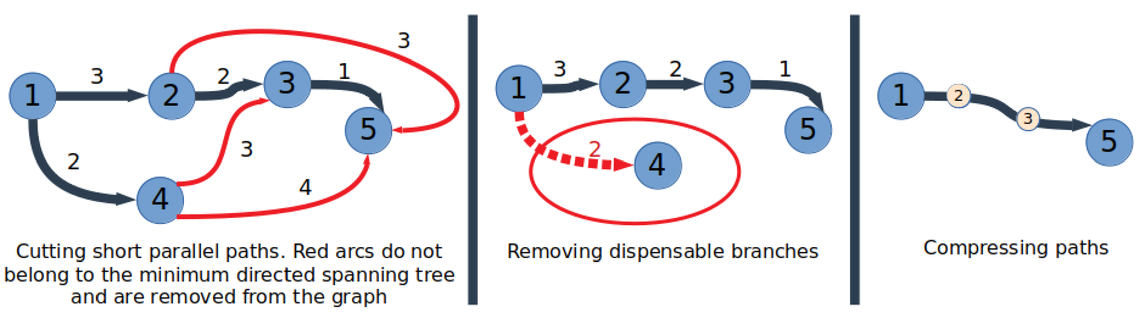
Overlap graph construction

In the first phase of the algorithm, the overlap graph is constructed. In order to reduce memory usage, ALGA creates the reverse of that graph instead and transposes it afterwards. By doing so, ALGA can efficiently recognize and remove transitive edges during the graph creation phase.



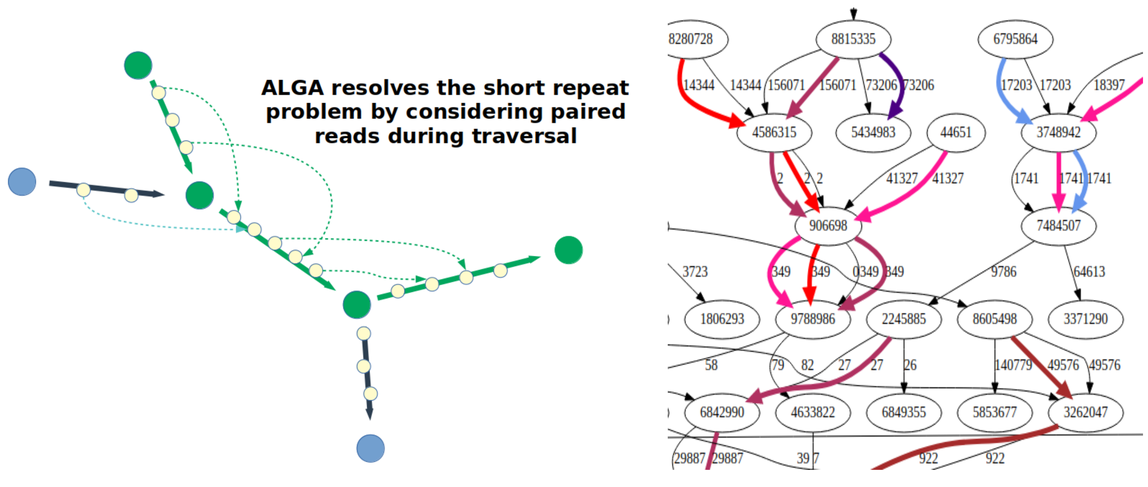
Graph transformation

The overlap graph needs to undergo a few simplification steps that transform it to a state ready for the traversal and creation of contigs. These steps include cutting short parallel paths by solving a variant of the minimum directed spanning tree problem in local subgraphs, trimming branches and compressing paths.



Contig derivation

Each contig is represented by some path in the simplified graph. Starting from a single edge, a path can be extended by appending some edges to its beginning or its end. Path extension is affected by local properties of the graph and connections between paired reads.



Arrows of the same color form a path that corresponds to a final contig

Quality of results

ALGA was tested on a few real data sets obtained for human, bacteria *M. parvicella*, algae *C. sorokiniana* and nematode *C. elegans*. Results were evaluated with the standard tool QUAST [1]. ALGA provides very good results according to metrics such as genome coverage fraction, length of resulting sequences and occurences of misassemblies.

Genome statistics	ALGA	SGA	SOAPdenovo2	MEGAHIT
Genome fraction (%)	90.297	90.538	85.723	91.707
Duplication ratio	1.009	1.108	1.018	1.04
Largest alignment	140 579	67 917	47 254	453 369
Total aligned length	2 775 879 684	3 055 094 153	2 658 682 524	2 892 744 943
NG50	11 495	4481	2495	41 177
NG75	3686	1468	687	14 935
NA50	13 834	4753	3264	39 249
NA75	6648	1813	1544	18 170
NGA50	11 453	4471	2490	35 275
NGA75	3648	1452	677	12 734
LG50	74 181	186 133	318 442	21 702
LG75	191 808	486 999	903 374	52 682
LA50	57 752	172 243	223 748	21 670
LA75	129 950	428 400	520 190	48 872
LGA50	74 402	186 503	318 873	25 104
LGA75	192 694	489 004	907 520	61 377
Misassemblies				
# misassemblies	2230	3688	739	30 456
# relocations	1161	1747	397	5815
# translocations	1034	1863	299	23 354
# inversions	35	78	43	1287
# misassembled contigs	2090	3560	705	27 715
Misassembled contigs length	13 097 797	6 953 129	1 447 334	705 421 771
# local misassemblies	4209	6296	2026	15 849
# scaffold gap ext. mis.	0	0	0	0
# scaffold gap loc. mis.	0	0	0	0
# unaligned mis. contigs	1189	1079	339	2398
Unaligned				
# fully unaligned contigs	17 690	44 272	10 928	115 368
Fully unaligned length	8 777 301	15 838 074	5 297 029	40 941 253
# partially unaligned contigs	2031	1684	752	2067
Partially unaligned length	3 193 085	2 313 312	939 636	3 325 133

Comparison of several assemblers for a whole human genome data set

Performance

ALGA is implemented with the use of different parallelization schemes, effective memory management and incorporation of cache-locality improvement techniques.

	Memory peak (GB) and elapsed time (hh:mm:ss)			
	M. parvicella	C. elegans	C. sorokiniana	H. sapiens
ALGA	1,7 00:01:29	19,3 00:24:48	27,8 00:48:11	247,3 15:31:50
GRASShopPER	17,6 02:02:28	361,6 57:12:58	638,9 53:33:33	out of memory > 750 GB
Velvet	9,3 00:08:52	21,0 02:05:00	107,6 14:42:04	out of memory > 750 GB
SGA	0,3 00:11:47	3,3 02:33:15	7,7 09:54:32	43,5 98:58:49
SOAPdenovo2	2,5 00:02:13	7,3 00:27:02	16,3 01:21:05	269,3 15:46:12
MEGAHIT	0,8 00:02:38	6,1 00:31:59	18,9 04:30:26	87,6 15:43:34
SPAdes	10,6 00:26:50	14,4 12:14:42	49,2 22:22:05	out of memory > 400 GB
Platanus	117,6 00:16:39	122,1 01:03:06	120,0 03:50:25	out of memory > 750 GB

Time and memory usage of tested assemblers for data sets obtained for *M. parvicella*, *C. sorokiniana*, *C. elegans* and H. sapiens

References and Acknowledgements

[1] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. Quast: Quality assessment tool for genome assemblies. *Bioinformatics*, 29:1072–1075, 2013.

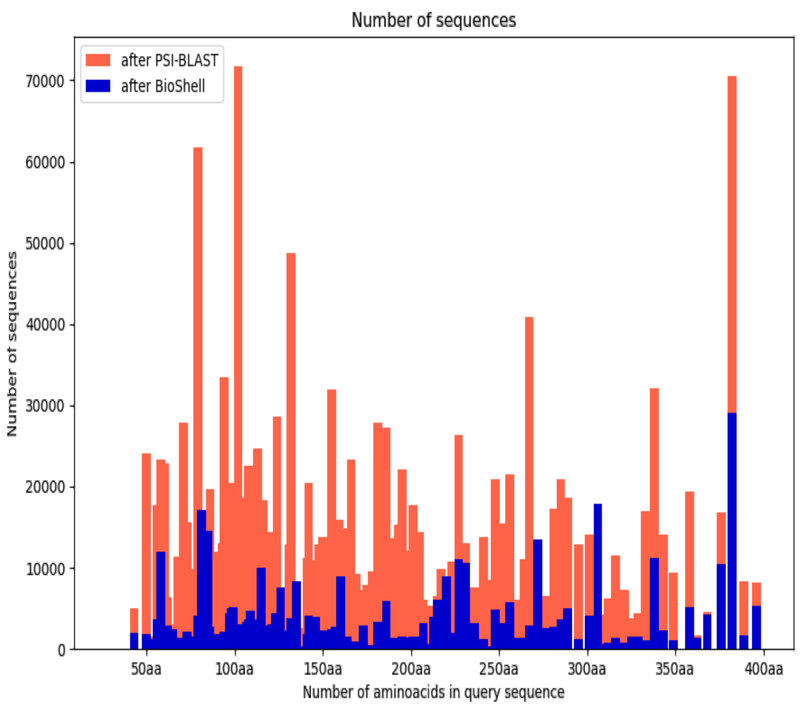
This work was supported by the European Regional Development Fund [grant POIR.04.02.00-30-A004/16] and carried out in the European Center for Bioinformatics and Genomics, Poznan University of Technology.

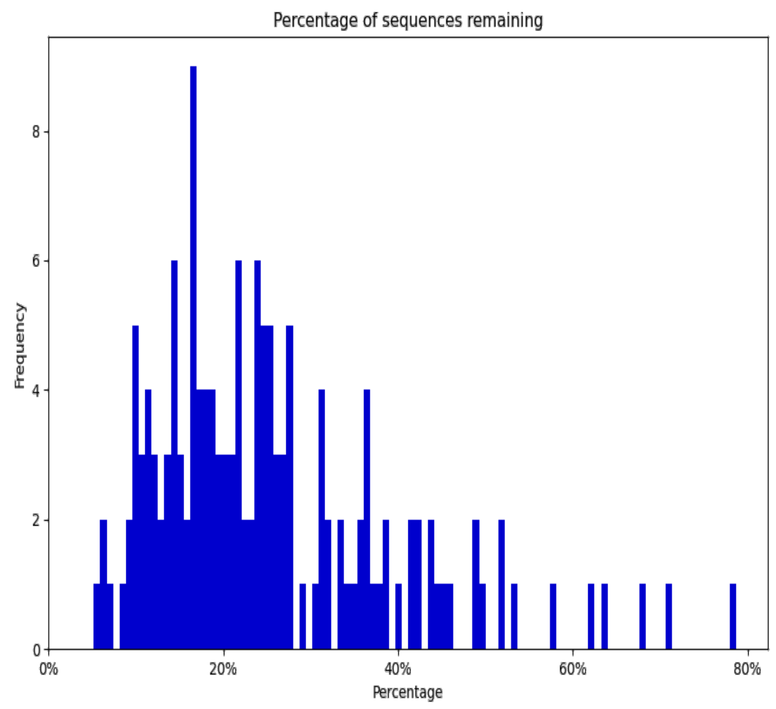
BioShell software reduce an overrepresentation of sequences, **increase quality** of a **MSA,** build **better sequence profile.**

Automated approach for sequence profile generation

Marcin Piwowar, Dominik Gront, Faculty of Chemistry, University of Warsaw

BioShell





Despite the recent progress in the field, construction of a multiple sequence alignment (MSA) still requires a considerable effort from a human expert. Automated methods can make various errors, that often result from an unfortunate selection of input sequences, e.g. when set of these sequences is redundant. In this contribution I used tools from BioShell package (*ap_blast_nonredundant*, *ap_filter_msa*) to filter an input sequence set and construct better MSA in an iterative fashion.

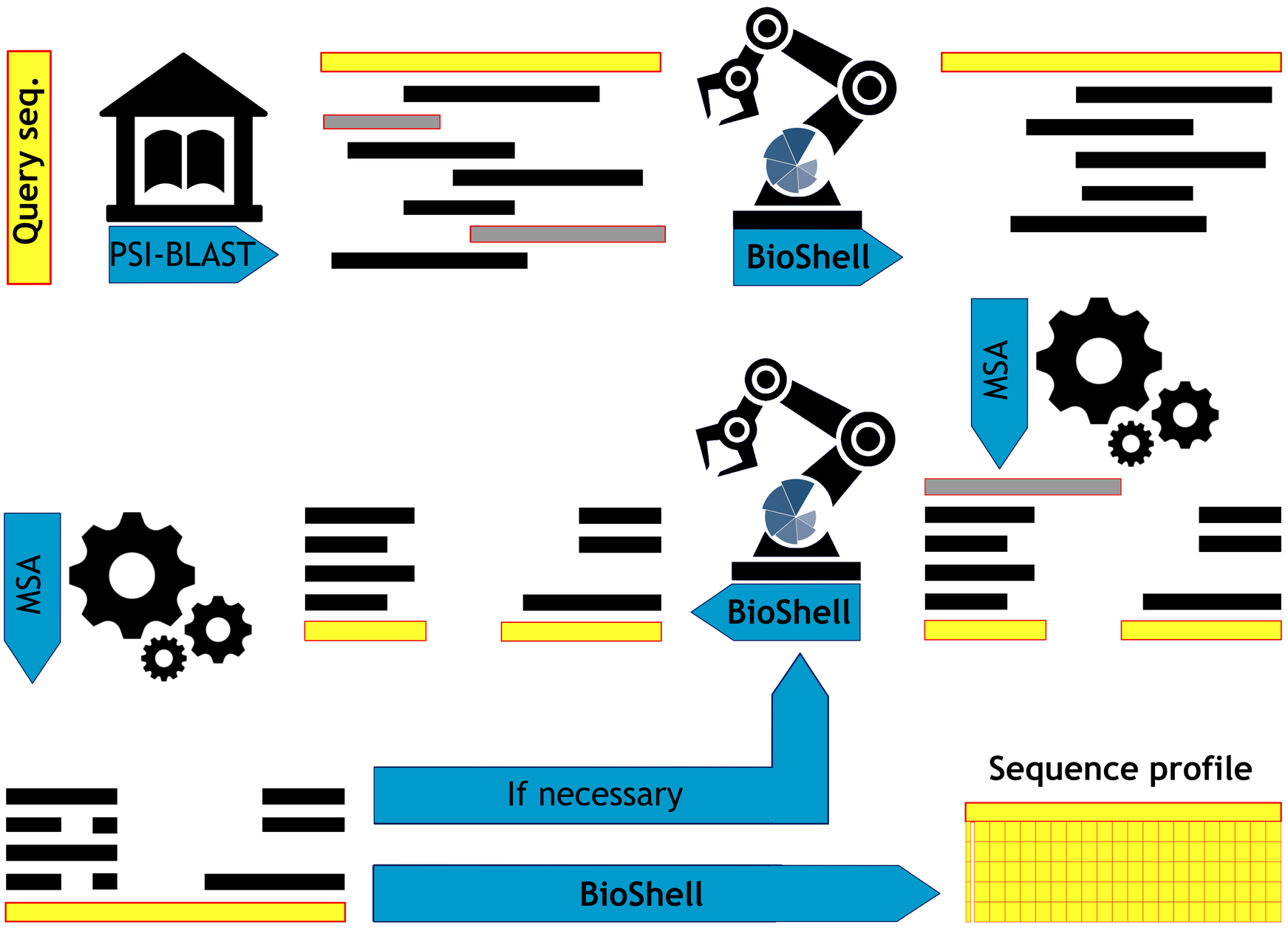
Results

The method has been tested and validated on a nonredundant set of sequences from HOMSTRAD and UniRef. BioShell for the identity parameter equal to 50% removes up to 99% of all found sequences. MSA is done with greater accuracy, because profile will be constructed from fewer, but more significant sequences.

Conclusions

- BioShell makes a set of sequences to be taken into account during MSA less redundant
- Protocol using BioShell with external software generates sequence profile with more biological information
- Because of less number of sequence, sequence profile is build up to 40 times faster
- Human expert applying BioShell is not forced to manually improve MSA

Applications will be tested on other databases.



UNIVERSITY OF WARSAW

Faculty of Chemistry
University of Warsaw

Take a picture to read more about BioShell

In silico evaluation of SARS-CoV-2 primers performance

Michał Kowalski[1], **Alina Frolova**[1,2], **Witold Wydmański**[1], **Wojciech Branicki**[1] **and Paweł Łabaj**[1]

1. Malopolska Centre of Biotechnology, Jagiellonian University University, ul. Gronostajowa 7A, 30-387 Krakow, Poland
2. Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine, 150, Zabolotnogo Str., Kyiv, 03143, Ukraine

Correspondence: m.kowalski@doctoral.uj.edu.pl

*The first three authors have equal contribution

Introduction

Emergence of novel coronavirus SARS-CoV-2 had became a global threat in a blink of an eye. Many research groups, coroporations and organizations had proposed sets of primers for RT-qPCR technology that ought to be reliable, primary source of diagnostic power all around the world. During the course of pandemic, studies and reports had shown that nowt every proposed set of primers can amplify the virus, thus false negative and false positive results had became a serious problem. Since global lockdown hadn't been handled properly in plethora of countries, evolution of SARS-CoV-2 had became region specific, which introduced mutations that altered performance of globally recommended primers. Our group inspired by diagnostic work of our collegues from *Human Genome Variation Research Group* from Malopolska Centre of Biotechnology and in collaboration with *MetaSUB* consortium had addressed those problems by performing a set of in-silico experiments for the evaluation of SARS-CoV-2 primers performance.
Those in-silico tests helped to establish what is the most recommended set of primers and their had been put all together as a *Python* library we called *pyprimer*, which will be available as an open-source solution applicable to benchmark performance of primers and to design them for PCR-family laboratory techniques.

Data and Methods

Global dataset was obtained from *GISAID* repository, then filtered with strict criterion concerning quality of sequences:
- Number of ambiguous nucleotides ("*N*") must be less than or equal to 5%
- No sequences with ambiguous nucleotides within primer bindning sites are allowed
- Metadata of sequences must be complete (or really easy to impute)

Regional Polish dataset was obtained from collection of sequences obtained in Małopolska Centre of Biotechnology by *Human Genome Variation Research Group*. Polish dataset hadn't required any filtering. Sequences of primer pairs were obtained from WHO and CDC websites.
Processing and analysis of data had been performed in following steps:

1. Multiple Sequence Alignment (for later construction of probability matrices)
2. Description of physical properties of sequences and primer pairs
3. Fuzzy matching of primer pairs with Levensthein distance set to zero
4. Filtering and selection of cannonical amplicons created by in-silico bindings
5. Evaluation of stability of primer pairs based on the Primer Pair Coverage metric

$$PPC = \frac{Fm}{Fl} \times \frac{Rm}{Rl} \times (1 - CVm)$$

$$CVm = \frac{\sigma_(Fm, Rm)}{\mu_(Fm, Rm)}$$

Where:
PPC - Primer Pair Coverage
Fm - Number of nucleotides that matched sequence in F primer
Fl - Total length of F primer
Rm - Number of nucleotides that matched sequence in R primer
Rl - Total length of R primer
CVm - Coefficient of variation for matched regions
σ - Standard deviation
μ - Arithmetic mean

6. Matching of probes to amplicons with same Levensthein distance criterion and discarding of ill fitted records.
7. Exploration of dimerization properties with *RNAfold*

As illustrated in the results, post-hoc analysis had been also performed to show in easy to percieve and graphic way, which primers are the onec that after in silico evaluation should be recommended for further use.

Results

Fig.-1 shows the Venn diagrams with four sets of primers that had the highest performance during in-silico evaluation. Although *US_CDC_2019-nCoV_N3* had the same in-silico performance as primers from *Institut Pasteur*, they had been retracted from global usage, hence they are not taken into account in discussion. Sets of primers shown at the global Venn diagram are recommended for in-laboratorium validation before applying them for diagnostic purposes.

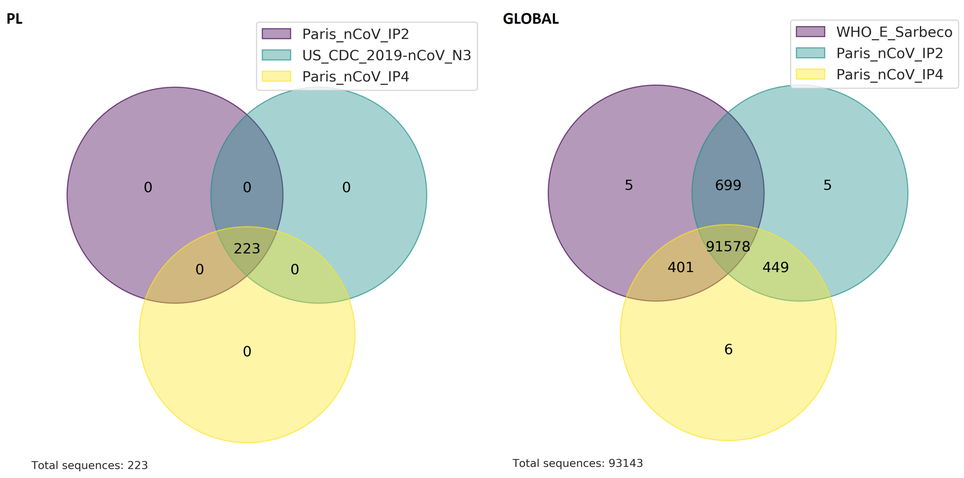


Figure 1: Venn diagram of three best primer pairs for diagnostic purposes of SARS-CoV-2 identification. On the left for Polish Sequences, on the right for global sequences downloaded from *GISAID*.

Fig.-2 shows the entire benchmark results in a form of horizontal bar plot, to underline the lack of performance in many of sets. Versioning of primers is kept due to ambiguous IUPAC coding in many of them.
To being able to determine whether chemical properties of primers will allow for the amplification of target, one must also consider occurrence of primer dimer problem

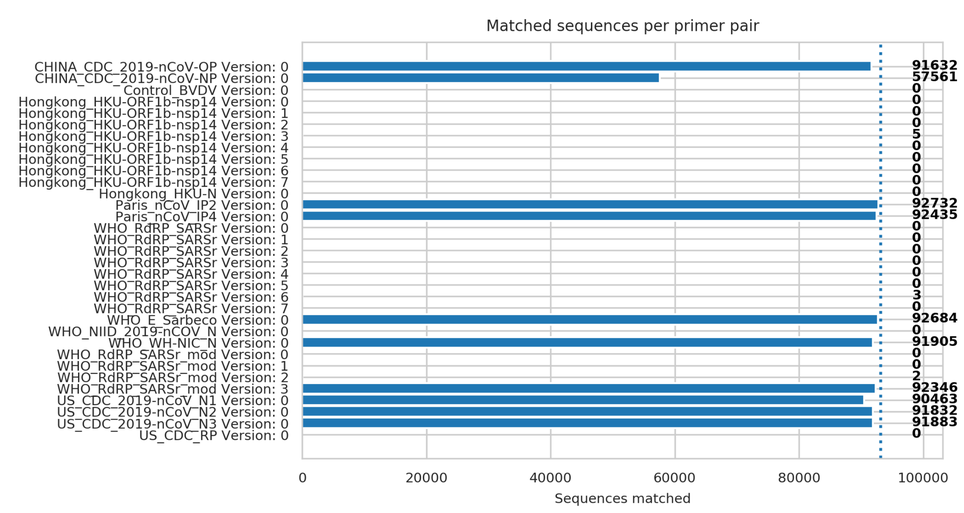


Figure 2: Horizontal bar plot that shows the overall performance of all primer sets. Length of the bars is determined by how much sequences given pair of primers had been able to match conservatively.

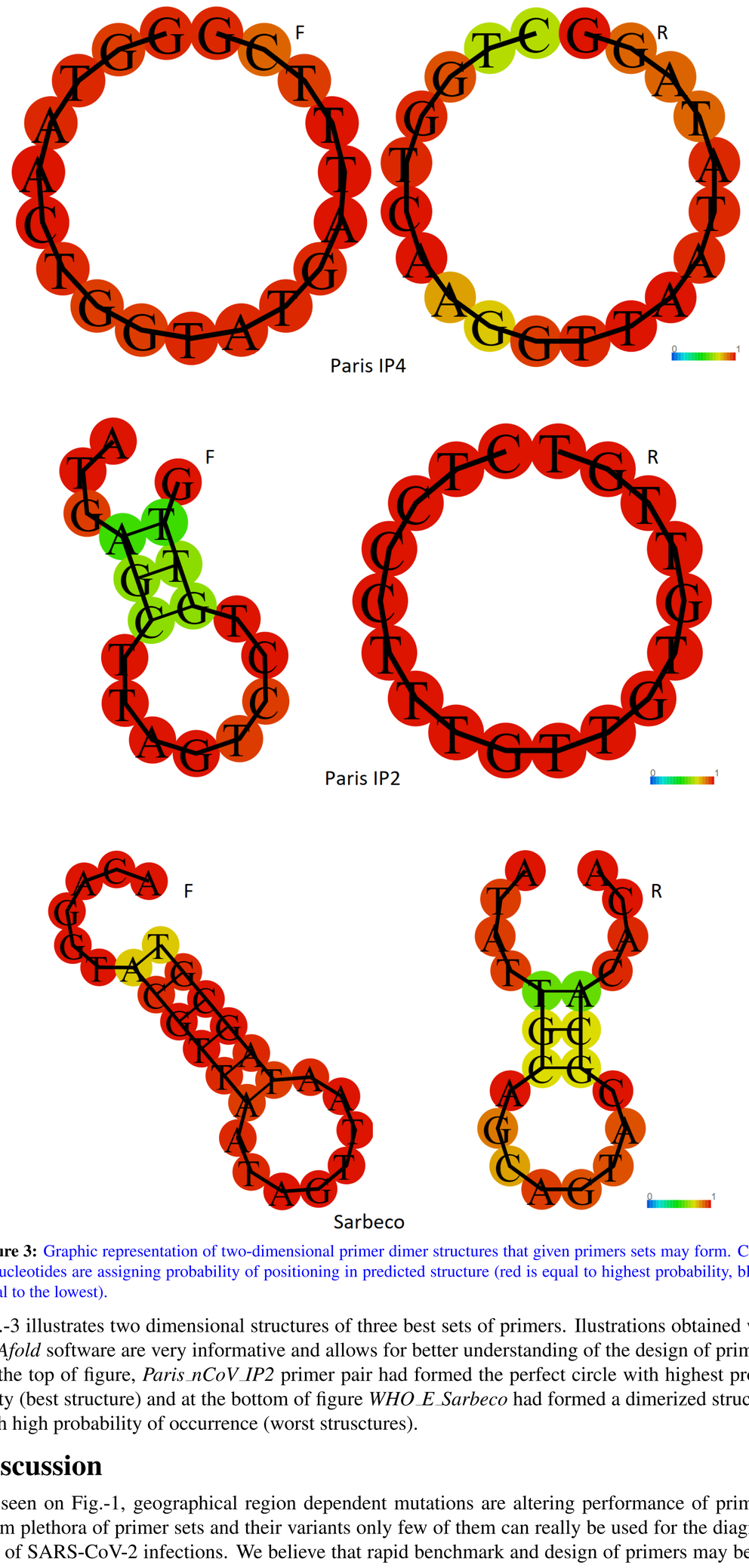


Figure 3: Graphic representation of two-dimensional primer dimer structures that given primers sets may form. Colors of nucleotides are assigning probability of positioning in predicted structure (red is equal to highest probability, blue is equal to the lowest).

Fig.-3 illustrates two dimensional structures of three best sets of primers. Ilustrations obtained with *RNAfold* software are very informative and allows for better understanding of the design of primers. At the top of figure, *Paris_nCoV_IP2* primer pair had formed the perfect circle with highest probability (best structure) and at the bottom of figure *WHO_E_Sarbeco* had formed a dimerized structure with high probability of occurrence (worst strusctures).

Discussion

As seen on Fig.-1, geographical region dependent mutations are altering performance of primers. From plethora of primer sets and their variants only few of them can really be used for the diagnostics of SARS-CoV-2 infections. We believe that rapid benchmark and design of primers may be the key for better diagnostic power, and that *pyprimer* python library may drastically improve the state of diagnostics while applied to design of primers precisely for geographical regions of interest (by avoiding generalization of the problem).

Acknowledgments

We would like to thank our colaborators from MetaSUB for expanding our research into other region stratified datasets and validating our work and conclusions with their own pipelines and methodologies. Special thanks to Emmanuel Dias-Neto and Israel Tojal da Silva from AC Camargo Cancer Center in Sao Paulo, Christopher E. Mason and Jonathan Foox from Weill Cornell Medicine, entire Human Genome Variation Research Group from Malopolska Centre of Biotechnology at Jagiellonian University and Krzysztof Pyré from ViroGenetics - BSL3 Laboratory of Virology at Jagiellonian University. We would also like to thank all of *GISAID* submitters.

Estimated nucleotide reconstruction quality symbols of basecalling tools for Oxford Nanopore sequencing

Wiktor Kuśmirek

Warsaw University of Technology, Institute of Computer Science, Warsaw, Poland

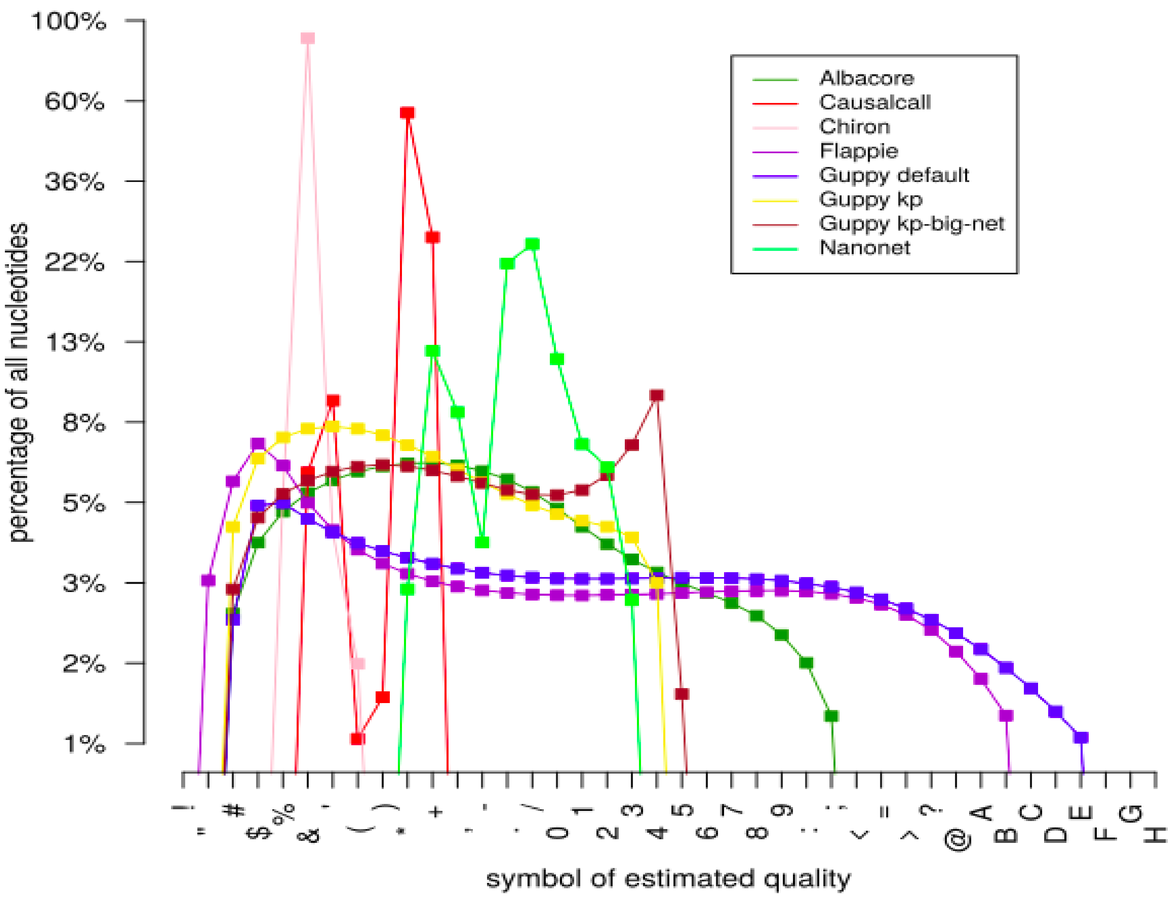
Abstract

Currently, one of the fastest growing DNA sequencing technologies is nanopore sequencing. One of the key stages of processing sequencer data is the basecalling process, which from the input sequence of currents measured on the pores of the sequencer reproduces the DNA sequences called DNA reads. Many of the applications dedicated to basecelling together with the DNA sequence provide the estimated quality of reconstruction of a given nucleotide.

Herein, we examined the estimated quality of nucleotide reconstruction reported by another basecallers. The results showed that the estimated reconstruction quality reported by different basecallers may vary depending on the tool used. In particular, for some tools, along with successive symbols of the estimated reconstruction quality (which theoretically should mean more and more accurate reconstruction), the real quality of the nucleotide increases (the number of matched nucleotides increaces and the number of errors decreases). However, there are tools that report the estimated reconstruction quality in the basecalling results, but these values are in no way interpretable. What is more, the estimated reconstruction quality reported in basecalling process is not used in any investigated tool for processing nanopore DNA reads..

Dataset

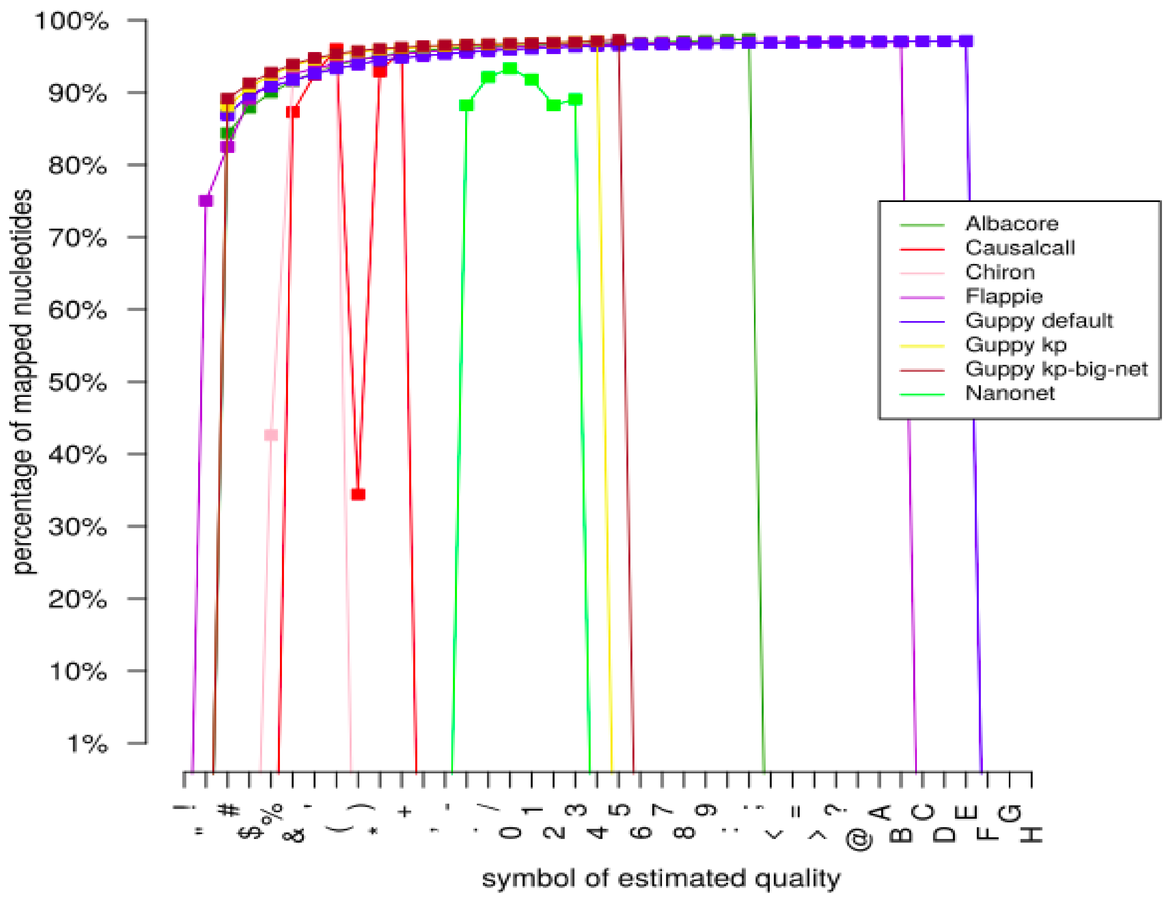
Basecaller	No. of reads	Sum [Mbp]	Mapped [%]	Match [%]
Albacore	4467	116.63	95.77	86.64
Causalcall	4467	115.12	92.21	84.36
Chiron	4467	85.44	81.88	80.43
Flappie	4467	115.04	95.44	89.66
Guppy default	4467	115.48	96.47	89.68
Guppy kp	4467	113.84	96.35	87.60
Guppy kp-big-net	4467	114.99	97.32	89.73
Nanonet	7702	118.18	67.33	84.05



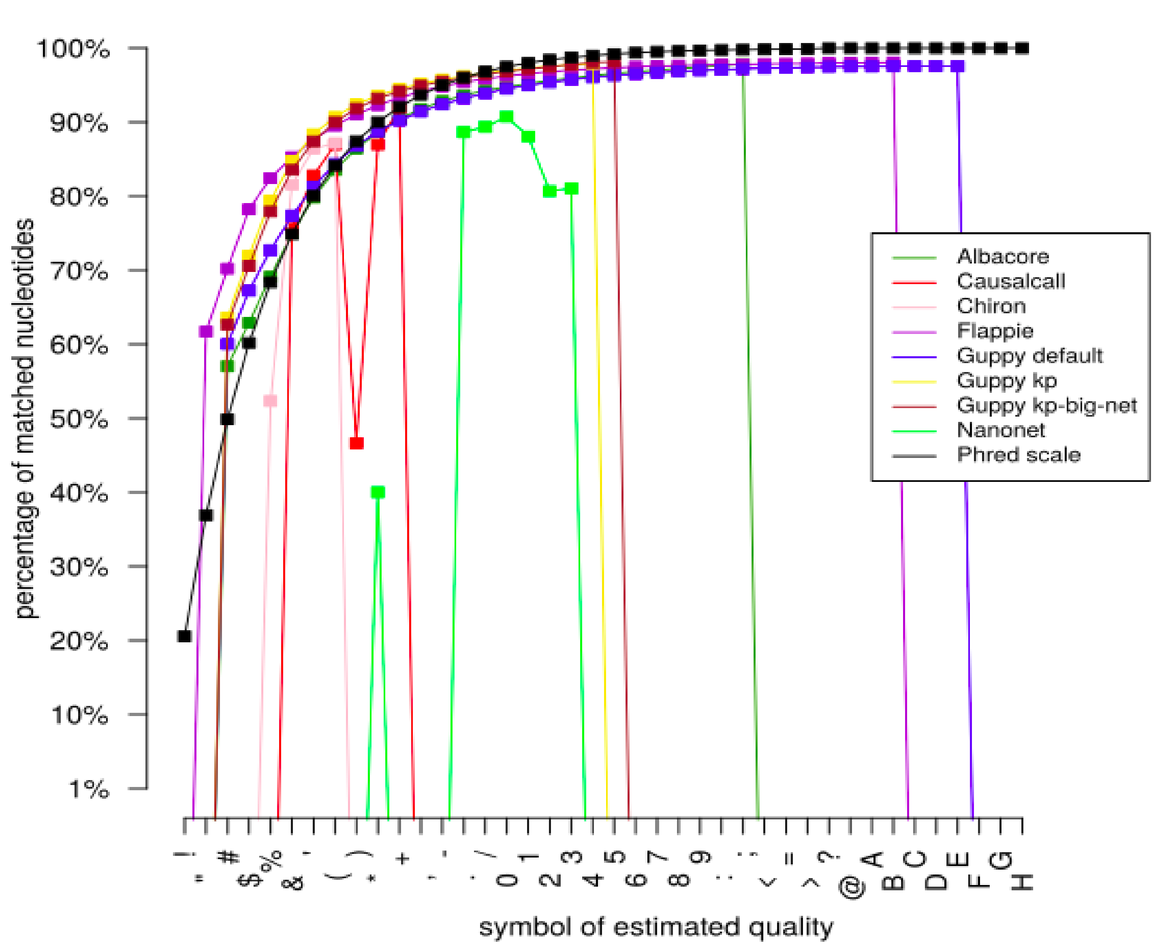
Acknowledgments

The project was funded by POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB). This work has been also supported by the Polish National Science Center grant Preludium 2019/35/N/ST6/01983.

Results



A



References

Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt. "Performance of neural network basecalling tools for Oxford Nanopore sequencing." Genome biology 20.1 (2019): 129.

David, Matei, et al. "Nanocall: an open source basecaller for Oxford Nanopore sequencing data." Bioinformatics 33.1 (2017): 49-55

Teng, Haotian, et al. "Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning." GigaScience 7.5 (2018): giy037.

Ḱoren, Sergey, et al. "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation." Genome research 27.5 (2017): 722-736.

Boža, Vladimír, Broňa Brejová, and Tomáš Vinař. "DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads." PloS one 12.6 (2017): e0178751.

Stoiber, Marcus, and James Brown. "BasecRAWller: streaming nanopore basecalling directly from raw signal." BioRxiv (2017): 133058.

Wrocław University of Environmental and Life Sciences

K. Kotlarz, M. Mielczarek, T. Suchocki, B. Czech, B. Guldbrandtsen, J. Szyda

Wrocław University of Environmental and Life Science, University of Bonn

THETA
Statistical Genetic Group
Institute of Animal Genetics

DIG (DEEP)ER

Deep learning algorithms for the imbalanced classification of correct and incorrect SNP genotypes from WGS pipelines

#1 MATERIALS

Whole-genome DNA sequence of four traditional Danish Red Dairy Cattle bulls:

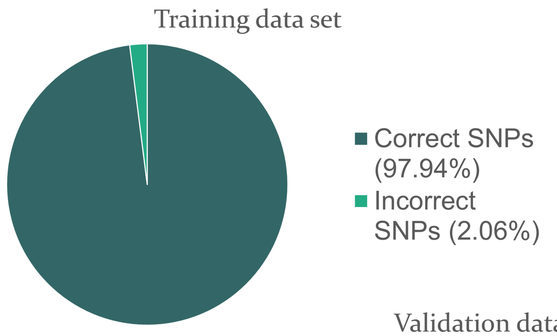
1) The training data set—**three animals**,
2) The validation data set—**the fourth animal**.

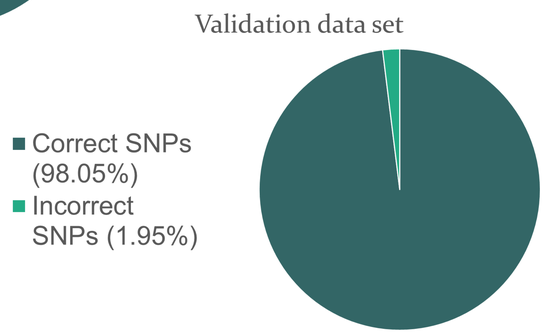
Correct SNPs (concordant WGS—Chip):

1) Training data set: 2 227 995 SNPs,
2) Validation data set: 749 506 SNPs.

Incorrect SNPs (discordant WGS—Chip):

1) Training data set: 46 920 SNPs,
2) Validation data set: 14 940 SNPs.



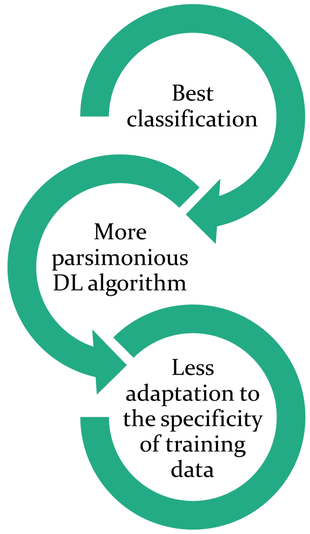


Contact:

Department of Genetics
Wrocław University
of Environmental
and Life Sciences

7 Kożuchowska Street
51-631 Wrocław
Poland
krzysztof.kotlarz@upwr.edu.pl

#4 CONCLUSIONS



#2 METHODS

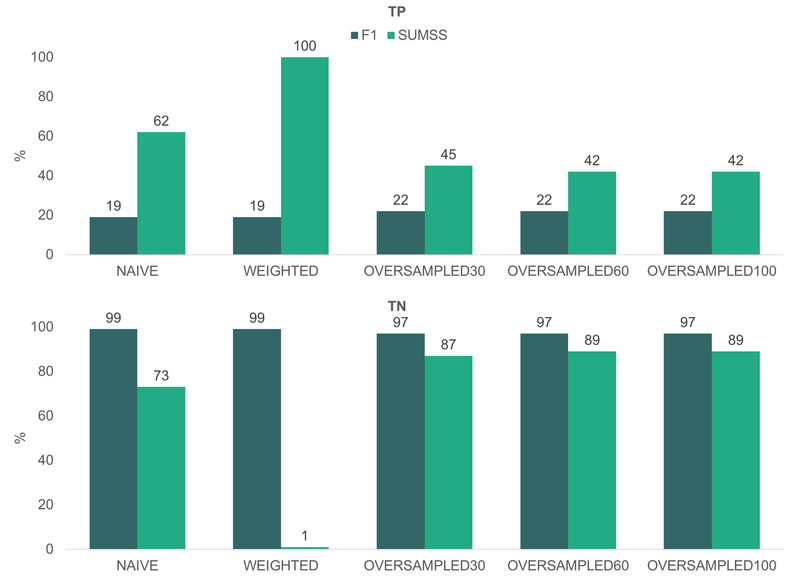
Deep Learning algorithms

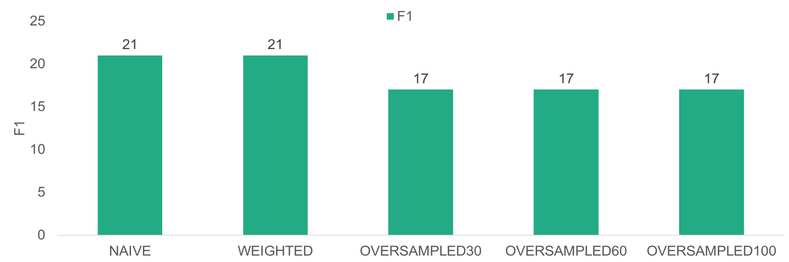
1) Naïve algorithm
2) Weighted algorithm
3) Oversampled algorithm; oversampled of the incorrect SNP:
   - 30%
   - 60%
   - 90%

Cutoff points

1) The estimated cutoff points for each model by:
   - $F1 = \frac{2TP}{2TP+FN+FP}$
   - $SUMSS = \frac{TN}{TN+FP} + \frac{TP}{TP+FN}$

#3 RESULTS





Classification of **validation data** by the algorithms, based on the cutoff thresholds for the **F1** or **SUMSS** metrics.

1) **True positive (TP)**—an incorrect SNP classified as incorrect,
2) **False negative (FN)**—an incorrect SNP classified as correct,
3) **True negative (TN)**—a correct SNP classified as correct,
4) **False positive (FP)**—a correct SNP classified as incorrect,
5) **F1**—values of the F1 metric.

DNA sequence features underlying large-scale duplications and deletions in humans

Mateusz Kołomański[1], Joanna Szyda[1,2], Magdalena Frąszczak[1], Magda Mielczarek[1,2]

[1] Biostatistics group, Department of Genetics, Wroclaw University of Environmental and Life Sciences
[2] National Research Institute of Animal Production

Objective

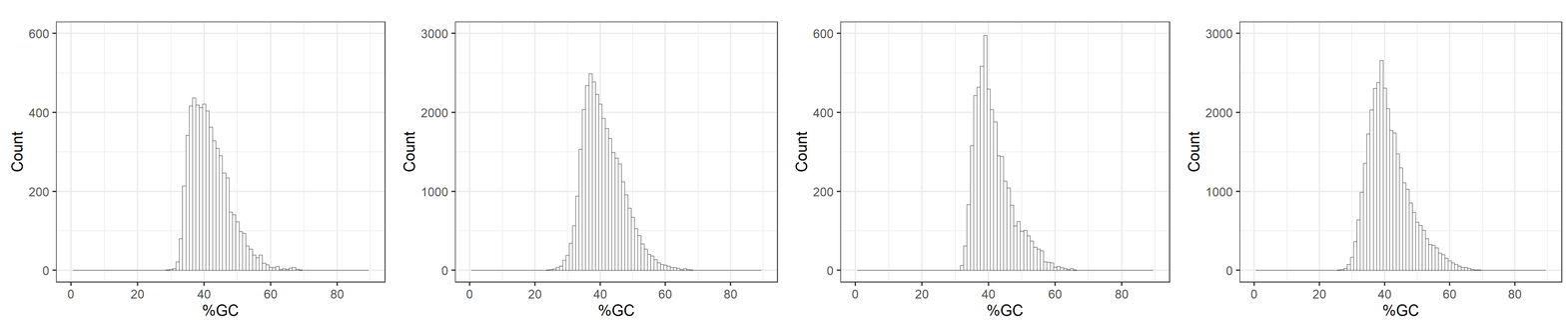
Characterizing regions of human genome that are susceptible to formation of Copy Number Variants.

Conclusions

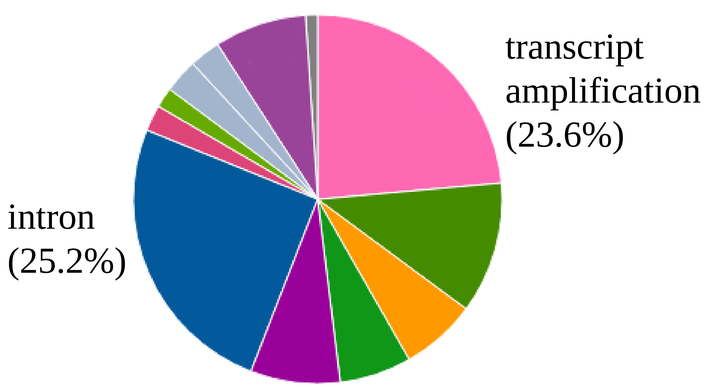
- **Deletions and sequences upstream of Copy Number Variants have low sequence complexity.**
- **Large proportion of CNVs overlap with introns.**

Results

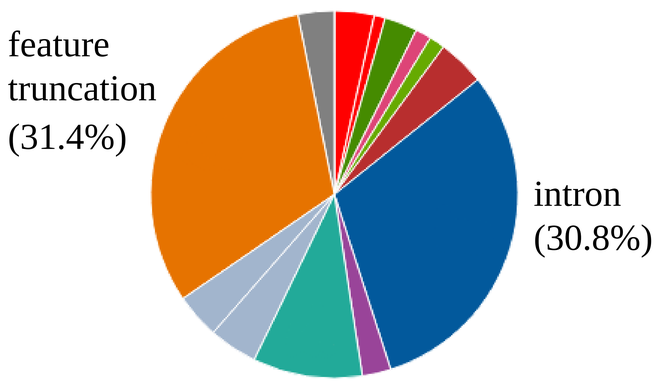
GC-pairs content



Functional annotation



duplications



deletions

Material & Methods

- Database of *1000 Genomes Project*
- 5 867 duplicated and 33 181 deleted regions
- 100 bp-long sequences flanking CNVs
- Random regions
- Seqences extracted from reference genome (GRCh38)

- Analysis regarded:
- Unknown nucleotide contents (14 CNVs)
- Guanine-Cytosine pairs content
- Sequence complexity → sDust software
- CNV-related and randomised regions comparison → Wilcoxon test
- Functional annotation → VEP

Canonical Correlation- based bioinformatic analysis for effective melanoma biomarker discovery

Sonia Wróbel [1], Ewa Stępień [1], Cezary Turek [2], Monika Piwowar [2]

1 Department of Medical Physics, Jagiellonian University, Marian Smoluchowski Institute of Physics, Kraków, Poland, 2 Department of Bioinformatics and Telemedicine, Jagiellonian University–Medical College, Krakow, Poland

ABSTRACT

- Here we introduce a new method based on canonical correlation analysis (CCA) that uses real-life dataset to meet the challenge of **melanoma biomarker discovery** [1-2]. The bioinformatics pipeline was successfully applied to human skin **melanoma multi-OMICS datasets** containing: **(1)** microvesicle micro-RNA transcriptomics, **(2)** microvesicle proteomics, **(3)** cell-total-RNA transcriptomics.
- The method applies a **sparse CCA (sCCA)** to three matrices, starting from features correlation across integrated experimental data [3].
- Validation using clinical data as well as supporting meta-data from extracellular vesicle dedicated databases allows the identification of evidence-based candidates for highly **significant molecular signatures** like **melanoma-associated microRNAs and oncoproteins**.

CHALLENGE

- Next Generation Sequencing (NGS) and other advanced large-scale experimental methods provide enormous amounts of **multi-dimensional biological data**. Understanding the interactions between transcriptomics, proteomics and other types of data generated using different platforms is fundamental. In such analyzes, the **integration of multiple OMICS** datasets together and selection of variables is a key to obtain **interpretable results**.
- **Canonical Coronation Analysis (CCA) is one of the most powerful method** for this bioinformatic challenge. Over the last years, a number of promising results for implementing CCA in the integration of OMICS data have been proposed [4-5].

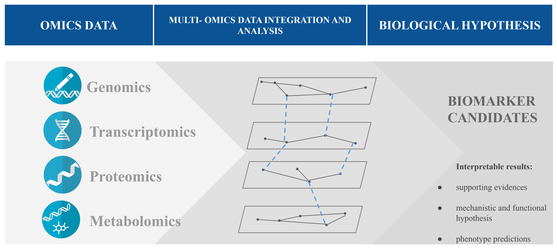


Fig. 1 Multi-omics data integration and analyses as effective method for identification of the biomarker candidates using information of biological interrelationships, bioactive molecules and their functions.

MELANOMA MODEL

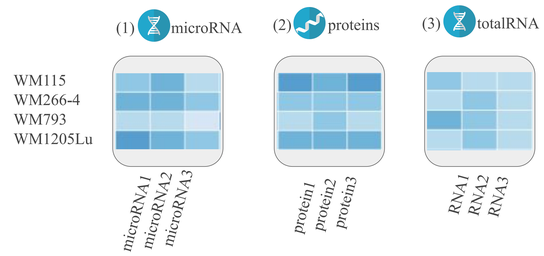
- We used two melanoma cell line models:
  - WM115: a primary vertical growth phase cell line and WM266-4: a lymph node metastasis vertical growth phase cell line. Bothe established from the same patient.
  - WM793: a primary vertical growth phase cell line and WM1205Lu: a metastatic vertical growth phase cell line. First established from patient and second from nude mice lung metastases.

Fig. 2 Melanoma cell lines: WM115, WM266-4, WM793, WM1205Lu originated from the European Searchable Tumour Cell Line and Data Bank (ESTDAB)- A Collection of Immunologically Characterised Melanoma Cell Lines and Databank (Tübingen, Germany).

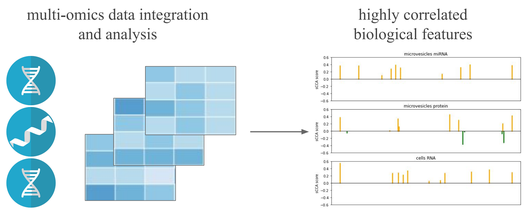
METHOD OVERVIEW

- As an input data we used proprietary microvesicle micro-RNA transcriptome and open source datasets for microvesicle proteome and cells total-RNA transcriptome [6-7]. Each data type was derived for standardized cell lines: WM115, WM266-4, WM793 and WM1205-Lu.
- Data analysis and interpretation was done using method based on sparse canonical correlation bioinformatics method developed in our research group (Fig. 2).
- To conduct sparse CCA we use matrices which represent different sets of features (1) microvesicle micro-RNA transcripts, (2) microvesicle proteins and (3) cell-total-RNA transcripts. on the same set of melanoma cell lines samples. Multi-OMICS dataset has samples in rows and the features on columns. Prepared matrices always had the same number of rows, but had different numbers of columns.
- In next step there was the visualization of highest correlated features and a list of this features with respective ranks.
- Last step provided pathways analysis and annotations supporting each functional insight from extracellular dedicated databases.

a. Input data:



b. Sparse Canonical Correlation Analysis (sCCA)



c. Results: biomarker candidates with supporting biological findings

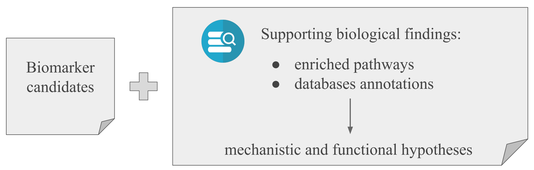


Fig. 3 Method overview. a) Method requires three input matrices for different genomics features for the same set of samples. In this study we used (1) microvesicle-micro-RNA transcripts, (2) microvesicle proteins and (3) cell-total-RNA transcripts for four melanoma cell lines models: WM115, WM266-4, WM793 and WM1207Lu. **b)** Method provides visualization of highest correlated features and a list of this features with ranks. **c)** Last step provides pathways analysis and annotations supporting each functional insight from extracellular dedicated databases for example: ExoCarta (www.exocarta.org), Vesiclepedia (www.microvesicles.org) [8].

RESULTS

- We identified highly correlated microRNA, proteins and totalRNA (Fig. 4 and Table 1). The top 30 highest ranked by the algorithm were selected for further analysis steps (five each with the highest negative and positive correlation from each of the data types).

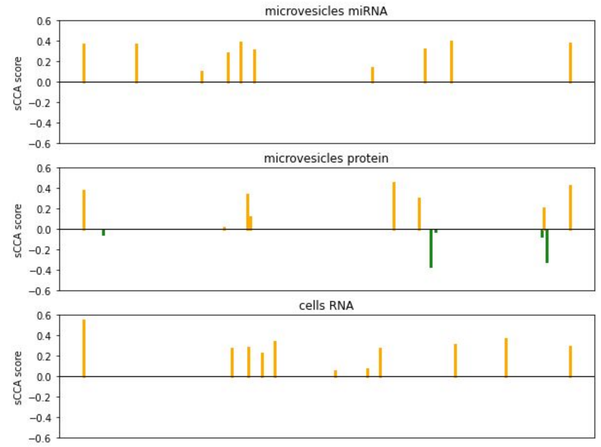


Fig. 4 Visualization of sCCA results for melanoma: microvesicles miRNA, microvesicles proteins and cell totalRNA. The x-axis shows features, while the y-axis shows the cCCA score. Presented bioinformatic method allows to adjust the number of displayed features, starting with the most important ones.

Table 1. Results for 30 top scored sCCA melanoma 1) microvesicles miRNA, 2) microvesicles proteins and 3) cell totalRNA with sCCA scores.

miRNA ID	sCCA score	protein ID	sCCA score	RNA (Gene) ID	sCCA score
MIMAT0002866	3,95E-01	Q15029	4,45E-01	AMIGO2	5,45E-01
MIMAT0002837	3,86E-01	Q14103	4,17E-01	SVEP1	3,60E-01
MIMAT0004687	3,73E-01	P25788	3,73E-01	IL31RA	3,38E-01
MIMAT0000724	3,67E-01	P27695	3,30E-01	RPS14P8	3,07E-01
MIMAT0000281	3,58E-01	Q6DD88	2,98E-01	ZNF812P	2,88E-01
MIMAT0002859_1	3,15E-01	O95232	1,99E-01	HEATR4	2,81E-01
MIMAT0002838	3,01E-01	P11717	1,15E-01	GFRA1	2,72E-01
MIMAT0002835	2,76E-01	Q9Y6E0	6,95E-02	NRP1	2,67E-01
MIMAT0002855	1,31E-01	P07195	3,17E-01	HRH1	2,24E-01
MIMAT0002833	9,78E-02	Q16186	3,61E-01	NCLP1	6,58E-02

- Selected top 30 highest ranked biological features were used for functional analysis starting with finding the most important interactions. We combine RNA interactome: http://www.rna-society.org/rnainter/ with protein interactome: https://string-db.org/ . We use only strongest experimental evidences with highest confidence score (>0.9).
- The three most important connection clusters were selected (Fig. 6). The clusters were supplemented with information from databases dedicated to extracellular microbes. Based on these data, two very significant protein with strong evidence for melanoma were found: IGF2R (protein ID: P11717, ExoCarta ID: ExoCarta_3482) and EFTUD2 (protein ID: Q15029, ExoCarta ID: ExoCarta_9343).
- The interactome study based on top 30 features also showed functional molecular enrichments like telomeric and damaged DNA binding or protein tyrosine kinase related pathways.

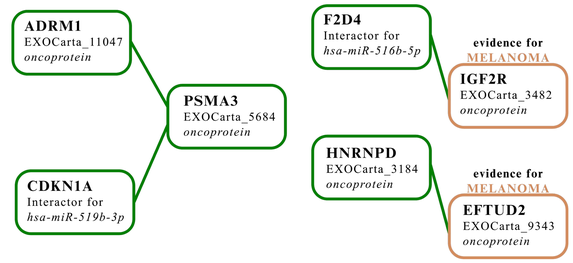


Fig. 5 Interactome analysis. We identify two oncoproteins with strong evidence for extracellular vesicles derived melanoma processes: IGF2R (protein ID: P1171) and EFTUD2 (protein ID: Q15029).

Table 2. Functional enrichments in study network.

Molecular Function (Gene Ontology)	
GO term	description
GO:0042162	telomeric DNA binding
GO:0004714	transmembrane receptor protein tyrosine kinase activity
GO: 0003684	damaged DNA binding
GO:0019955	cytokine binding
GO:0004713	protein tyrosine kinase activity

DISCUSSION

- Proposed method detected important signatures in multi-omics datasets and identified biomarkers candidates like circulating cancer-associated microRNAs and oncoproteins.
- Pipeline ranked significant biological features using sCCA score.
- Method allowed to examine the biological processes related with melanoma progression by selecting molecular signatures that have supporting evidence in databases.
- Method is dedicated to extracellular melanoma biomarker identification but it is elastic and can be adapted to research on other data and cancer types.

REFERENCES

1. Ryan Van Laar, Mitchel Lincoln, and Barton Van Laar. Development and validation of a plasma-based melanoma biomarker suitable for clinical use. British Journal of Cancer, 118(6):857–866, January 2018.
2. Su Yin Lim, Jenny H. Lee, Russell J. Diefenbach, Richard F. Kefford, and Helen Rizos. Liquid biomarkers in melanoma: detection and discovery. Molecular Cancer, 17(1), January 2018.
3. Daniela Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics, 10(3):515-534, July 2009.
4. Theodoulos Rodosthenous, Vahid Shahrezaei, and Marina Evangelou. Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. Bioinformatics, 36(17) May 2020.
5. Helian Feng, Nicholas Mancuso, Alexander Gusev, Arunabha Majumdar, Megan Major, Bogdan Pasaniuc and Peter Kraft. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improve the power of transcriptome-wide association studies. July 2020.
6. Magdalena Surman , Sylwia Kędracka-Krok, Dorota Hoja-Łukowicz, Urszula Jankowska, Anna Drożdż, Ewa Ł. Stępień and Małgorzata Przybyło. Mass Spectrometry-Based Proteomic Characterization of Cutaneous Melanoma Ectosomes Reveals the Presence of Cancer-Related Molecules. International Journal of Molecular Sciences , 21(8), 2934, March 2020
7. Dieudonne van der Meer, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E Francies and Mathew J Garnett. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. Nucleic Acids Research, 47(D1):D923–D929, January 2019.
8. Website: http://www.microvesicles.org/ and http://www.exocarta.org/ . Date of access: 11.11.2020

REST and ZBTB33 in glioma REST ChIP-seq peaks - partners or competitors?

Marta Jardanowska [1,2 *], Bartosz Wojtas[3], Małgorzata Perycz[3], Bożena Kaminska[3], Michal J. Dabrowski [1]

1 Institute of Computer Science, Computational Biology Lab, Polish Academy of Sciences, Warsaw, Poland
2 Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
3 Nencki Institute of Experimental Biology, Warsaw, Poland
* To whom the correspondence should be addressed: marta.jardanowska@ipipan.waw.pl

Introduction

Gliomas are one of the most common and deadly cancers and because of that are intensively studied. At the same time, one of the most promising and still unfathomable issue is the role of the REST transcription factor in brain carcinogenesis processes. On the other hand, the canonical role of REST is regulation of neurogenesis and glial cells development and participation in the neurosecretion process. REST is the main repressor of transcription in neurodegenerative diseases and is associated with the regulation of ion channels and cytoskeletal proteins, but also other transcription factors (TFs). Therefore REST is described as both, activator and repressor of transcription depending on physiological or pathophysiological context. The purpose of this study was to check whether any TF motifs overlap or are in close proximity to REST Transcription Factor Binding Sites (TFBS).

Materials and methods

For REST ChIP-seq peaks from U87 cell line we assigned their summits within the 200bp sequence around the summit (+/- 100bp), using open source bioinformatic tools. For that purpose we used Position Weight Matrices (PWMs) of TF motifs from HOCOMOCO[1] database and 14 additional REST PWMs, mainly from ENCODE[2].
The search of TF motifs was performed using PWMEnrich[3] Bioconductor R package. To identify specific transcription factor binding sites with the corresponding q-values, we used online FIMO[4] tool from MEME Suite 5.0.5. Additionally, peaks were assigned to gene promoters and based on TCGA glioma RNA-seq and in-house REST ChIP-seq data it was specified whether REST represses or activates the expression of the particular genes based on the correlation results, negative or positive, respectively.

Results

Rank	Target	PWM	P-value
1.5	KAISO_HUMAN.H11MO.0.A		0
1.5	KAISO_HUMAN.H11MO.1.A		0
3	KAISO_HUMAN.H11MO.2.A		1.39e-139
4	E2F4_HUMAN.H11MO.1.A		3.8e-82
5	REST_HUMAN.H11MO.0.A		4.85e-78
6	REST_m14_known_matrix		1.67e-77
7	FEV_HUMAN.H11MO.0.B		3.52e-73
8	ZBED1_HUMAN.H11MO.0.D		2.47e-72
9	E2F5_HUMAN.H11MO.0.B		1.44e-68
10	ZBT14_HUMAN.H11MO.0.C		4.7e-67
11	E2F2_HUMAN.H11MO.0.B		1e-65
12	E2F1_HUMAN.H11MO.0.A		3.24e-64
13	E2F4_HUMAN.H11MO.0.A		6.39e-64
14	SP1_HUMAN.H11MO.1.A		7.2e-64
15	REST_m4_GM12878_encode		2.36e-62

Fig. 1 Ranking of TOP15 motifs for REST activated genes.

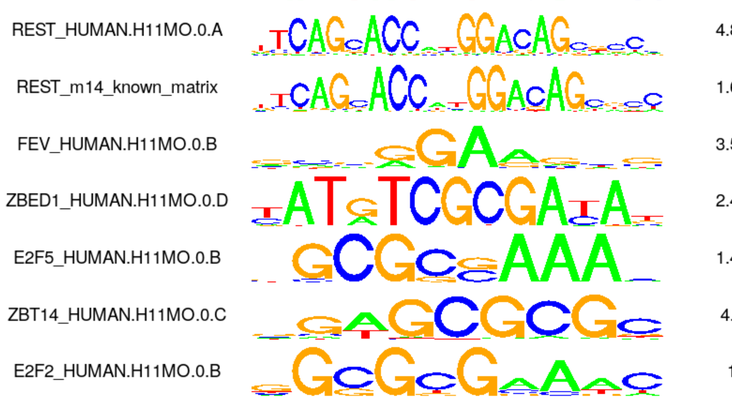


Fig. 2 Clustering of TF motifs characteristic for REST activated genes, REST repressed genes and common motifs based on DNA sequences.

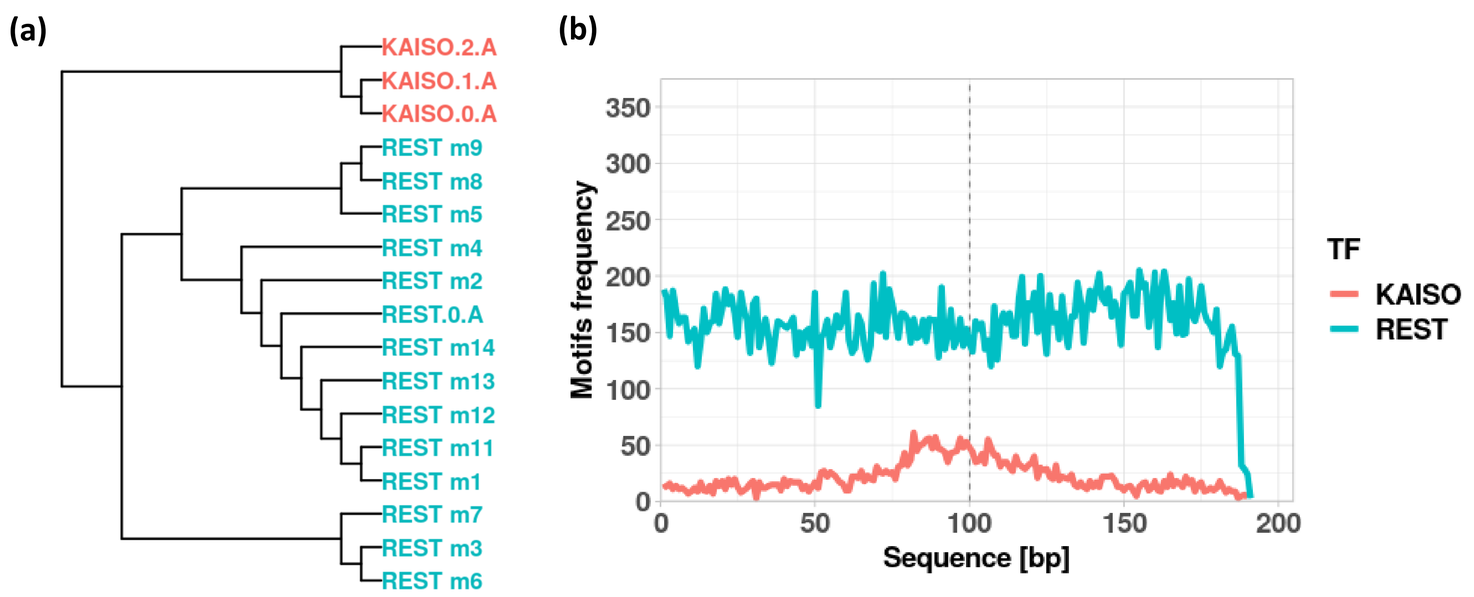


Fig. 3 REST and KAISO motifs (a) clustering based on DNA sequences (b) occurrence dependent from the localization in the activated genes sequences.

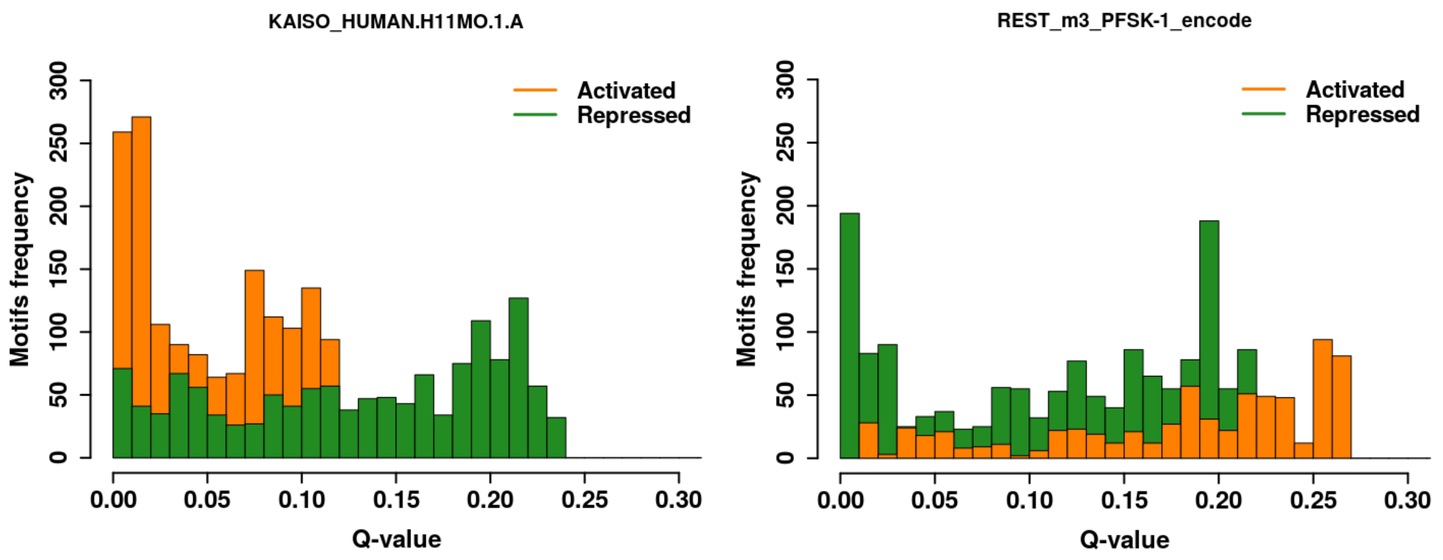


Fig. 4 Q-value and frequency relation for selected KAISO and REST motif for REST ChIP-seq peaks for activated and repressed genes.

Conclusions

- We identified 202 TF motifs (12 REST motifs) in the 200bp sequences surrounding REST ChIP-seq peaks for the activated genes sequences and 237 TF (14 REST motifs) motifs for the repressed genes sequences. Top places in the motifs ranking for the REST activated genes were occupied by the KAISO motifs, characteristic for the ZBTB33 transcription factor. (Fig. 1)
- Motifs characteristic for activated (n = 21) and repressed (n = 56) genes clustered separately. (Fig. 2)
- Analysis of the nucleotide sequences of the identified motifs showed that they significantly differed between REST and ZBTB33, meaning that the co-occurrence of these TF motifs within the examined sequences was not due to sequence similarity. (Fig.3a)
- We observed that in the REST activated genes, KAISO motifs were significantly more frequent in the proximity to the peak summits than in the rest of the examined 200bp sequence. (Fig. 3b)
- ZBTB33 motifs occurred with higher frequency and lower q-value in the REST activated genes, while the majority of REST motifs were within the repressed genes. (Fig. 4)
- These results may suggest that while the main REST role may be repressive, its role within the activated genes promoters can be at least co-dependent on ZBTB33.

Supported by National Science Centre, Poland grant number DEC- 2015/16/W/NZ2/00314.

References

[1] I. V Kulakovskiy, I. E Vorontsov, I. S Yevshin, R. N Sharipov, A. D Fedorova, E. I Rumynskiy, Y. A Medvedeva, A. Magana-Mora, V. B Bajic, D. A Papatsenko, F. A Kolpakov, V. J Makeev, *HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis,* Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D252–D259 **[2]** https://www.encodeproject.org **[3]** R. Stojnic, D . Diez, *PWMEnrich: PWM enrichment analysis. R package version 4.18.0,* 2018. **[4]** C. E. Grant, T. L. Bailey and W. Stafford Noble, *FIMO: Scanning for occurrences of a given motif*, Bioinformatics 27(7):1017–1018, 2011.

DNA methylation patterns of active enhancers specific for *pilocytic astrocytoma* and Higher Grade Glioma samples

Agata Dziedzic[1], Marta Jardanowska[1], Marcin Grynberg[2] and Michał J Dąbrowski[1]

[1] Institute of Computer Science, Polish Academy of Sciences, Poland
[2] Institute of Biochemistry and Biophysics, Polish Academy of Sciences

Aim

- To study molecular differences in enhancers of different glioma grades: *pilocytic astrocytoma* and Higher Grade Glioma.
- To detect specific methylation sites in Transcription Factor motifs responsible for changes of its transcription factors binding affinity and as a result - changes of target gene expression.

Fig.1. Schematic representation of target gene expression regulation via enhancer.

Materials & Methods

Experiment	Type of data	Analysis maperformed on data
Chip-seq for H3K27ac	Genome coordinates of active enhancers	Motif search
Bisulphite seq	Methylation level per single cytosine (~3.5 mln sites per sample)	**DM** cytosines calling
RNA-seq	Read counts per gene	**DE** genes calling

Tab.1. Analysis performed on three layers of biological information for the set of 7 PA and 10 HGG samples.

Results

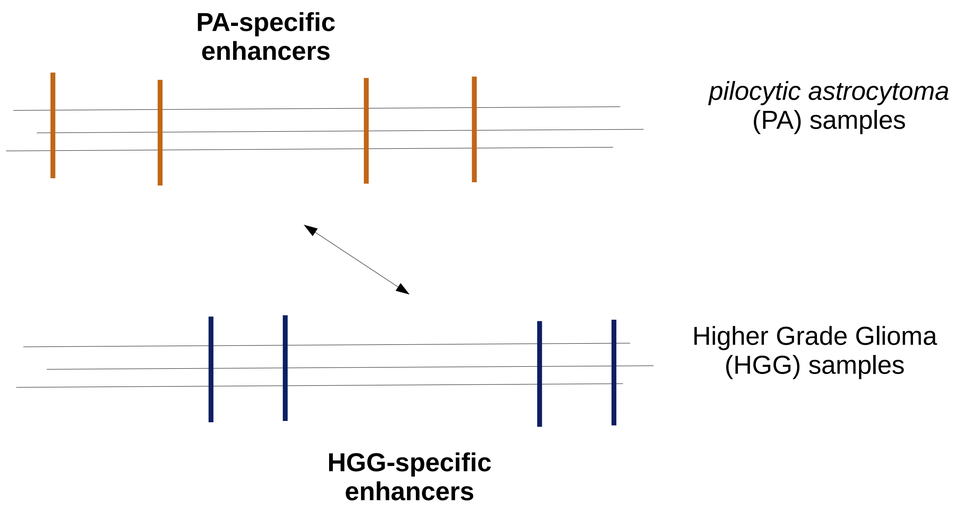


Fig.2. Schematic representation of enhancers methylation levels comparisones.

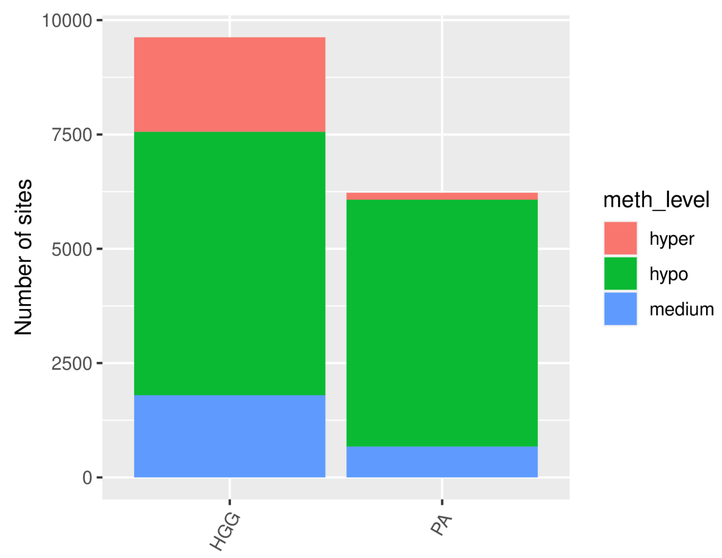


Fig.3. Number of CpG sites devided into three ranges of methylation level.There are more hypermethylated sites in HGG-spec. Enhancers comparing to PA-spec. enhancers (X-squared = 1309.9, df = 1, p-value < 2.2e-16).

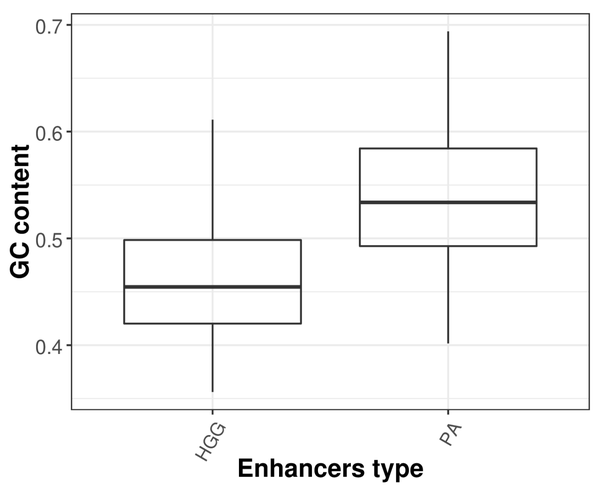


Fig.4. Mean GC content was 46 % for HGG and 54 % for PA – difference was statiscally important (HGG n = 124, PA n = 114, Mann-Whitney U test: p-value = 4.292992e-17, W = 11528)

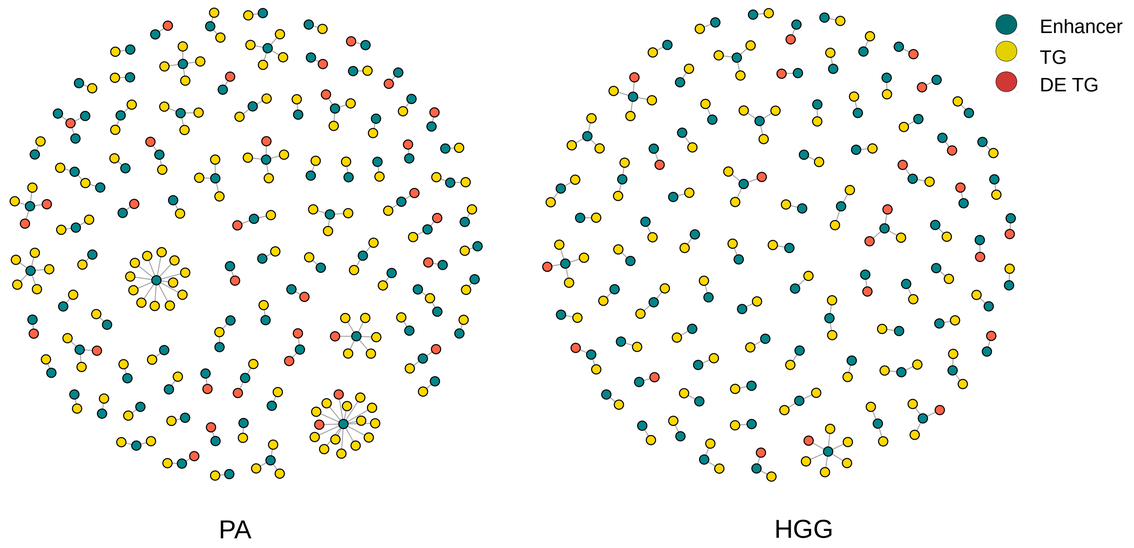


Fig.5. PA: 92 enhancers targeting 161 TG (32 DE). HGG: 84 enhancers targeting 120 TG (22 DE).

Conclusions

- HGG-specific enhancers had **lower frequency of guanine and cytosine nucleotides** then PA-specific enhancers and higher global DNA methylation level.
- Methylation pattern of **14 TF motifs was confirmed to be consequently hypermethylated in HGG** compared to PA samples and all of this motifs were found in at least one ehnancer with differentially expressed target gene.
- These results indicate specific TF motifs whose **methylation may have an influence on regulation of TG expression** and therefore contribute to gliomagenesis.

a)

b)

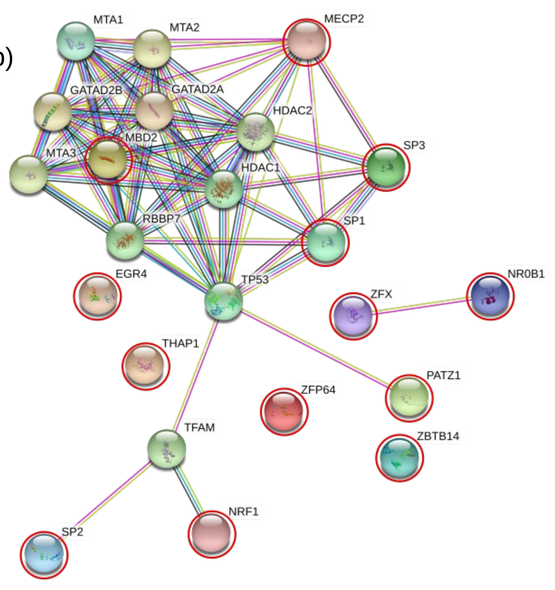


Fig.6. a) Selected 14 TF motifs & their nucleotide sequence; b) Graph of 13 TFs together with additional proteins they interact with.

Funding:
National Science Center, SYMFONIA3: DEC 2015/16/W/NZ2/00314

Novel alternative splicing events detection in human genome with Spladder

Agata Muszyńska[1,2],Paweł Łabaj[1,3]

[1] Małopolska Centre of Biotechnology UJ, Kraków, Poland
[2] Institute of Automatic Control, Electonics and Computer Science, Silesian University of Technology, Gliwice, Poland
[3]Chair of Bioifnormatics RG, Boku University, Vienna, Austria

Abstract

Although human genome is widely studied since many years its complexity remains not fully understood. One of the mechanisms that stands for that is alternative splicing, which is a process of joining exons in multiple ways, so that novel mRNA and, in fact, novel proteins are produced. Currently we are not fully aware of all of the splicing events that might be present in a given genome. One of the tools that provides the possibility to investigate that is Spladder. It builds an augmented splicing graph, based on current annotation and than expands it with novel events. Currently Spladder supports detecting six different types of such events. We used Spladder software on data from SEQC consortium project [1][2].

We investigated 3 samples (A- mixture of 10 different cancer cell lines, B- healthy individual and C- A and B samples mixed in 1:1 ratio) run on different RNA targeting panels, as well as on whole transcriptome sequencing data obtained with two protocols- ribo-depletion and polyA selection. Preliminary results show that there is a fraction of genes containing novel events, which seems to be cancer or sample specific, but majority is the same irrespective of sample. It seems that the current gene model can be extended by this data. Spladder also revealed that the fraction of intron retention events is higher for whole transcriptome sequencing data than for targeted approach and is higher for ribo-depletion protocol than for polyA selection, what is expected after comparing sample processing and library preparation for these approaches.

These results show that there is still a lot of work ahead of us to fully describe our genome but at the same time that Spladder might be a good tool, not only for that challenge, but also for others like detecting cancer specific events.

Study design

4 x A, B, C → Panel1, Panel2, Panel3, WTS PolyA, WTS riboZero

Samples

A: mixture od 10 different cancer cel lines
B: healthy individual
C: A and B mixed in 1:1 ratio

Panels

Panel1: around 1000 genes
Panel2: around 2000 genes
Panel3: around 2000 genes

Spladder pipeline

Annotattion + Alignment Data → Augumented Splicing Graph → Detected Splice Events → Quantified Splice Events → Differentail Analysis/ sQTL Tests

Events detected

- Alternative 5 prime end
- Alternative 3 prime end
- Exon skip
- Intron retention
- Mutually exclusive exons
- Multiple exon skip

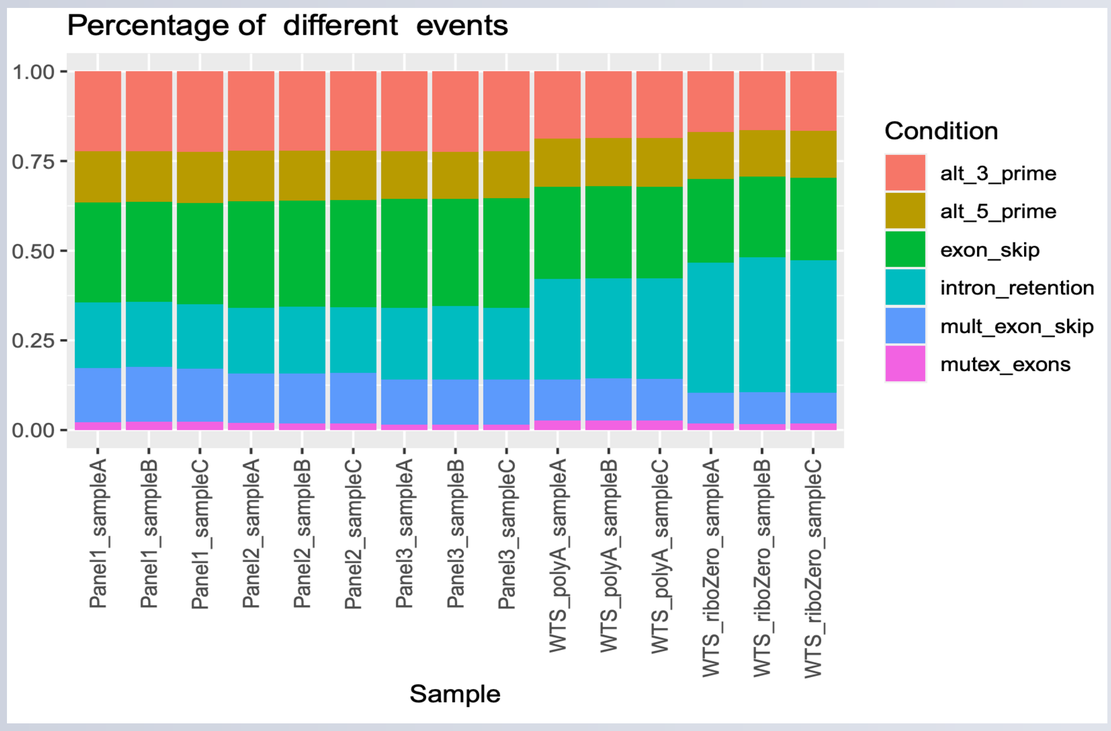
PSI value

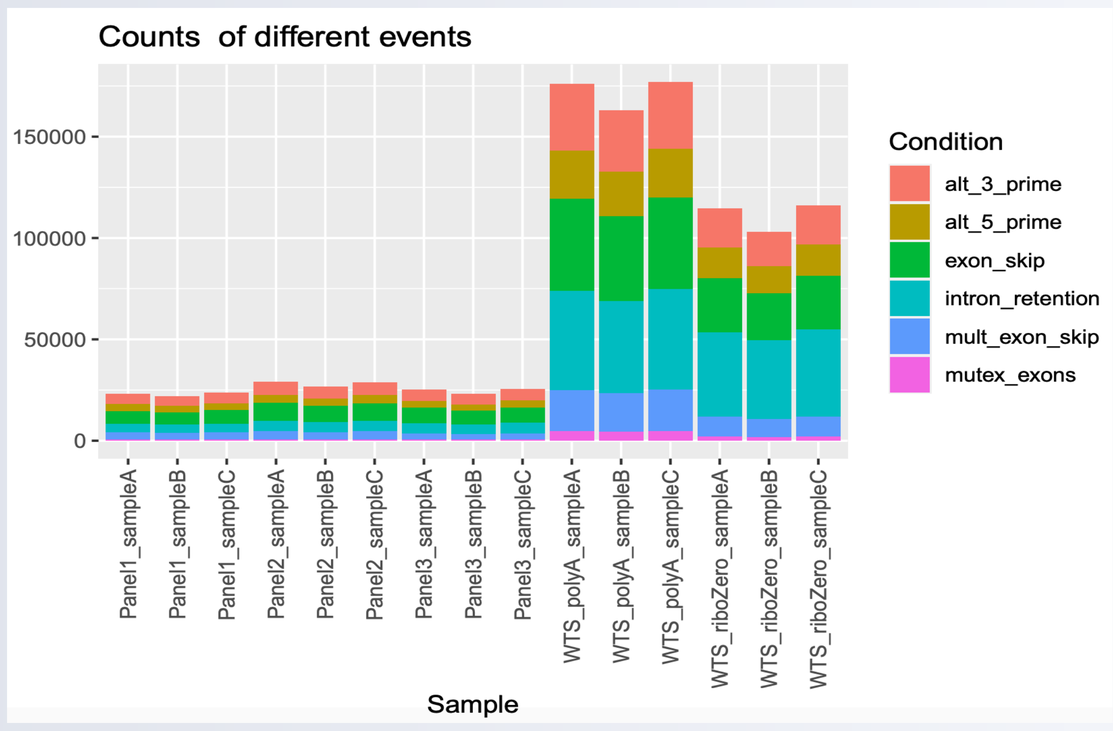
$$PSI = \frac{IR}{IR+ER}$$

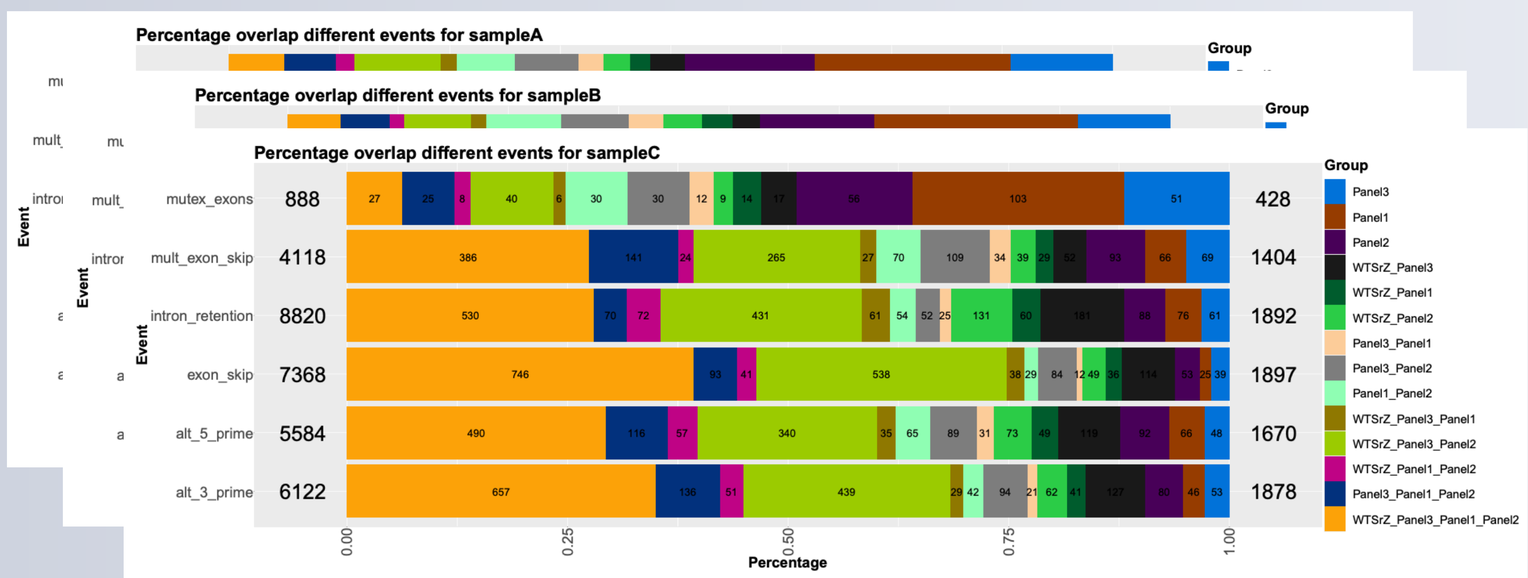
Valid events

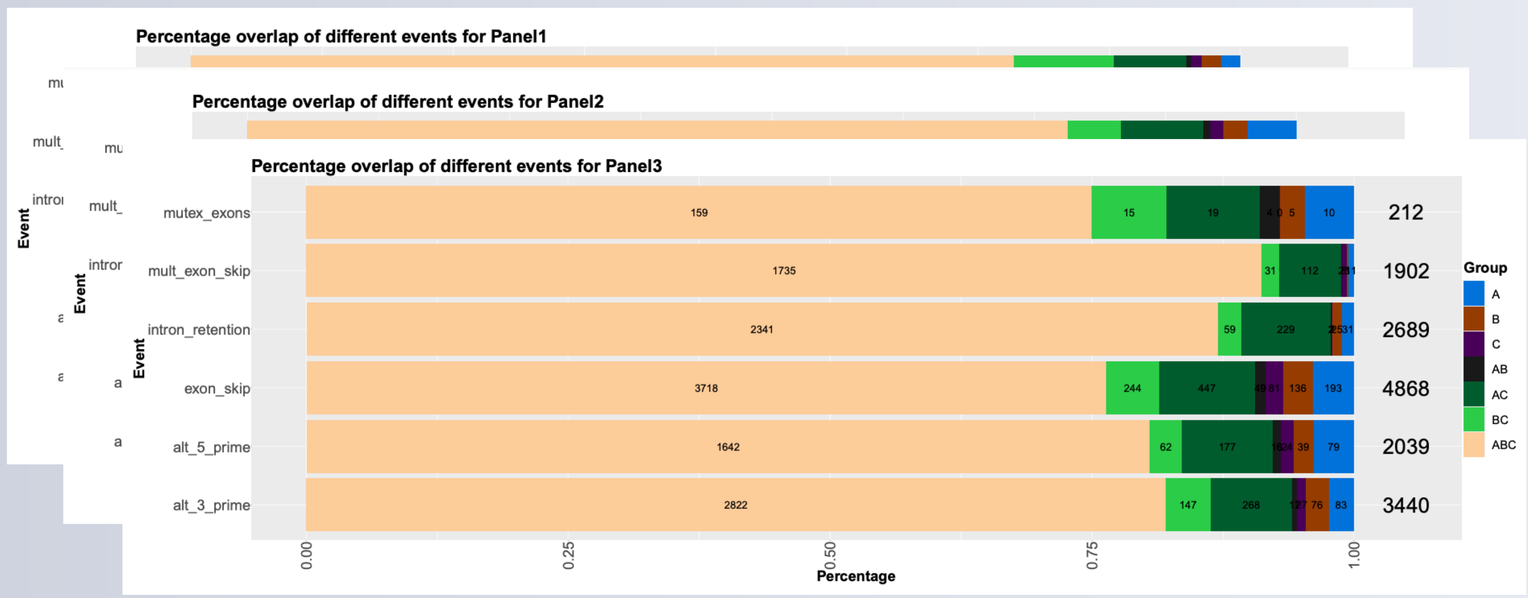
Events with PSI value above 0.1 for all 4 replicates

Results









Conclusions

- We were able to detect all splicing events in our data, among which the most prevalent were exon skip and intron retention, whereas the least- mutually exclusive exons.
- Although there were some events, which seems to be cancer or sample specific, majority is common- this suggest that current gene model might be expanded.
- Intron retention events occur more often in whole transcriptome sequencing data, than in any of the panels and also often in ribo-depletion than in polyA. This reflects differences in library preparation for these approaches.
- WTS with polyA protocol detects more events than riboZero.

References

[1] Su, Zu et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Consortium. Nature Biotechnology 32, 903-914(2014)
[2] Kahles A, Ong CS, Zhong Y, Rätsch G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics. 2016 Jun 15;32(12):1840-7.

BioMetaNet: Meta-Network model for human lymphoblastoid cell lines representing complete biological interactome

Kaustav Sengupta[1,2], Michał Denkiewicz[1,3], Anup Kumar Halder[4], Subhadip Basu[4], Dariusz Plewczyński[1,3,5]

Email : k.sengupta@cent.uw.edu.pl, m.denkiewicz@cent.uw.edu.pl, dariuszplewczynski@cent.uw.edu.pl

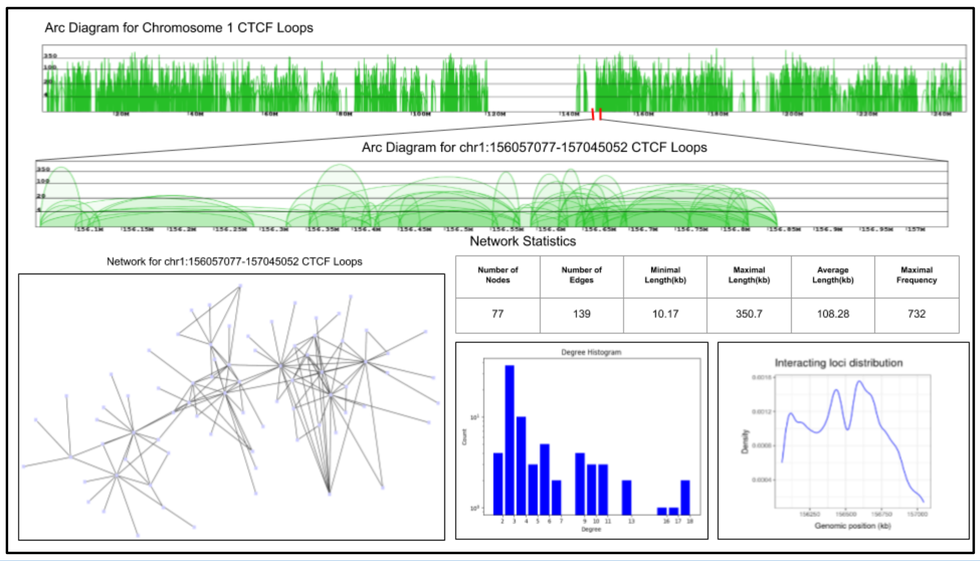
1. *Center of New Technologies, University of Warsaw, Poland*
2. *Faculty of Mathematics Informatics and Mechanics, University of Warsaw, Poland*
3. *Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland*
4. *Department of Computer Science and Engineering, Jadavpur University, Kolkata, India*
5. *Computer Science Department, University of California, Davis, CA, United States*

Introduction

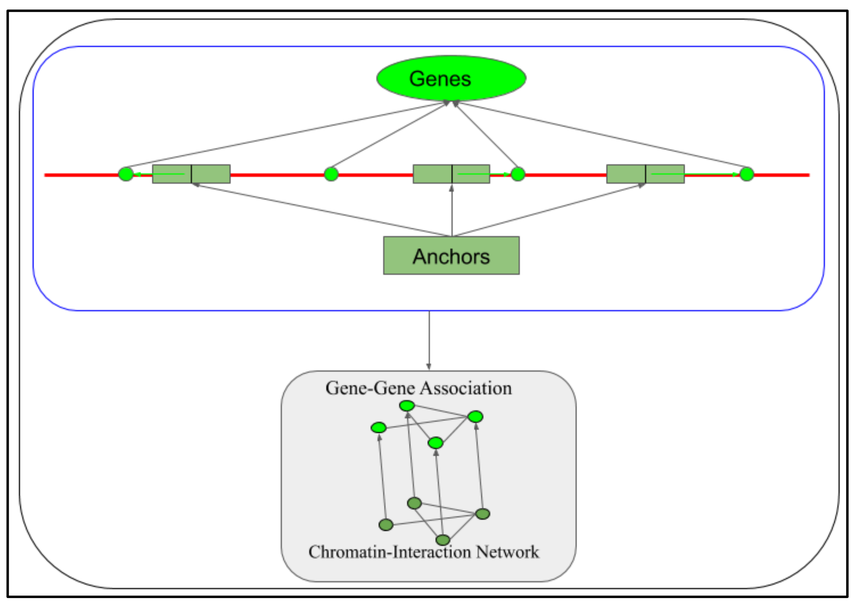
In recent years, with the development of high throughput methods, researchers obtained access to a vast array of biomolecular interaction data. Most of these biological data can be represented as networks or graphs. Thus, network analysis is becoming a powerful tool for modeling biological systems. We propose a meta-network representation of the complete map of DNA pairwise interactions for human lymphoblastoid cell lines combined with information about encoded proteins and metabolic pathways. In a single graph (meta-network) we integrate multiple biological networks, namely, Chromatin Interaction Network (CIN), Genomic Association Network (GAN), Protein-Protein Interaction Networks (PIN), Gene Ontology (GO) terms, and metabolic pathways. Thus cheating the meta-network connecting 3D chromatin interaction to functionality.

Methods

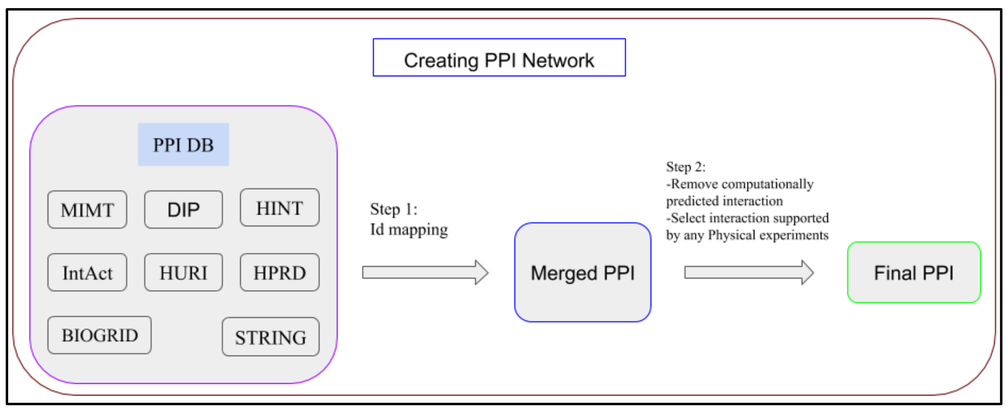
Chromatin Interaction Networks (CIN)

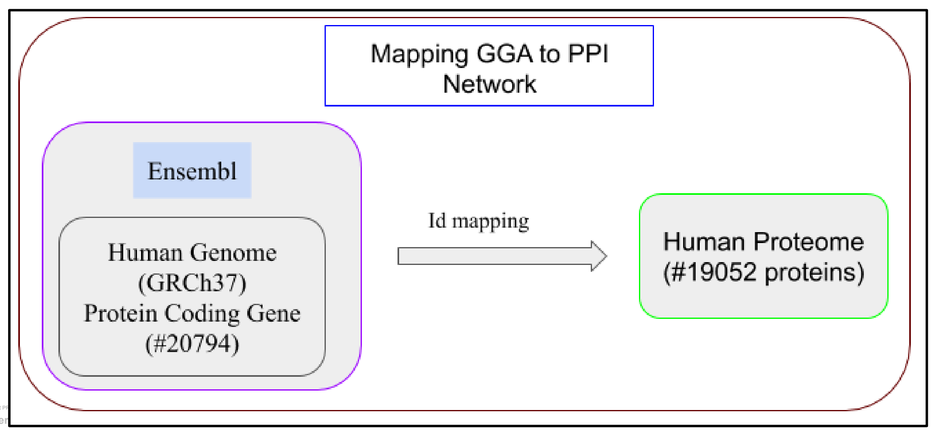


Gene-Gene Association Networks (GAN)

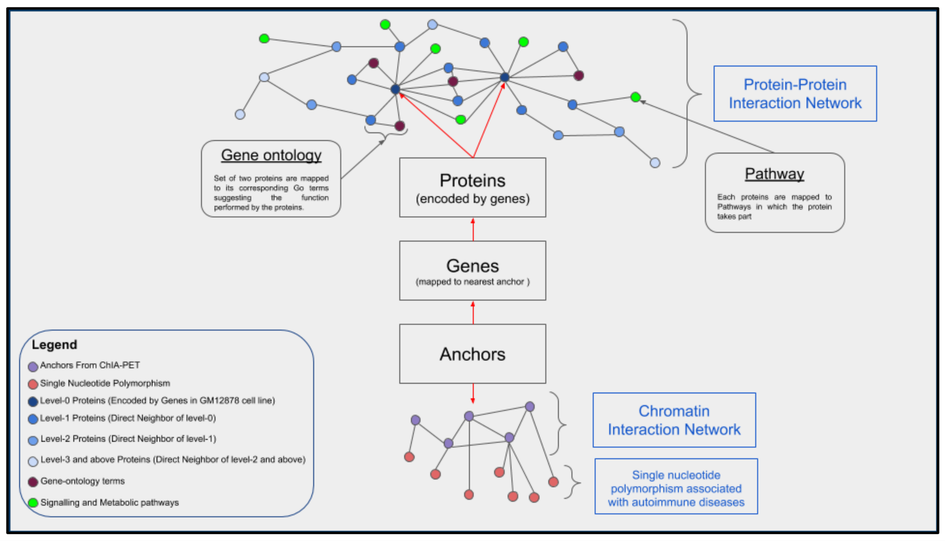


Protein-Protein Interaction Network (PIN)

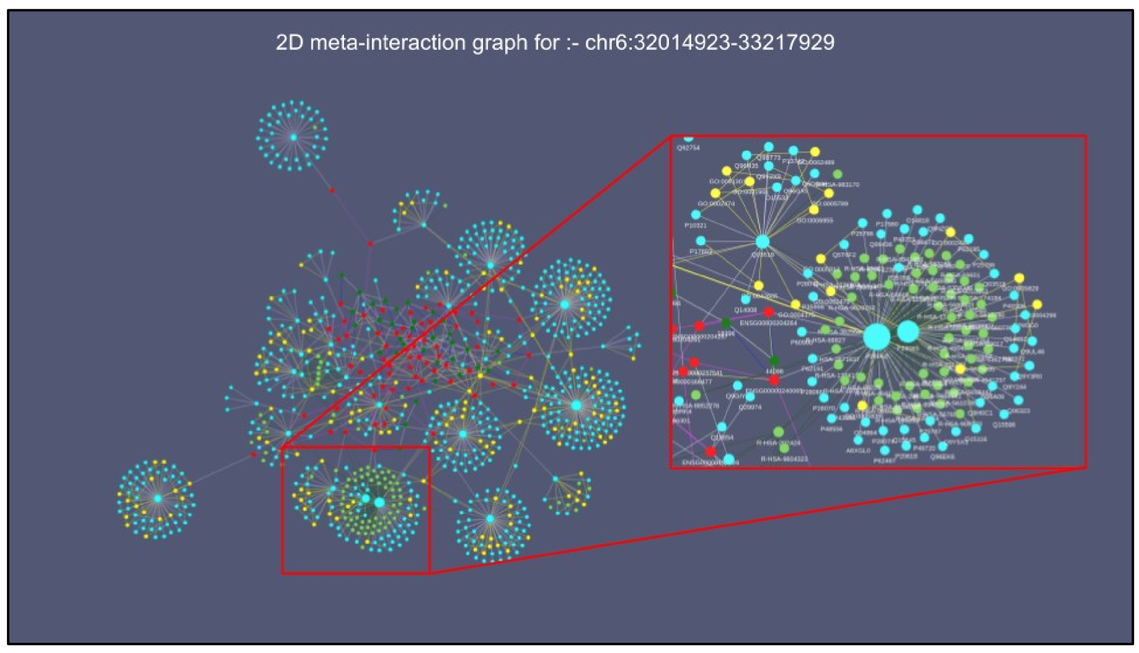




Gene-Ontology (GO), Pathways and Single Nucleotide Polymorphism (SNPs) Mapping



Results



We analyzed the meta-network and found proteins P28062 and P28065 encoded by genes PSMB8 and PSMB9, present in location chr6:32014923-33217929, share around 60 pathways which are higher than the average concentration of metabolic pathways shared between two proteins.

Critically, the genes PSMB8 and PSMB9 are also connected by proximity with HLA genes and TAP genes using the proteomic networks. The protein P28062 and P28065 are two of the 17 essential subunits (alpha subunits 1-7, constitutive beta subunits 1-7, and inducible subunits including beta1i, beta2i, beta5i) that contribute to the complete assembly of the 20S proteasome complex.

Conclusion

The meta-network can give us insigths into the interactions between genomic, proteomic and chromatin (structural) networks. In particular: the proteins P28062 and P28065, due to a large numer of shared pathways and the proximity of their encoding genes to the known autoimmune-related genes, can be critical for studies of autoimmune disease. Moreover, the presence of essential genes and proteins, the study of genome rearrangements in from of structural variants in this region can give us novel insights into the study of autoimmune diseases.

In conclusion, our meta-network model can be instrumental in getting a complete picture of biological functionality linked with 3D chromatin interactions. The network can also be extended to incorporate Structural Variants which can provide an idea of how functionality varies with the larger genome rearrangement.

Acknowledgement

This work has been supported by Polish National Science Centre (2019/35/O/ST6/02484), Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to DP). The work was co-supported by European Commission Horizon 2020 Marie Skłodowska-Curie ITN Enhpathy grant 'Molecular Basis of Human enhanceropathies'; and National Institute of Health USA 4DNucleome grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation".

References

1.Anup Kumar Halder, Michał Denkiewicz, Kaustav Sengupta, Subhadip Basu, Dariusz Plewczynski,Aggregated network centrality shows non-random structure of genomic and proteomic networks,Methods,2019,ISSN 1046-2023,https://doi.org/10.1016/j.ymeth.2019.11.006.

2.Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, Michalski P, Piecuch E, Wang P, Wang D, Tian SZ, Penrad-Mobayed M, Sachs LM, Ruan X, Wei CL, Liu ET, Wilczynski GM, Plewczynski D, Li G, Ruan Y. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell. 2015 Dec 17;163(7):1611-27. doi: 10.1016/j.cell.2015.11.024. Epub 2015 Dec 10. PMID: 26686651; PMCID: PMC4734140.

3.Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., … Clamp, M. (2004). An overview of Ensembl. Genome research, 14(5), 925–928. https://doi.org/10.1101/gr.1860604 UniProt Consortium (2008). The universal protein resource (UniProt). Nucleic acids research, 36(Database issue), D190–D195. https://doi.org/10.1093/nar/gkm895

MASTERMIND: The Best Linear Model to Accurately Determine Monoisotopic Mass

Piotr Radziński, Michał Startek, Anna Gambin

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Abstract

Nowadays, monoisotopic mass is used to be an important feature in top-down proteomics. Knowing the exact monoisotopic mass enables precise and quick protein identification in large protein databases. However, only in spectra of small molecules monoisotopic peak is visible, for bigger molecules position of the peak have to be predicted. By improving prediction of the peak, we contribute to more accurate identification of molecules, what is crucial in fields such as chemistry and medicine. In this work we present MASTERMIND algorithm, that is a two-step procedure to predict monoisotopic mass for proteins with 8-400 kDa mass range. The first step is to approximate monoisotopic mass by linear regression based on average mass and variance of a given spectrum. The second step rounds linear prediction to the closest point which is reliable to be a peak in the spectrum. For 96.6% of proteins, prediction error is below 0.2 ppm, what is approx. 30% better than in recently proposed MIND tool. Our algorithm was implemented in python, data analysis was performed in R. Proteins to learn the model comes from Uniprot database, their theoretical spectra were calculated by use of IsoSpec structure calculator.

MASTERMIND algorithm

I. Initial prediction

At the beginning, we calculate initial prediction of monoisotopic mass, by use of spectrum's average mass and variance:

$$\hat{M}_{\text{mono}} = \beta_0 + \beta_{\text{avg}} \cdot M_{\text{avg}} + \beta_{\text{var}} \cdot M_{\text{var}}.$$

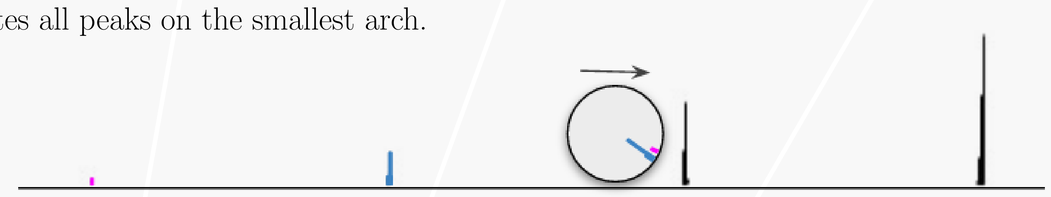
Prediction is not good enough for practical use, however, for 96.6% proteins prediction error is smaller than 0.5 Da, what is crucial for our algorithm. We want to round initial prediction to closest point on the grid

$$\mathcal{W}(\zeta, \Delta) = \{\zeta n + \Delta : n \in \mathbb{N}\},$$

which determine where peaks that are not visible on spectrum should be.

II. Estimation of the grid step ζ

Grid step ζ, is equivalent to circumference of circle, that rolled through spectrum concentrates all peaks on the smallest arch.



Mathematicaly, we have

$$\zeta_0 = \operatorname*{argmin}_{\zeta \in \mathbb{R}} \operatorname{Var} P_\zeta(\mathcal{S}),$$

where

$$P_\zeta(z) = \frac{\zeta}{2\pi i} \log\left[\exp\left(\frac{2\pi i z}{\zeta} - i\operatorname{Im}\left[\log\left(\sum_{p \in \mathcal{S}} p^{\text{prob}} \cdot \exp\left(2\pi i p^{\text{mass}}/\zeta\right)\right)\right]\right)\right].$$

To avoid long calculation for each protein, we trained linear model that gives ζ_0 based on protein average mass

$$\hat{\zeta} = \gamma_0 + \gamma_{\text{avg}} \cdot M_{\text{avg}}.$$

III. Estimation of the grid shift Δ

When we have $\hat{\zeta}$, we calculate grid shift, to fit the grid into spectrum

$$\hat{\Delta} = \operatorname*{argmin}_{\Delta \in [0, \hat{\zeta}]} \sum_{p \in \mathcal{S}} p^{\text{prob}} \cdot \min_{w \in \mathcal{W}(\hat{\zeta}, \Delta)} |p^{\text{mass}} - w| = \operatorname{Re}\left[\frac{\hat{\zeta}}{2\pi i} \log\left(\sum_{p \in \mathcal{S}} p^{\text{prob}} \cdot \exp\left(2\pi i p^{\text{mass}}/\hat{\zeta}\right)\right)\right].$$

IV. Final prediction

To obtain final prediction, we round initial prediction to closest point on the fitted grid, and apply slight correction

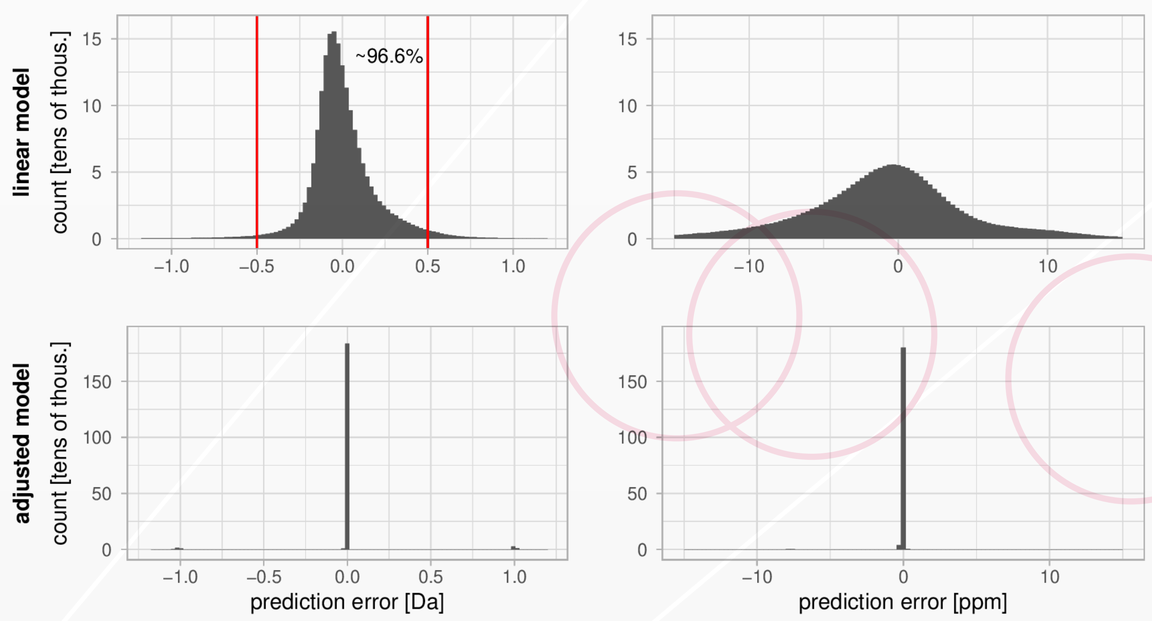
$$\hat{\hat{M}}_{\text{mono}} = \operatorname*{argmin}_{w \in \mathcal{W}(\hat{\zeta}, \hat{\Delta})} \left|w - \hat{M}_{\text{mono}}\right| + \lambda \cdot \hat{M}_{\text{mono}}.$$

Data & Tools

- Chemical formulas used to train models comes from **Uniprot** database;
- Their spectra were calculated by **IsoSpec** structure calculator;
- MASTERMIND algorithm was implemented in python, data analysis was performed in R. To calibrate linear models we used 10-fold cross-validation;

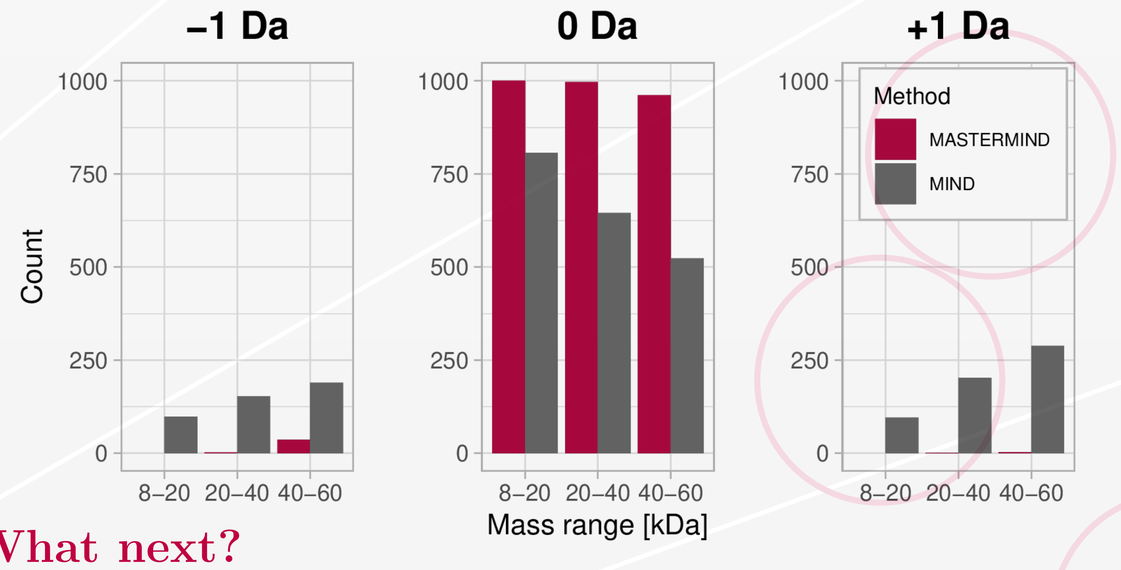
This research is supported by the Polish National Science Center grants 2018/29/B/ST6/00681 and 2017/26/D/ST6/00304.

How rounding improves prediction?



Comparison with MIND

- MIND prediction is based on the most-abundant peak, MASTERMIND is based on average peak and variance;
- MASTERMIND is close to true monoisotopic mass in **96.6%** versus 66.5% for MIND;
- MASTERMIND is better in every mass range it was compared with MIND, and is trained on bigger mass range;
- MASTERMIND loses accuracy fast, when spectrum resolution is getting worse;



What next?

- Elaborate a method, that finds average mass and variance regardless of spectrum resolution;
- Test MASTERMIND on real spectra;

References

Mateusz K. Łącki, Michał Startek, Dirk Valkenborg, Anna Gambin, 2017, *IsoSpec: Hyperfast Fine Structure Calculator*, Analytical Chemistry, vol. 89(6).

Frederik Lermyte *et al.*, 2019, *MIND: A Double-Linear Model To Accurately Determine Monoisotopic Precursor Mass in High-Resolution Top-Down Proteomics*, Analytical Chemistry, vol. 91(15).

Contact us!

pmradzinski@mimuw.edu.pl michal.startek@mimuw.edu.pl

A novel approach to search for interdigitated proteins - unusual domain swapped topology

Mateusz Skłodowski[1], Joanna M. Macnar[1,2], Dominik Gront[1]

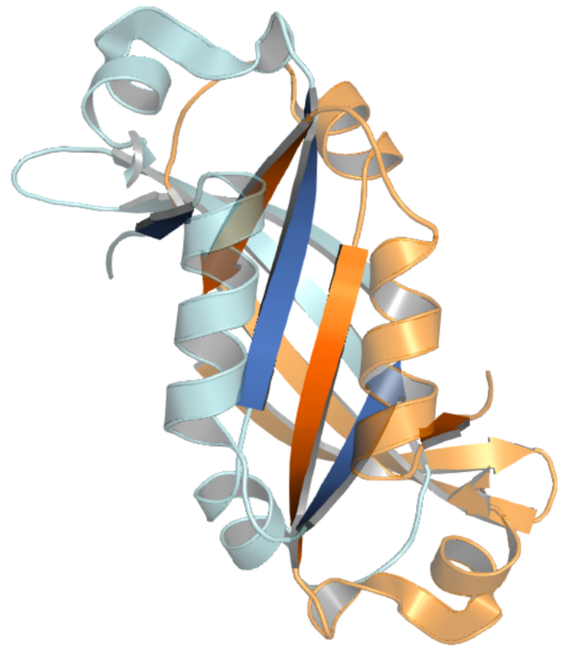
1 Faculty of Chemistry, University of Warsaw, Pasteura 1, Warsaw, Poland

2 College of Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Stefana Banacha 2C, Warsaw, Poland

Introduction

Interdigiteted motives are specific cases of protein domain swapping [1] including secondary structures from two different polypetide chains creating a single beta sheet. Additionally, interdigiteded stuctures consist of interchangeable occurrence of beta strands from different chains in beta-sheet.
In our work we search Protein Data Bank[2] for proteins that have the motive described earlier. For this task we used BioShell [3], [4] and graph theory. For further analysis, a group of proteins with the longest six-element beta sheet was adopted, in which their structural, sequential and functional similarity was studied.

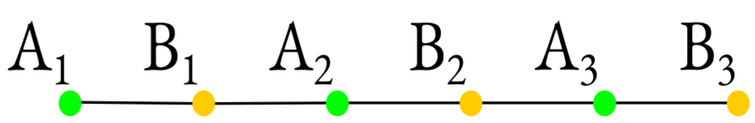
Interdigited protein example



Protein with six-element interdigitated beta sheet - AF2331[5]. *Darker colors represent secondary structures involve in motive.*

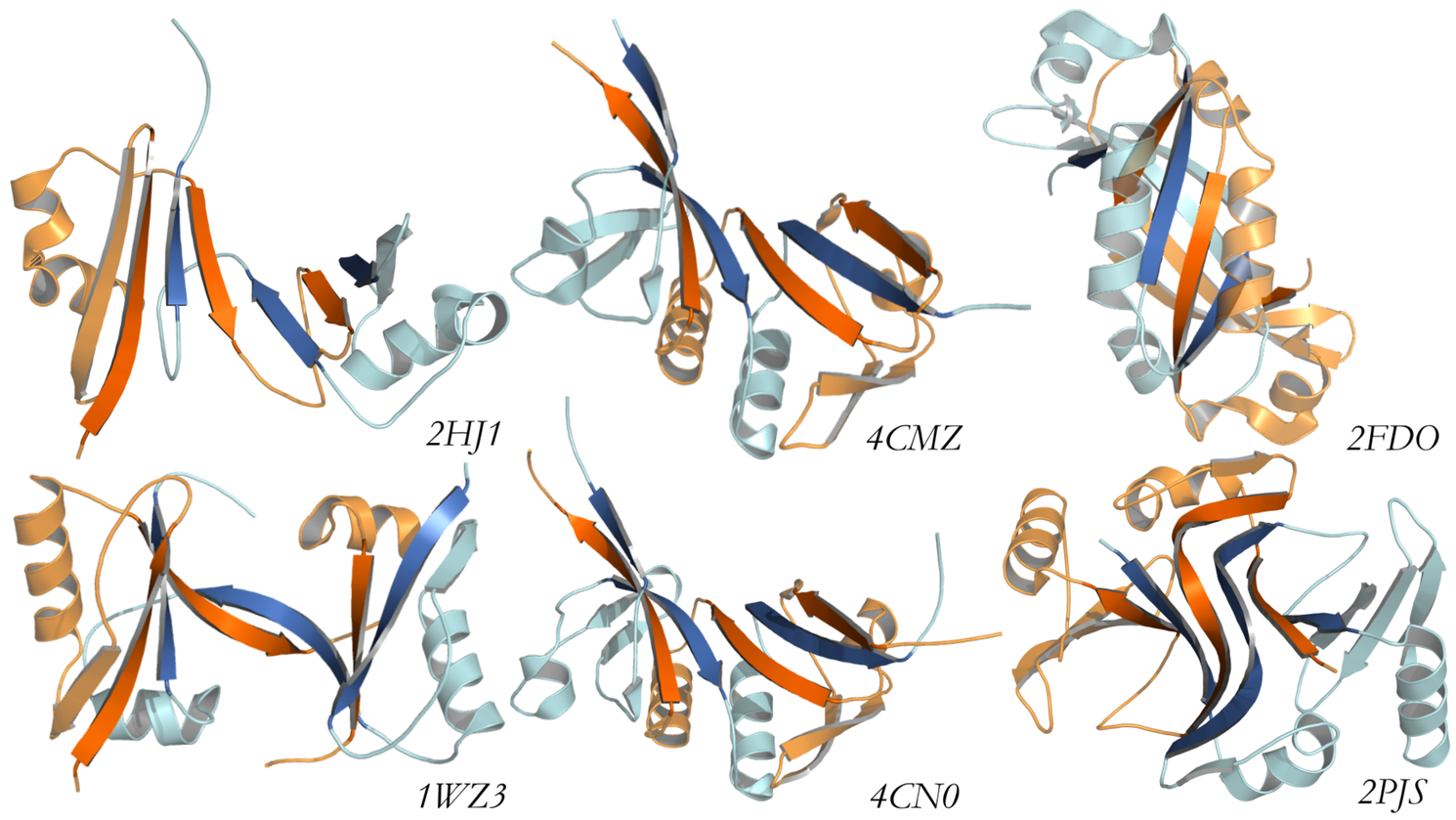
Graph theory application

In our project we applied graph theory to describe interactions between beta strands. For this work we state that each vertex of a graph is single beta strand. If the stands create a hydrogen bond, we assume an edge of the graph between them. To check if the beta sheet is interdigitated, we color the graph depending on the assignment of a beta strand to its protein chain. At this point, the depth-first search algorithm is used to gather information if interacting strands belong to different chains. The information collected also enables analysis in relation to the length of the motif.



Schematic application of the algorithm on the example of protein AF2331

Interdigited protein - examined group of proteins



Six proteins with six-element interdigitated beta sheet obtained by analysis

Conclusions

- Our approach has allowed us to identify new interdigitated proteins.
- We identify six proteins with six-element intedigiteted beta sheet.
- All of them are homodimers and their length does not extend beyond 120 aminoacids.
- We also identified a group of proteins with a smaller beta card. However, more research is needed in this subject.
- Another interesting topic is proteins, in which interdigitated beta sheets are formed by interactions of secondary elements from more than two chains.

Basic informations about examined group of proteins

Protein (PDB id.)	Year of publication	Sequence length [aa]	Homodimer?	Original organism	Crystal system	Resolution of measurement [Å]
1WZ3	2005	96	Yes	*Arabidopsis thaliana*	C2	1,8
2HJ1	2006	97	Yes	*Haemophilus influenzae*	C2	2,1
2PJS	2007	119	Yes	*Agrobacterium fabrum*	C2	1,85
4CN0	2014	97	Yes	*Homo sapiens*	C2	1,75
4CMZ	2014	92	Yes	*Homo sapiens*	C2	2,7
2FDO	2005	94	Yes	*Archaeoglobus fulgidus*	C2	2,4

References

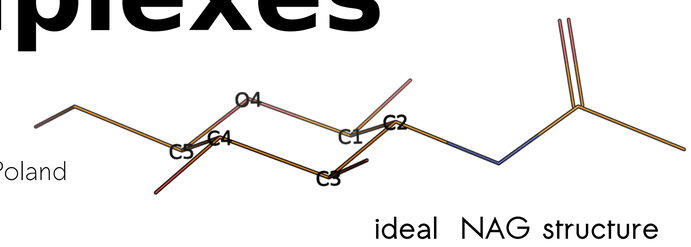
1 M. J. Bennett, S. Choe, and D. Eisenberg, “Refined structure of dimeric diphtheria toxin at 2.0 Å resolution,” Protein Sci., 1994, doi: 10.1002/pro.5560030911.
2 H. M. Berman et al., “The Protein Data Bank,” Nucleic Acids Research. 2000, doi: 10.1093/nar/28.1.235
3 D. Gront and A. Kolinski, “BioShell - A package of tools for structural biology computations,” Bioinformatics, 2006, doi: 10.1093/bioinformatics/btk037.
4 J. M. Macnar, N. A. Szulc, J. D. Kryś, A. E. Badaczewska-Dawid, and D. Gront, “Bioshell 3.0: Library for processing structural biology data,” Biomolecules, 2020, doi: 10.3390/biom10030461.
5 S. Wang et al., “The crystal structure of the AF2331 protein from Archaeoglobus fulgidus DSM 4304 forms an unusual interdigitated dimer with a new type of α + β fold,” Protein Sci., 2009, doi: 10.1002/pro.251.

BioShell software can efectivelly analyze rings in small compounds

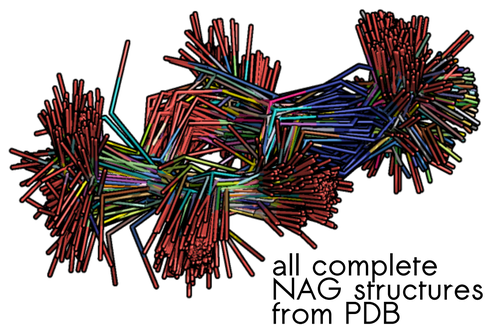
Analysis of small molecules parameters in ligand-protein complexes

Joanna M. Macnar [1, 2], *Wladek Minor* [3], *Dominik Gront* [1]

1. Faculty of Chemistry, Biological and Chemical Research Center, University of Warsaw, Warsaw, Poland
2. College of Inter Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Poland
3. Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, USA



ideal NAG structure



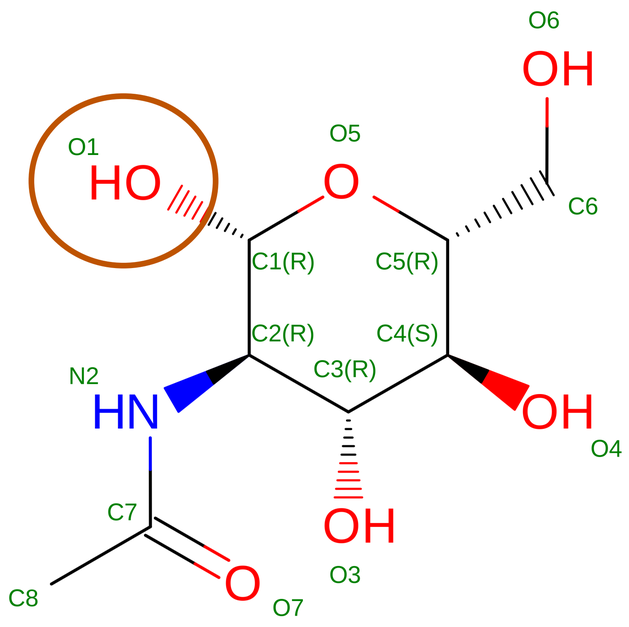
all complete NAG structures from PDB

Intro

Structural information about ligand-macromolecule complexes is critical for biomedical sciences. This analysis will lead to an improved library of restraint parameters and subsequently better refinement of ligand-protein complexes which contain 2-acetamido-2-deoxy-beta-D-glucopyranose (NAG).

Methods

We chose the most common small molecule from PDB which participates in a biological pathway and has one aliphatic ring. We found 5673 deposits and used BioShell package to analize their geometry.



Results

We analyzed 271 structures, thet were complete and determined by X-ray crystallography out of 5673 deposites that contained NAG ligands. A a referenceing X-ray crystallography. As a refence structure the `ideal.sdf` file form PDB was used.

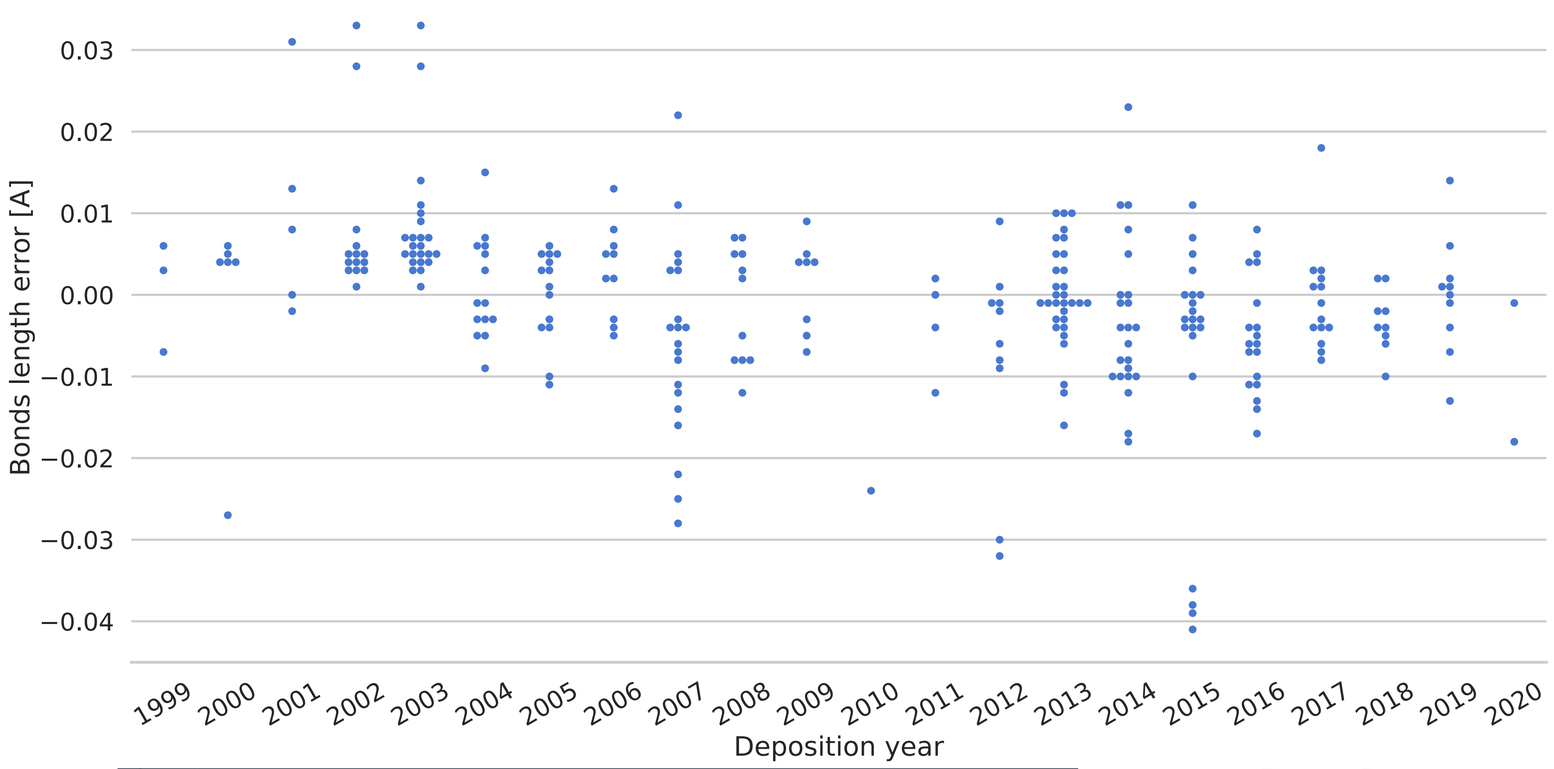


Figure 1

The average deviation of the bond lenght comparing to the ideal structure

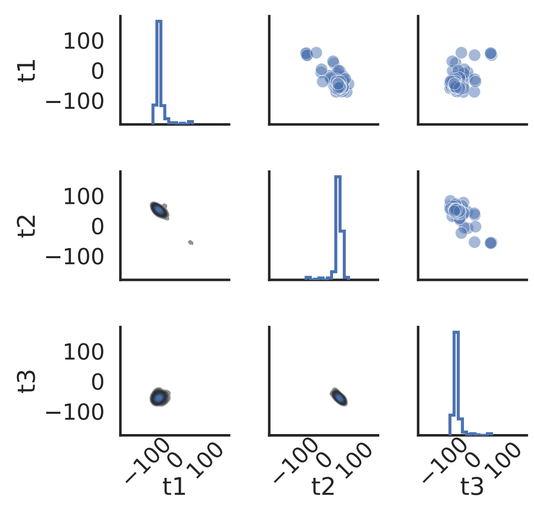


Figure 2

Scatter, KDE plots and histograms showing three subsequent torsion angles from a six member ring.

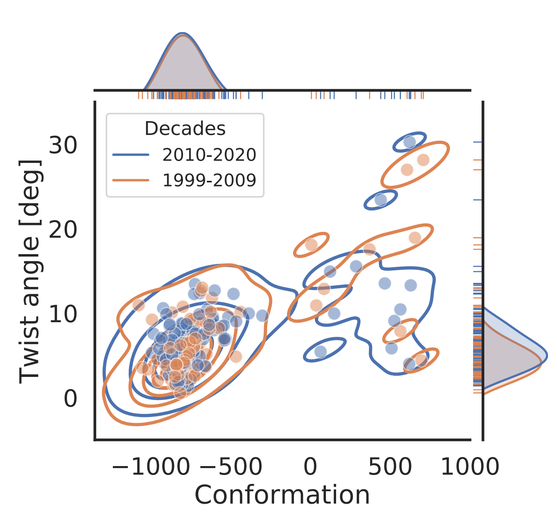


Figure 3

Conformational analysis of NAG rings showing improvementdeposit quality over time.

Conclusions

More reaserch is needed:

- The quality of NAG structures has remained roughly constant for 20 years
- Correlation to electron density map should be included for better analysis
- Missing ligand atoms are a common problem in deposits
- BioShell is a suitable package for ligand geometry analysis.

NATIONAL SCIENCE CENTRE POLAND

This research was funded by the National Science Centre (Poland) Grant No.2019/35/N/ST6/04459

UNIVERSITY OF WARSAW

Take a picture to **read more**

Cost-sensitive feature selection - information theory approach

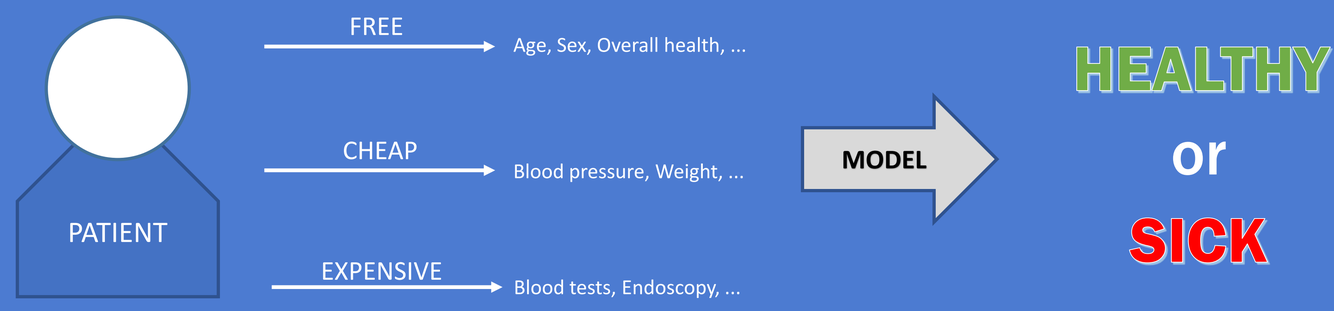
Tomasz Klonecki - Institute of Computer Science, Polish Academy of Sciences

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.

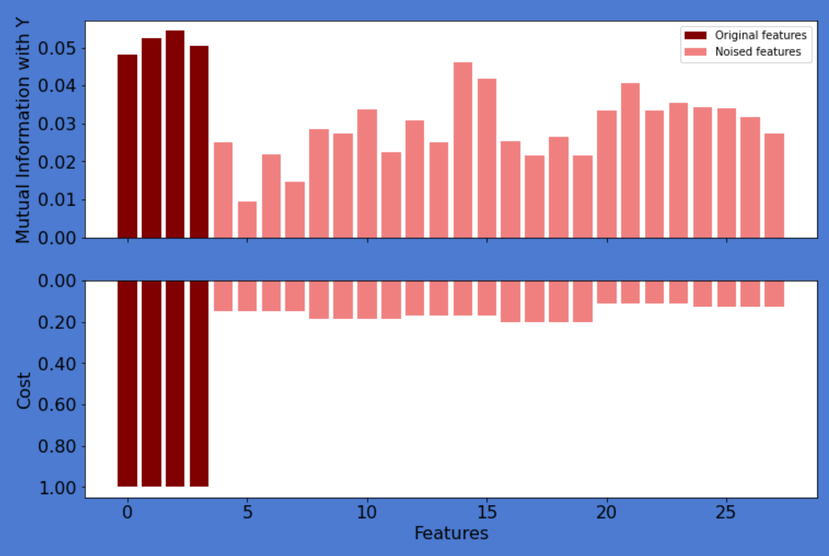
The existing feature selection methods usually ignore costs associated with the considered features. The goal of cost- sensitive feature selection is to select a subset of features which allow to predict the target variable (e.g. occurrence of the diseases) successfully within the assumed budget.

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.

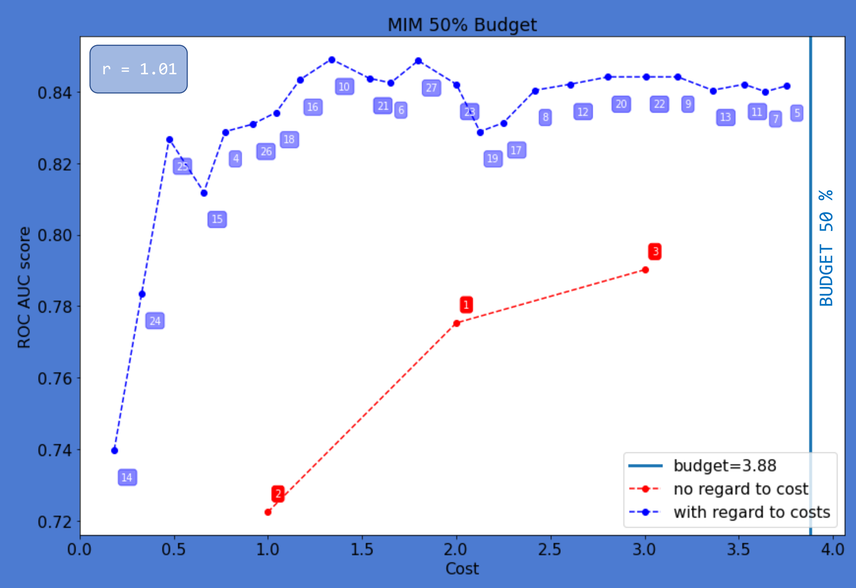


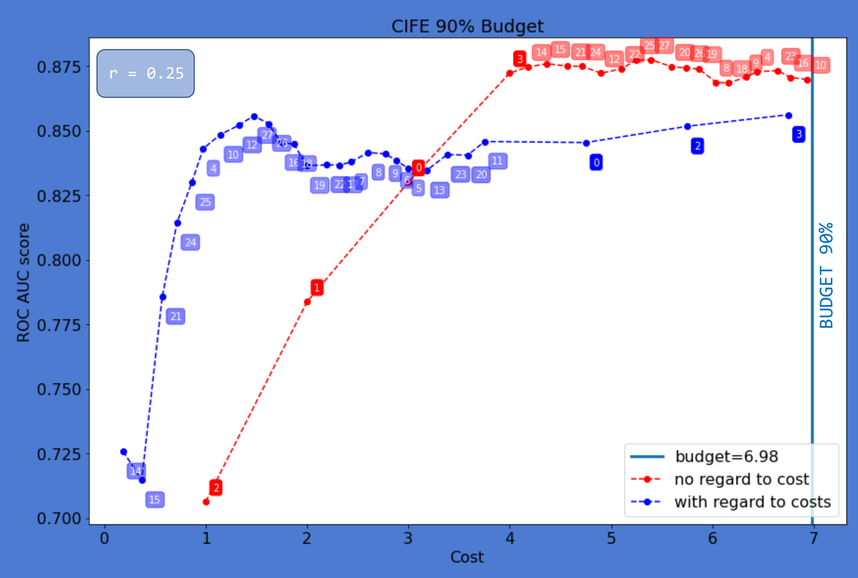
Artificial Dataset

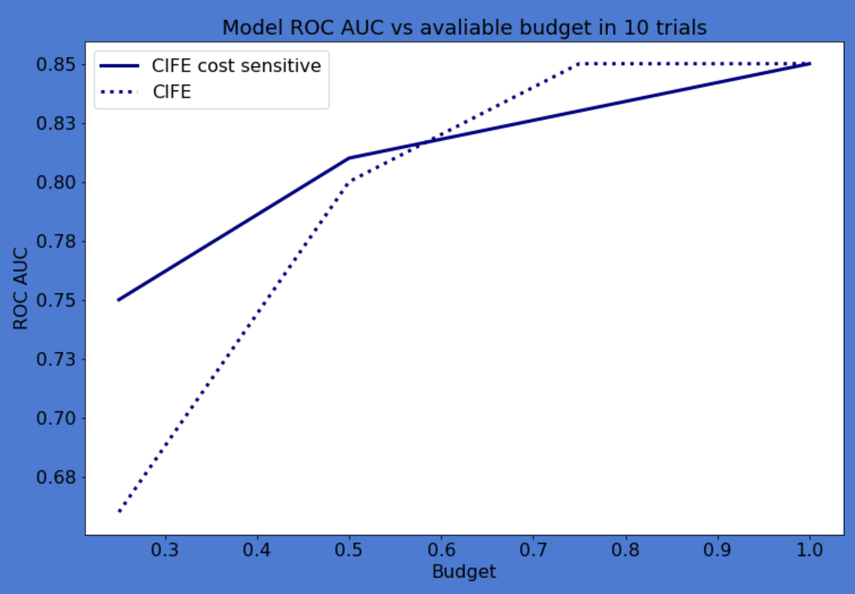
1. Generate original features from normal distribution $X_1, X_2, \ldots, X_p \sim N(0,1)$
2. Generate target variable Y based on $X_1, X_2, \ldots, X_p$ with binomial distribution.
3. Generate noised features $X_i' = X_i + E_j$ where $E_j \sim N(0, \sigma_j)$.
4. Assign cost to each feature $c_i = 1$ 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

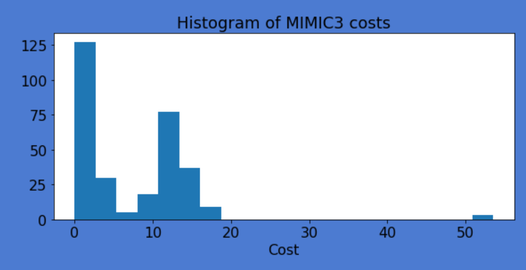
MIMIC III is one of the most popular medical datasets in the world. For experiments we use data of 6500 patients.

Types of features:

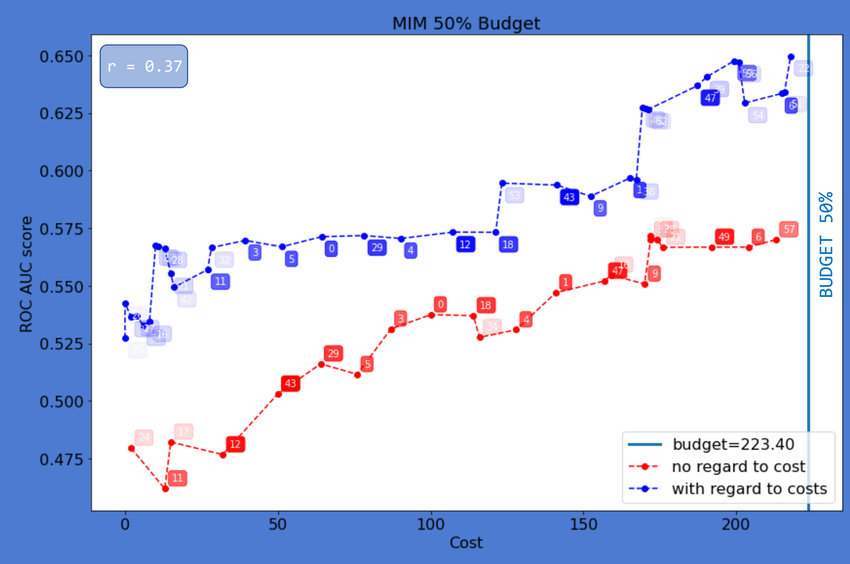
- basic patient information (Age, gender, ...)
- basic medical tests (HR, Blood pressure, ...)
- advanced medical tests (Blood tests, Urine tests, ...)

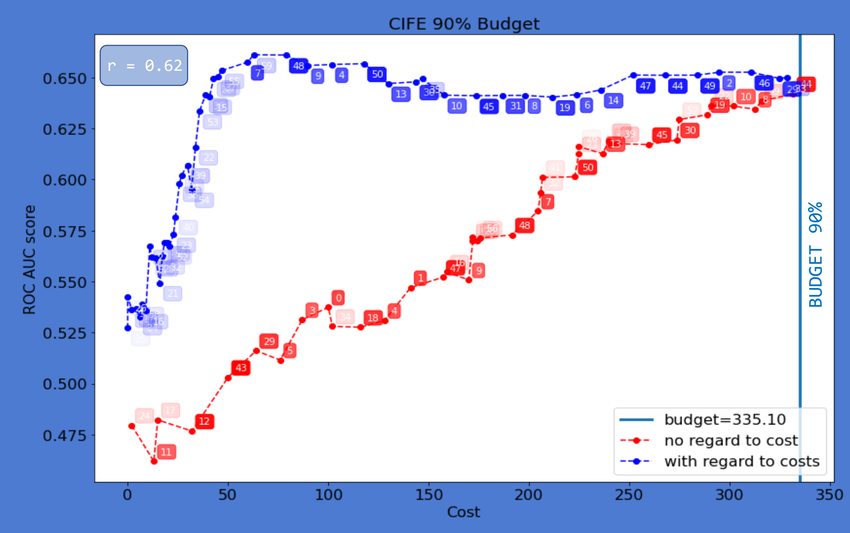
Target variable:

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

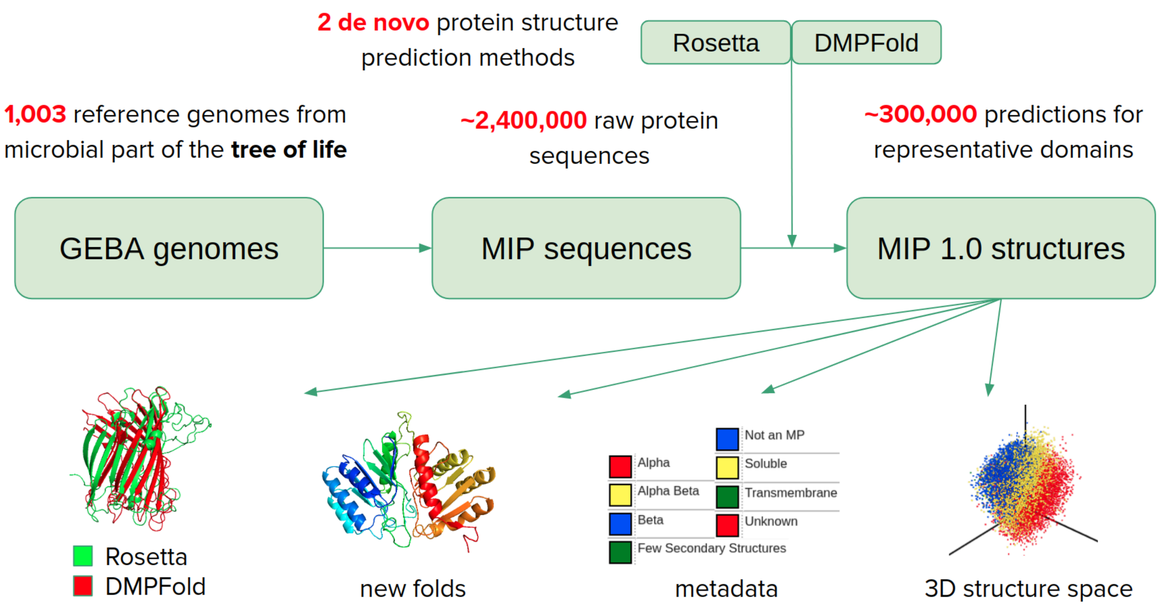
Paweł Szczerbiak[1], Douglas Renfrew[2], Julia Koehler Leman[2], Daniel Berenberg[2],
Chris Chandler[2], Vladimir Gligorijević[2], Richard Bonneau[2], Tomasz Kościółek[1]

[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.

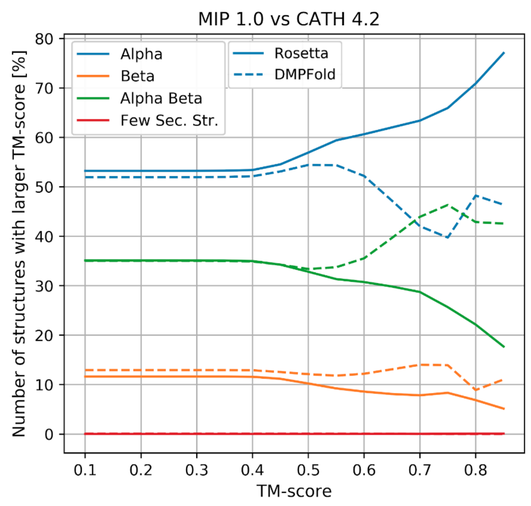
Rosetta vs DMPFold

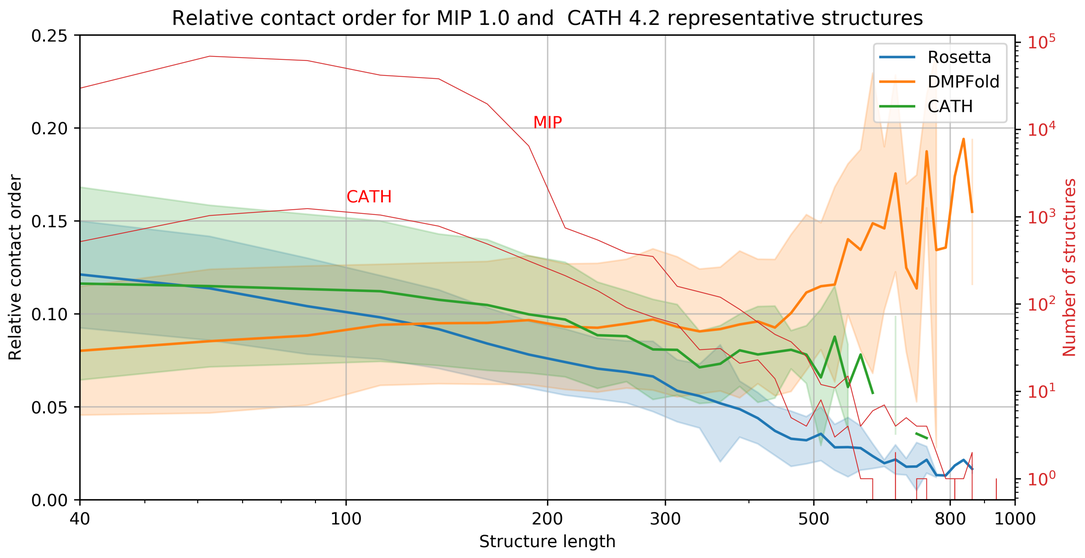
Rosetta

- 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**

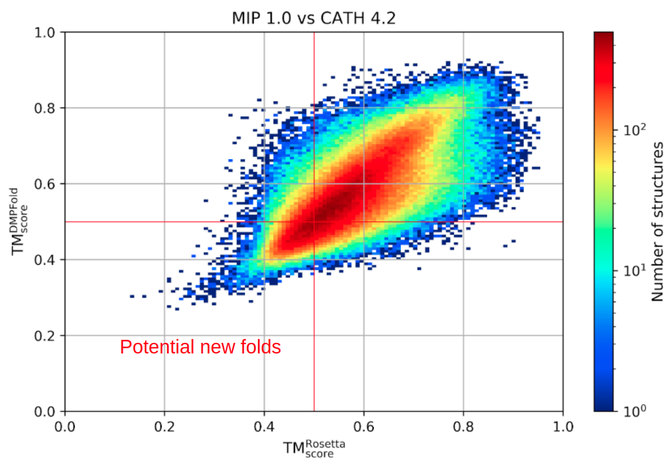
TM-score:

]0 – 0.5[different folds

[0.5 – 0.1] same folds

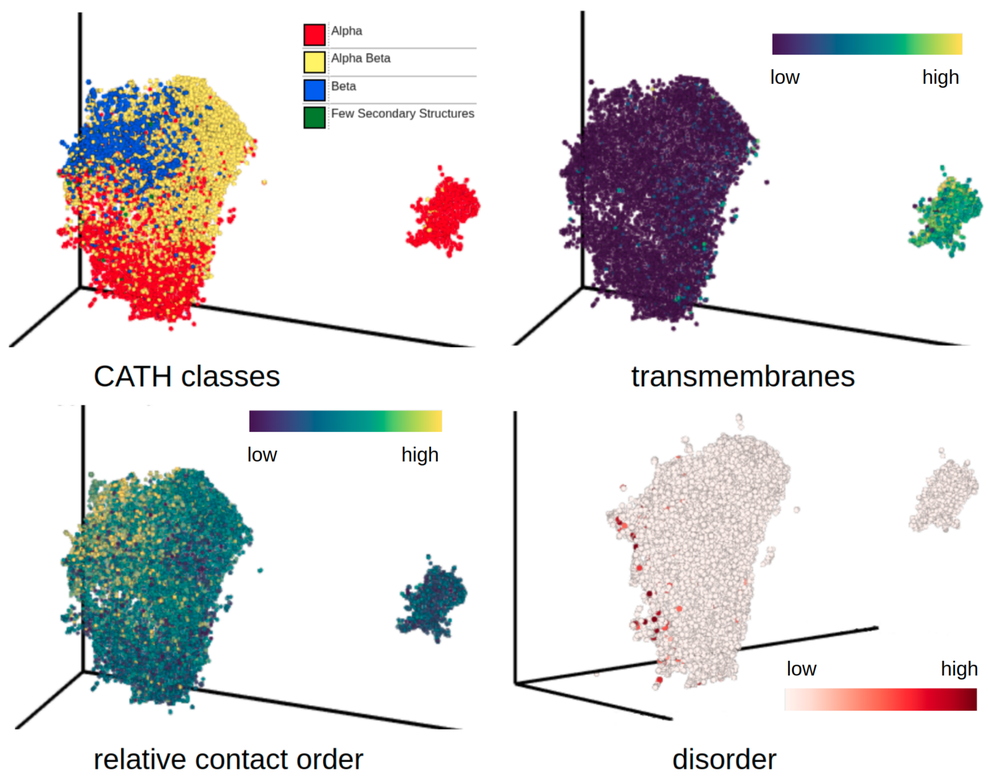
TM-score = **0.3**

TM-score = **0.7**



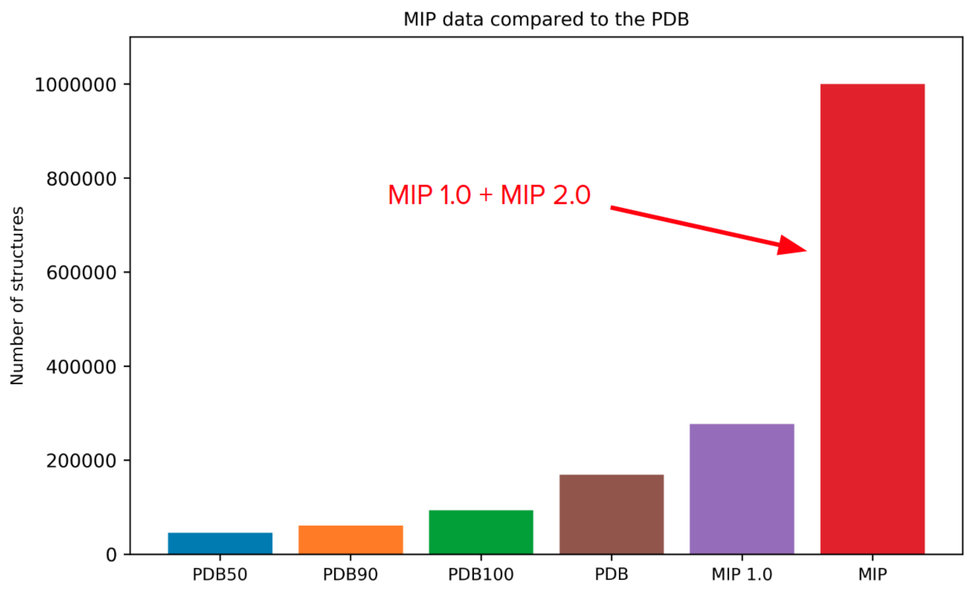
Structure space visualization

- 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

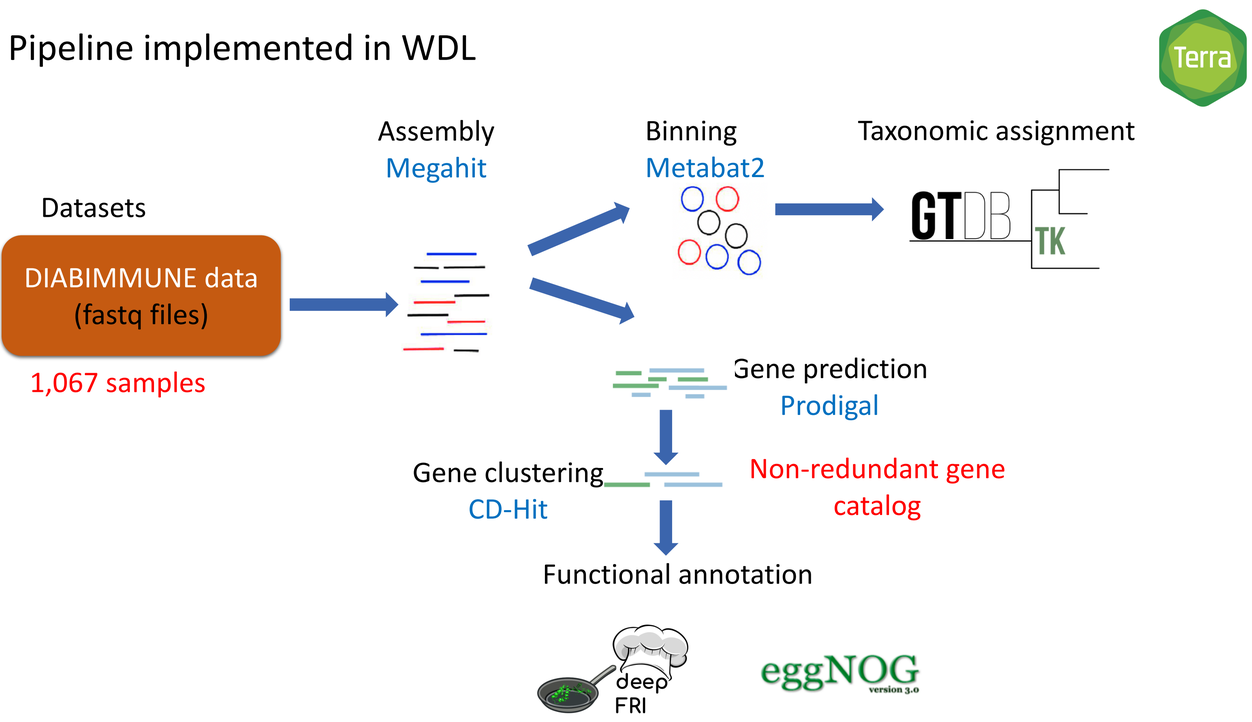
- The human gut microbiome contributes to the development and persistence of diseases such as type-1 diabetes (T1D), ulcerative colitis, obesity and many others.
- Exact mechanisms of how gut microbiota influences health remains poorly understood.
- Only 50% of microbial protein-coding genes may be functionally 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

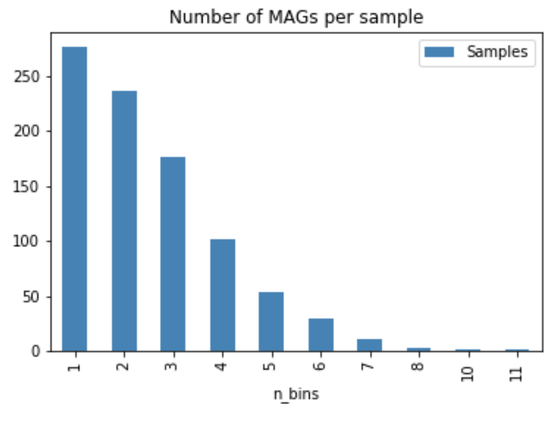
Pipeline implemented in WDL

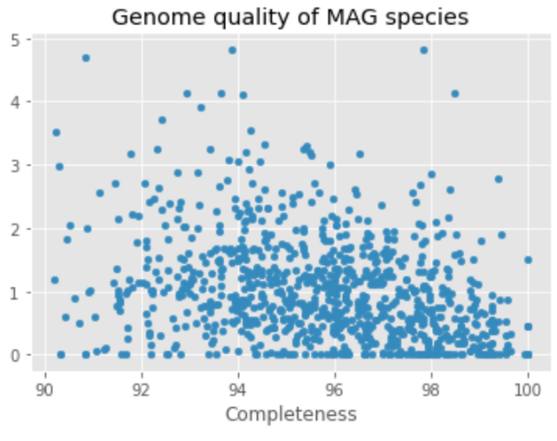


Predicted Gene ontology terms

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

Quality and completeness of metagenome assembled genomes





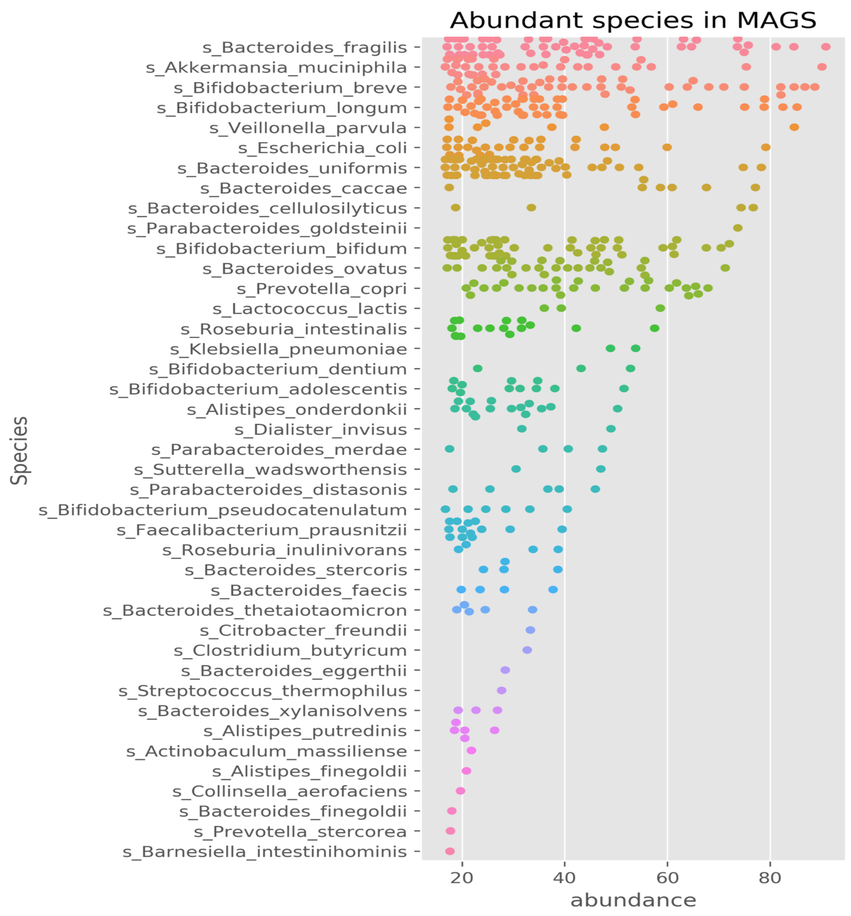
Genome quality threshold of >90% genome completeness and <5% contamination, the final genomes matching these criteria were 2,256

Results

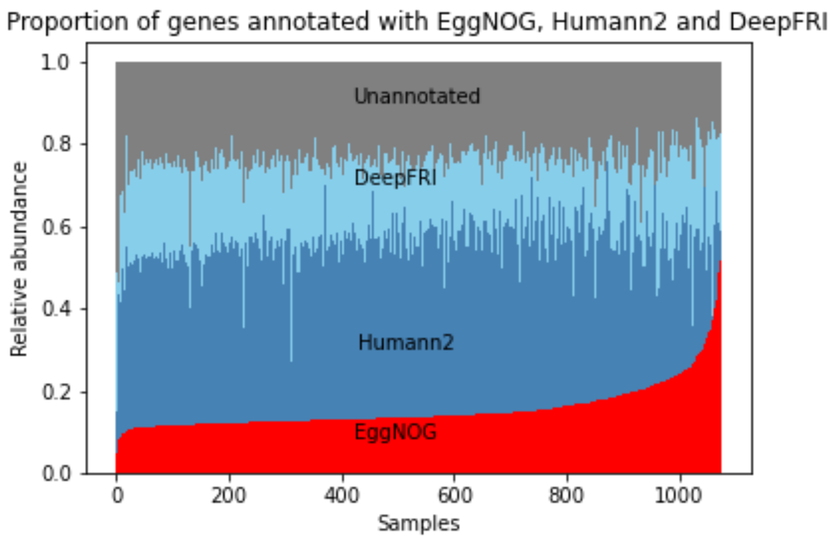
Assembly and gene prediction statistics

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.

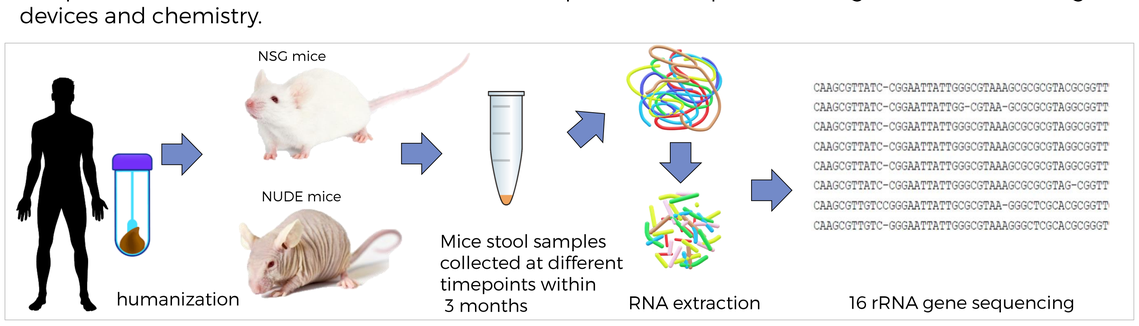


Figure 1

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.

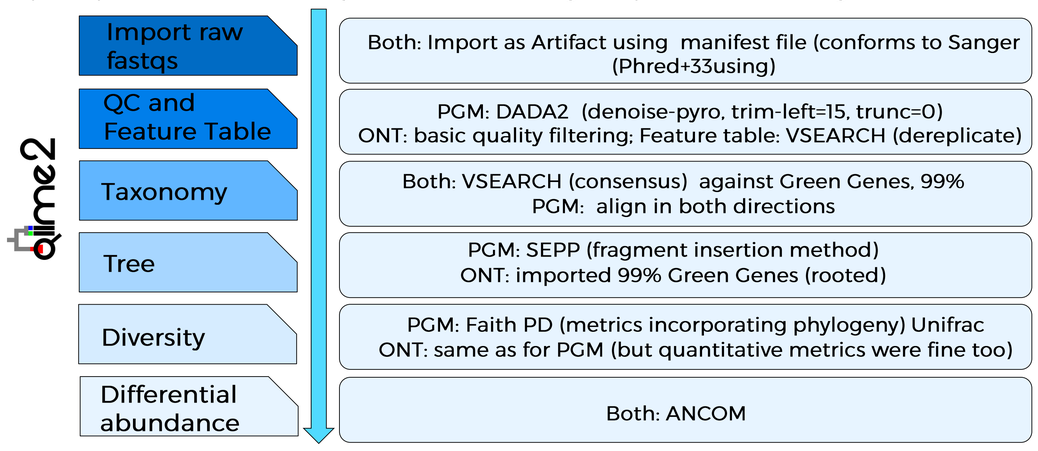


Figure 2

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

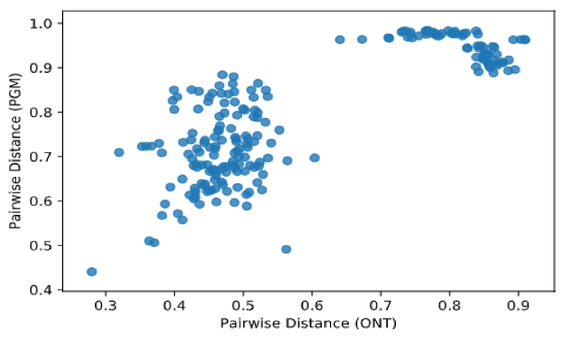
	Ion Torrent	Oxford Nanopore
# Samples	126 (123 mice, 1 human, 2 mock)	23 (22 mice, 1 human, 2 mock)
Unique features	9,877	3,543
Total features	17,908,604	1,130,914
Features per sample (median)	129,097	41,628
Reads per feature (median)	83	11
Features per sample at even sampling depth	45,000	9,000
Features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

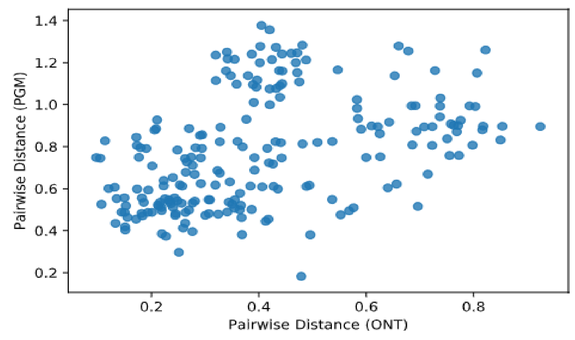
Mantel Test for unweighted Unifrac



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

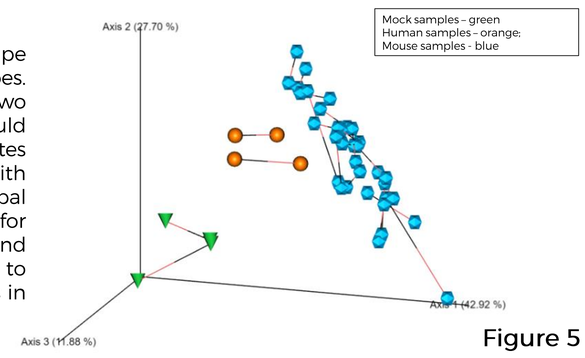


Figure 5

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.

Patterns of DNA variation between the autosomes, the X chromosome and the Y chromosome in *Bos taurus* genome

Bartosz Czech[1*], Bernt Guldbrandtsen[2,3], Joanna Szyda[1,4]

[1] Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

[3] Department of Animal Sciences, University of Bonn, Bonn, Germany

[4] Institute of Animal Breeding, Balice, Poland

[*] bartosz.czech@upwr.edu.pl http://theta.edu.pl

CONCLUSIONS

- fewer extreme variants are consistent with purging due to the homozygous state in males
- accumulation of nonsynonymous mutations on the BTY could be associated with loss of recombination
- variants in transcription regions on BTX have less severe consequences as compared to BTY and autosomes

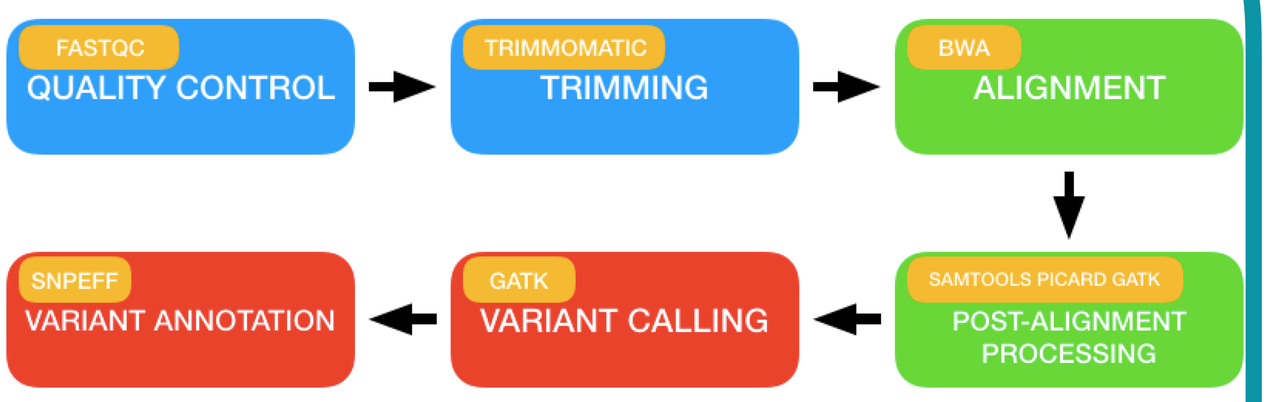
MATERIAL

- 217 individuals of 7 Danish cattle breeds
- WGS – Illumina HiSeq 2000
- assembly: ARS-UCD1.2_Btau5.0.1Y
- Btau_5.0.1 and ARS-UCD1.2 GFFs

RESULTS

- 23,655,295 SNPs / 3,758,781 InDels
- numbers of SNPs and InDels not uniformly distributed across 100kb non-overlapping windows ($P < 0.001$)
- Ka/Ks ratio: BTA = 0.79 BTX = 0.62 BTY = 2.00

METHODS



Statistical analysis:
- variant density on each chromosome
- InDel length • Ka/Ks ratio • nucleotide divergence
- Tajima's D • SIFT score

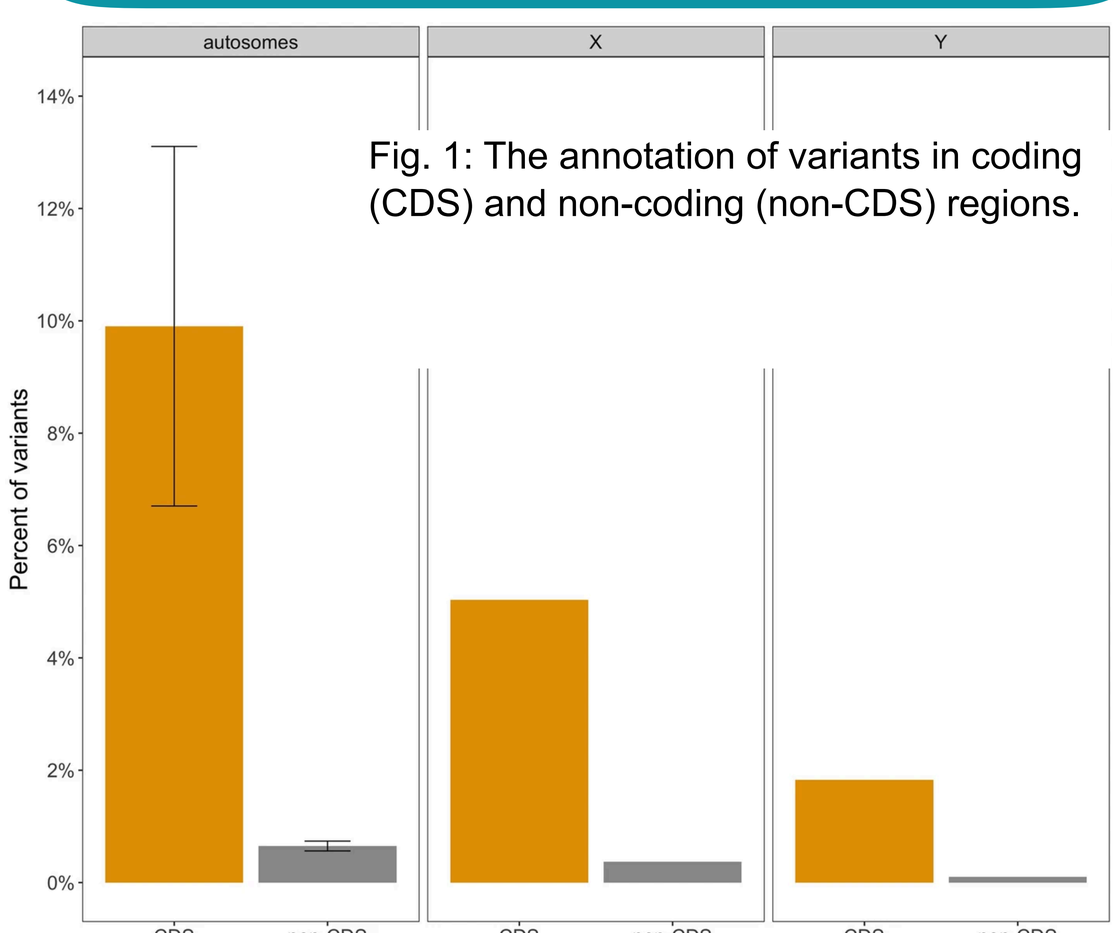


Fig. 1: The annotation of variants in coding (CDS) and non-coding (non-CDS) regions.

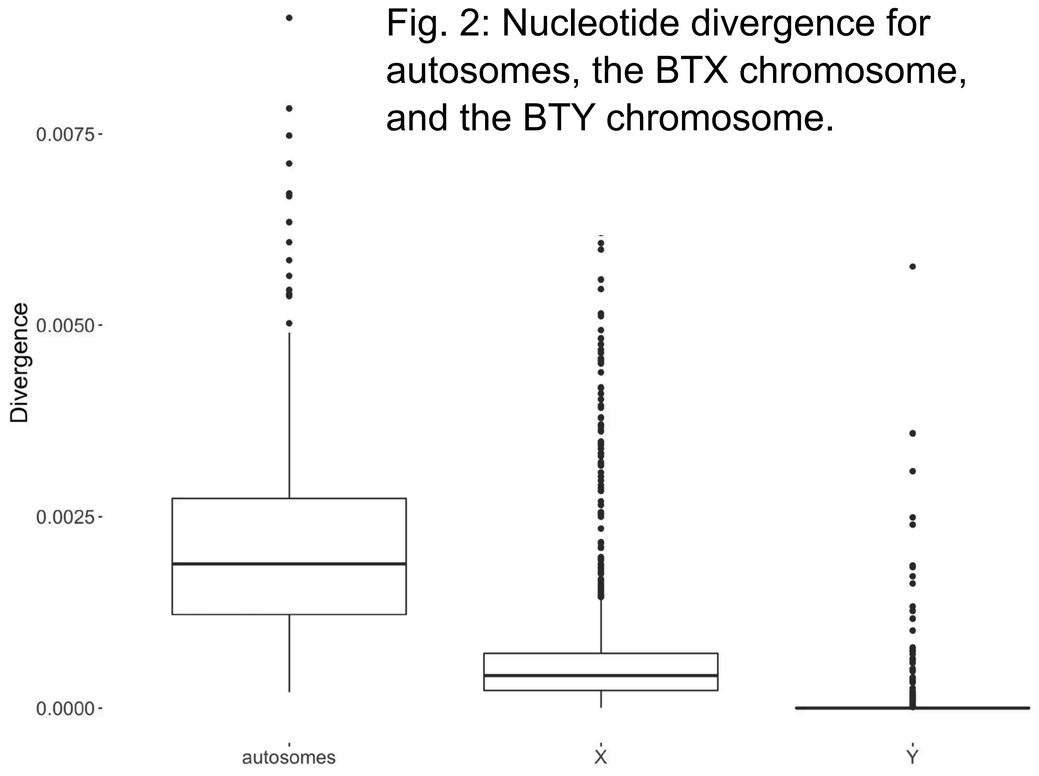


Fig. 2: Nucleotide divergence for autosomes, the BTX chromosome, and the BTY chromosome.

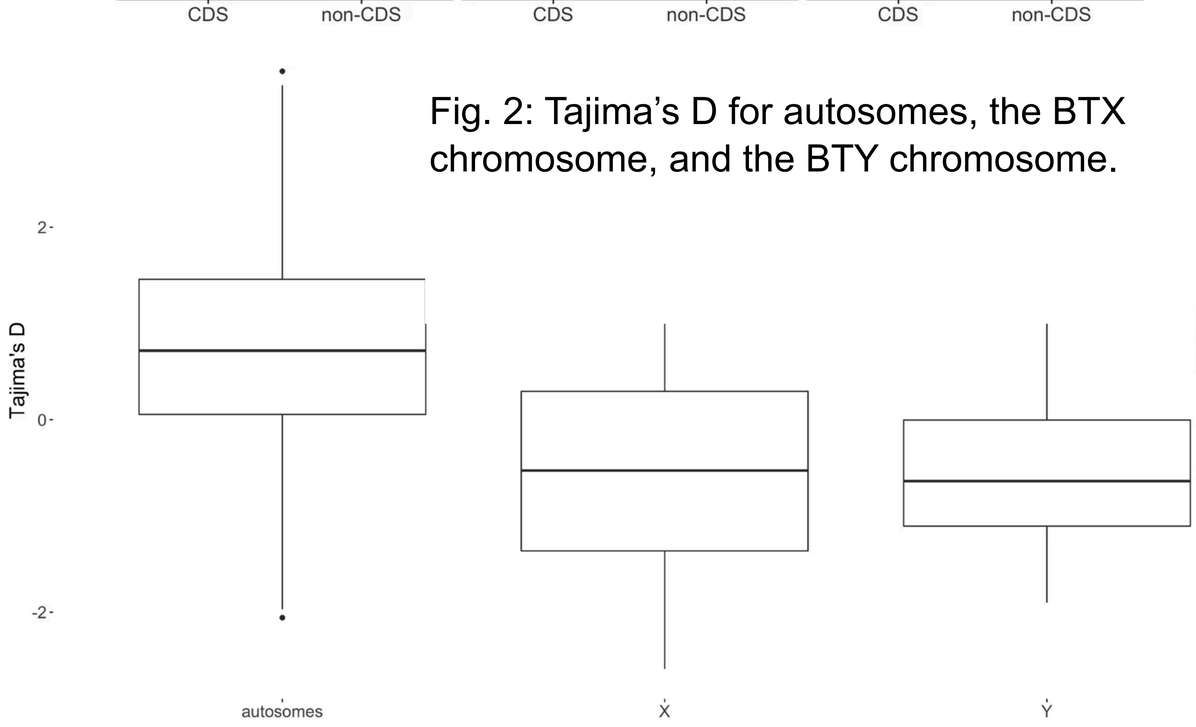


Fig. 2: Tajima's D for autosomes, the BTX chromosome, and the BTY chromosome.

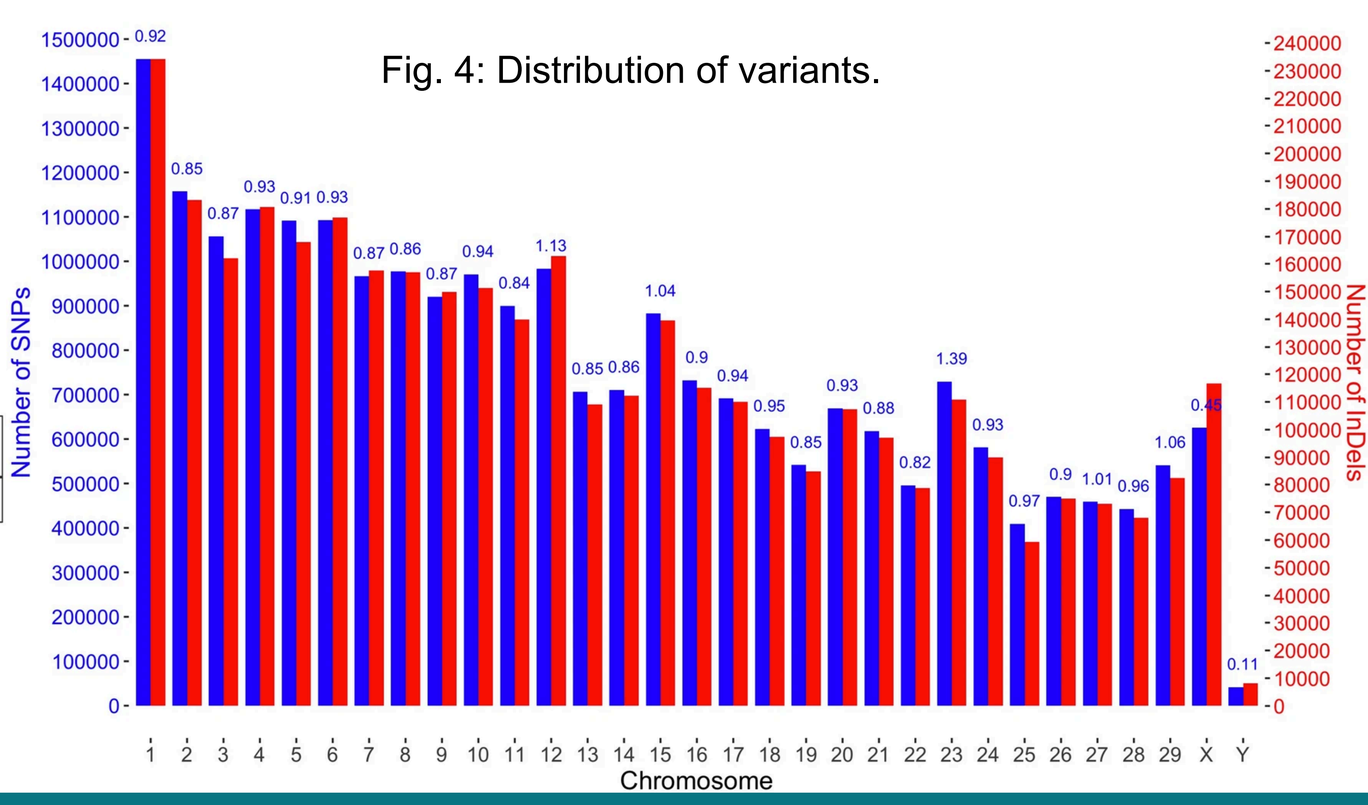


Fig. 4: Distribution of variants.

Mulitple sequence alignment analysis
master thesis

author: Paulina Dziadkiewicz (MiNI PW), advisor: dr hab. Norbert Dojer (MIMUW)
pedziadkiewicz@gmail.com, dojer@mimuw.edu.pl

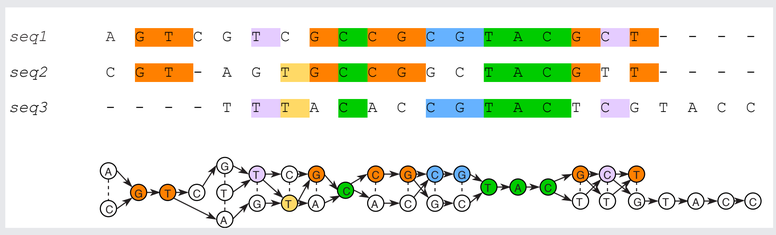
Introduction

Constant growth of genomic data leads to arising of a new research field called **pangenomics**. It is focused on delivering methods for joint multiple sequences processing. In this work a tool called *PangTree* is introduced.
The purpose of this tool is to extend currently used methods – **multiple sequence alignment**, **consensus search**, **multialignment graph representation** into new concept called **Affinity tree**. It is designed to be used as a taxonomic study or a reference genome for aligned sequences.

Multialignment as a graph

Graph representation of multiple alignment is based on partial order alignment graph.[1] The transformation is executed as follows:

1. Process multialignment column by column;
2. Merge identical nucleotides into single nodes;
3. Add directed edges between subsequent nodes and undirected for aligned nodes.



The representation is concise and intuitive. It is suitable to represent both short-length mutations and longer rearrangements, e.g. inversions or duplications.

Consensus idea

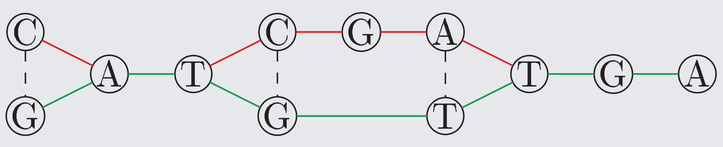
Typically, a consensus is determined by voting procedure on multialignment columns:

```
CATCGATGA
GATG-TTGA
CATG-TTG-

        ↓

CATG-TTGA
```

However, for multialignment given as a graph, Lee[1] proposed to find consensuses as minimum set of paths which describe all sequences.



Using Lee's approach we can build a graph model of multialignment and find a flat division of its component sequences into subgroups. Each of them has a consensus sequence assigned.

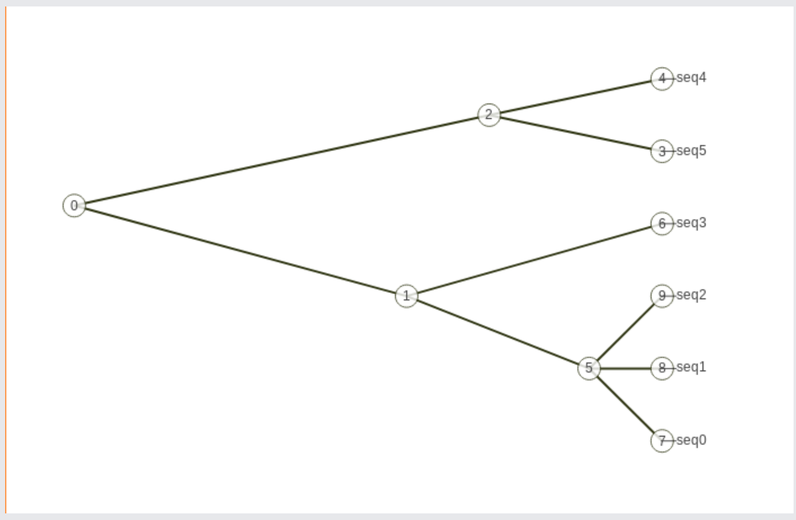
References

[1] Lee C. *Generating consensus sequences from partial order multiple sequence alignment graphs*, Bioinformatics. (2003) 22;19(8):999-1008.

[2] Dziadkiewicz, P., Dojer, *N. Getting insight into the pan-genome structure with PangTree.* BMC Genomics 21, 274 (2020).

Affinity tree

The introduced data structure is called Affinity tree. It serves as an extension of Lee's methods into hierarchical division of aligned sequences joint with consensus paths generation.
The root node has all input sequences assigned. Each non-leaf node has at least two children nodes that form a partition of the sequences assigned to their parent into more homogeneous subsets.



Each node has the following attributes assigned:

- a subset of input sequences,
- a linear consensus sequence being their common representation,
- a *minComp* (minimum compatibility) - value which reflects this node's homogeneity level.

Figure 1: An example of a Affinity tree

Affinity tree can be used as a reference genomes source, an evolution model or an assessment of heterogeneity for given dataset.

Simulated dataset

In order to evaluate the proposed solution a simulated multialignment was prepared using Evolver and evolverSimControl software. This alignment was based on a phylogenetic tree presented in Figure 2. It can be easily compared with the obtained Affinity tree which is shown in Figure 3.

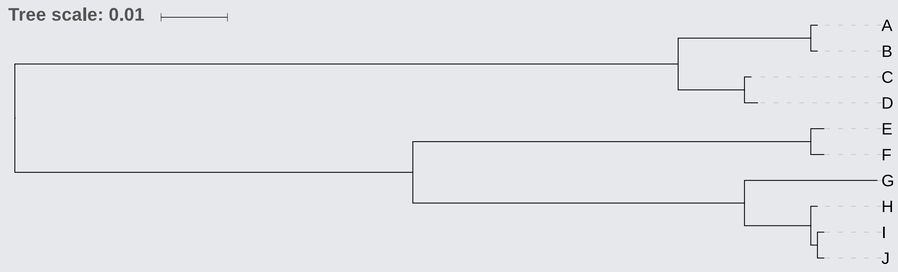


Figure 2: Phylogenetic tree for simulated data

The trees have similar forms which means, that the evolution pattern was correctly discovered by pangtree. However, the result includes not only the tree but also a consensus sequence assigned to each node. This is the main advantage of the Affinity tree over a phylogenetic tree.

For further simulations please follow the article[2].

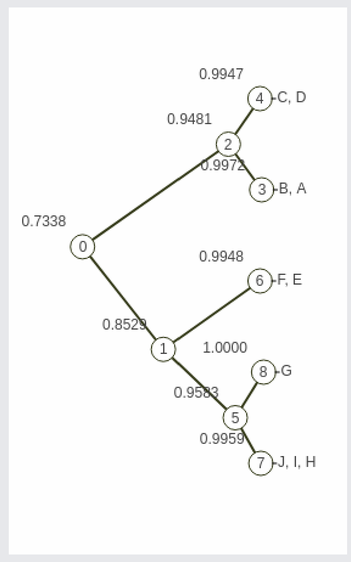


Figure 3: Affinity tree for simulated data

Ebola virus dataset

The proposed approach was also applied to Ebola virus alignment. The multialignment file was built using 160 genomes and is available in UCSC Ebola Portal together with associated studies.

The relationships between aligned sequences were correctly discovered.

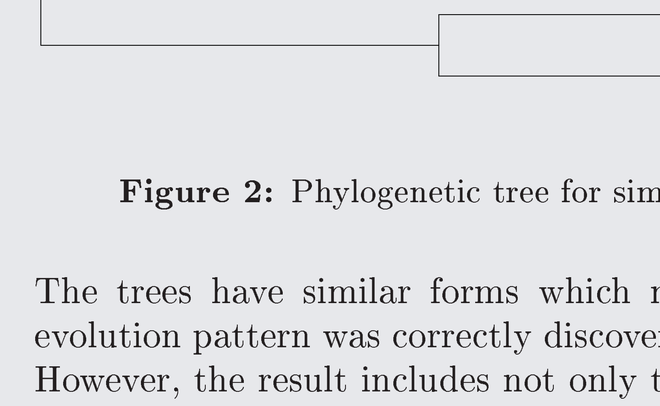


Figure 4: Local compatibilities between consensus sequences of ebolavirus species

Plots on Figure 4 show compatibilities with the consensus sequence of the species from the caption. Dark background indicates coding sequences, respective genes are listed below.

Figure 5: Ebola – Affinity tree

Cerebral Microbleeds detection on MR images with hybrid neural network

Aleksandra Suwalska[a], Yingzhe Wang[b], Ziyu Yuan[c], Yanfeng Jiang[c,d], Jinhua Chen[e], Mei Cui[b], Xingdong Chen[c,d], Chen Suo[c,f], Joanna Polanska[a]

[a] Silesian University of Technology, Department of Data Science and Engineering, 44-100 Gliwice, Poland
[b] Department of Neurology, Huashan Hospital, Fudan University, Shanghai, People's Republic of China
[c] Fudan University Taizhou Institute of Health Sciences, Taizhou, People's Republic of China
[d] State Key Laboratory of Genetic Engineering and Collaborative Innovation Centre for Genetic and Development, School of Life Sciences, Fudan University, Shanghai, People's Republic of China
[e] Taizhou People's Hospital, Taizhou, People's Republic of China
[f] Department of Epidemiology & Ministry of Education Key Laboratory of Public Health Safety, School of Public Health, Fudan University, Shanghai, People's Republic of China

Objective

The aim was to develop a novel tool for the automated detection of cerebral microbleeds (CMBs) based on magnetic resonance (MR) images. The system is expected to increase the sensitivity of CMB detection and to improve the accuracy of the diagnosis of the disease.

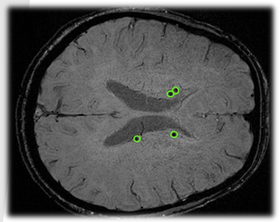


Figure 1: Example of brain image slice with CMB marked by an expert.

Introduction

Cerebral microbleeds are caused by structural abnormalities of the brain's small vessels. CMBs are linked with many neurological diseases; they can even lead to cognitive impairment, disability or death. They are visible on Susceptibility Weighted Imaging (SWI) sequences as round or elliptical areas with lower signal intensity and diameter up to 10 mm. Their manual detection is prone to errors and time-consuming.

Materials

In the study, MRI images from Taizhou People's Hospital were collected for a group of 304 patients and were used to train and test the system (Dataset 1). MR images from another 70 patients (Dataset 2) were used as an external independent validation. The process scheme is presented in Fig.2.

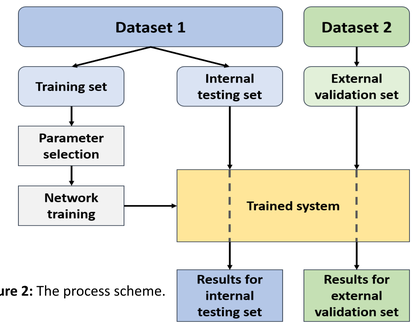


Figure 2: The process scheme.

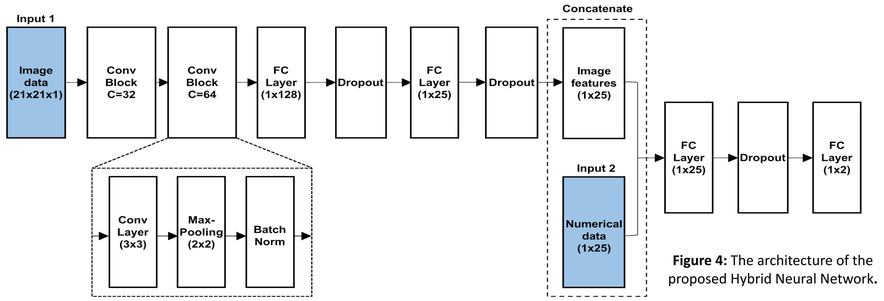


Figure 4: The architecture of the proposed Hybrid Neural Network.

Methods

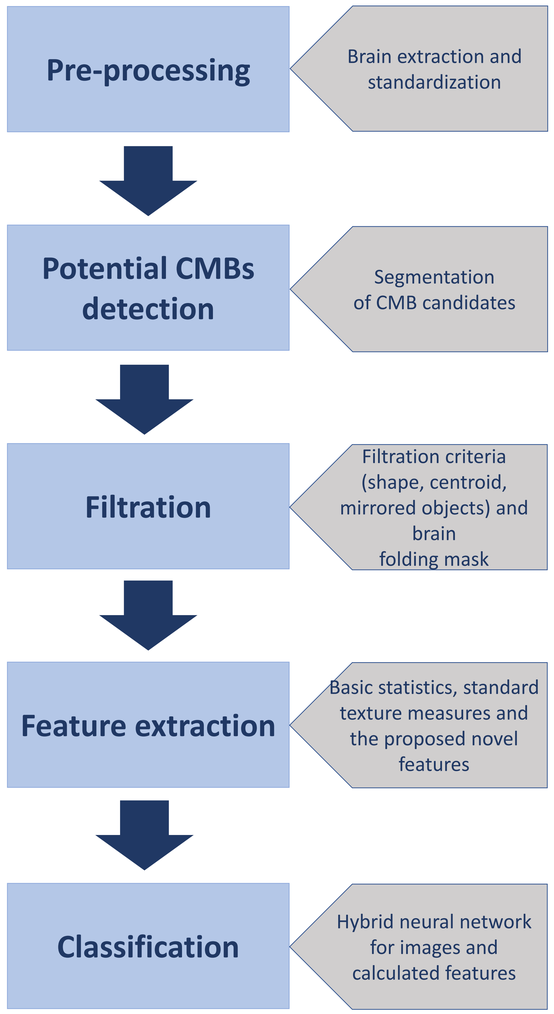


Figure 3: The pipeline of the CMB detection algorithm.

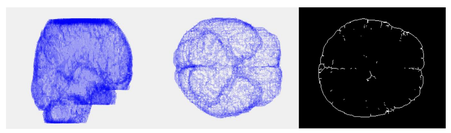


Figure 5: Brain folding mask in 3D and 2D.

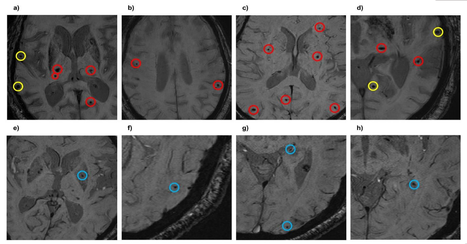


Figure 6: Exemplary positive samples. Red - CMBs identified correctly; yellow - CMBs lost by the system; blue - exemplary false positives.

Results

Dataset 1: The network reached a weighted accuracy of 94.48% with a sensitivity of 90.00% and specificity of 98.95%. The number of objects incorrectly classified as CMBs was 32 which gives an average of 0.54 false positives (FP) per patient.
Dataset 2: The system was able to detect 108 from 118 CMBs which resulted in the sensitivity of 91.5%. The number of false positives was 117 which gives 1.92 FPs per patient and the specificity of 95.2%.

		Training set			Test set					
Author	Modality	Patients without CMBs	Patients with CMBs	No. of CMBs	Patients without CMBs	Patients with CMBs	No. of CMBs	Sensitivity	Specificity	FPs/patient
Barnes et al. (2011)	SWI	-	6	120	-	6	6	81.70%	95.90%	107.50
Bian et al. (2013)	mIP SWI	-	5	116	-	10	304	86.50%	-	44.90
Chen, Yu et al. (2015)	SWI	-	15	62	-	5	55	89.13%	-	6.40
Van den Heuvel et al. (2016)	SWI+T1	18	23	491	-	10	136	89.00%	-	25.90
Dou, Chen et al. (2016)	SWI	-	270	270	-	50	117	93.16%	-	2.74
Ateeq et al. (2018)	SWI	-	14	104	-	6	63	93.70%	-	56.00
Chen et al. (2018)	SWI+echo scans	-	61	2458	-	12	377	94.70%	-	11.60
Liu et al. (2019)	SWI+phase	-	179	1473	10	31	168	95.80%	-	1.60
Suwalska, Wang et al. (2020) - Dataset 1	SWI	213	30	134	52	9	10	90.00%	98.95%	0.54
Suwalska, Wang et al. (2020) - Dataset 2	SWI	-	-	-	40	21	118	91.50%	95.20%	1.92

Table 1: Comparison with existing solutions (not all details were always available). Our results are marked red.

Conclusions

The use of both SWI images and numeric features allowed for the CMB's identification with high sensitivity and specificity without the need for additional imaging or complex models. On both test data, the developed system outperforms existing methods in terms of the number of false positives (FP) per patient. Our research confirms the usefulness of deep learning solutions to the problem of CMB detection based only on single MRI modality.

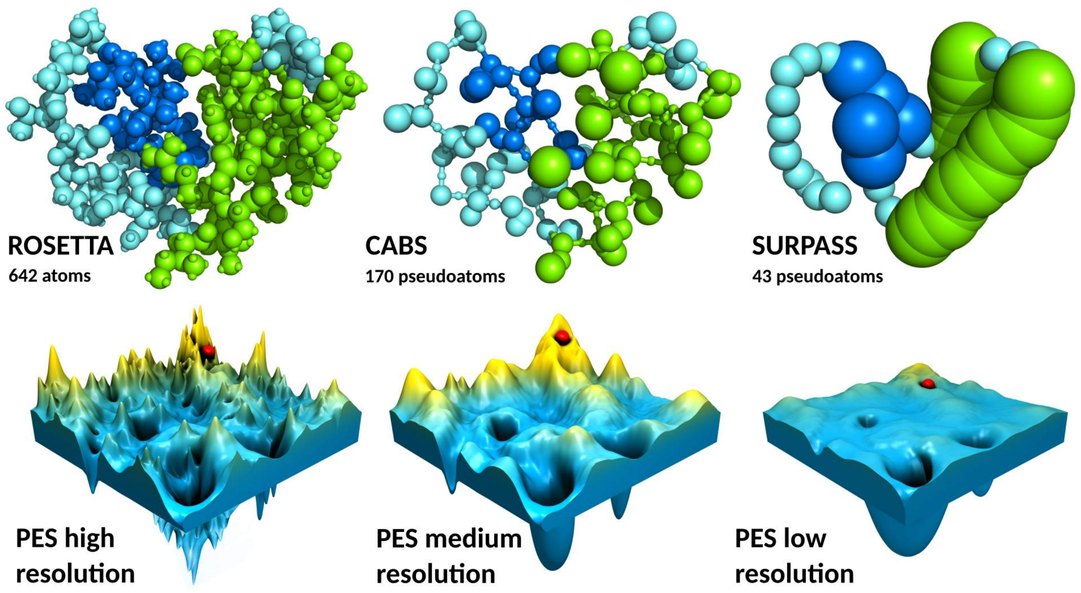
This work was partially financed by SUT BK-277/RAU4/2020. Calculations were carried out using infrastructure of GeCONiI (POIG.02.03.01-24-099/13).

Multiscale Modeling of Protein Structure and Dynamics Using Coarse-Grained Models of Various Resolution

Aleksandra Elżbieta Badaczewska-Dawid

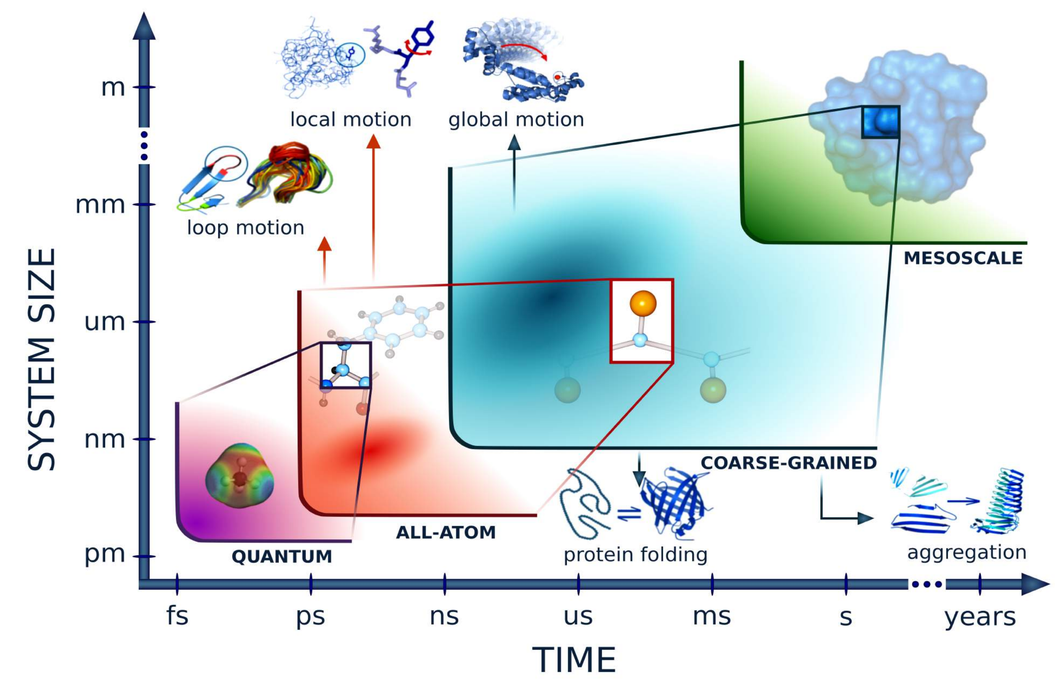
Department of Chemistry, Iowa State University, Ames, 50010 IA, U.S.; Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

COARSE-GRAINED PROTEIN MODELS



The coarse-grained models, their representation of protein chains, force fields, and sampling techniques must be carefully designed. In all coarse-grained models, the main purpose was to reduce the number of degrees of freedom. For this reason, pseudo atoms replace amino acid fragments or even entire amino acids. A broad spectrum of coarse-grained protein chain representations was proposed, starting with the simple lattice protein-like HP models or structurally more realistic low-resolution models like SICHO, by intermediate resolution coarse-grained models (e.g., CABS, UNRES) to almost exact coarse-grained protein models, like Rosetta or PRIMO. Medium-resolution CG models significantly expand the time scale and system size of molecular modeling. However, they struggle with de novo modeling of larger structures. Therefore, an efficient tool is needed to expand the range of de novo modeling of protein structure and dynamics by fast and efficient simulations of low-resolution structures.

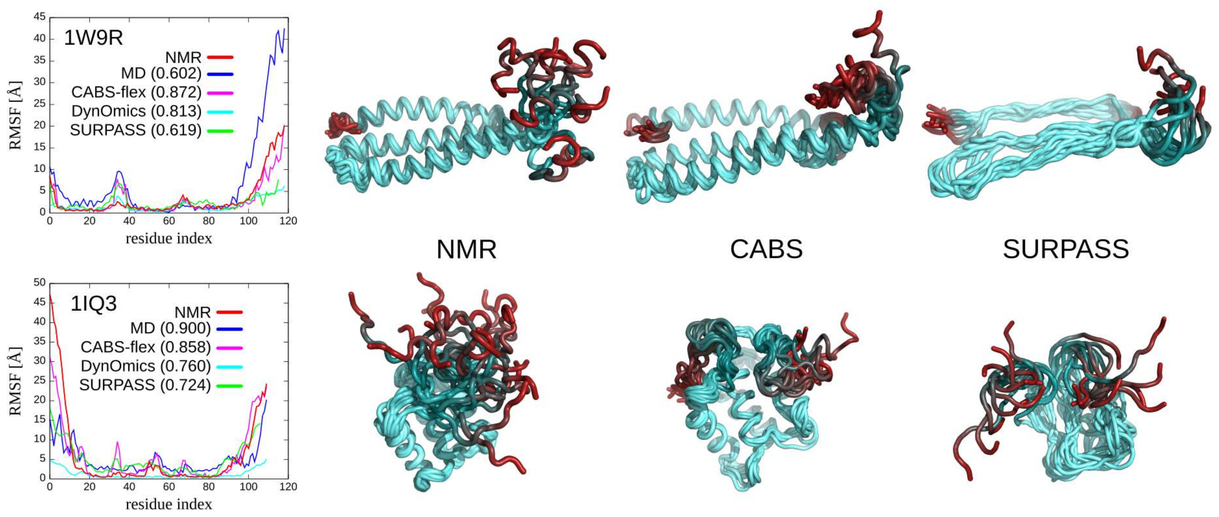
MULTISCALE MODELING



Classical atom-level molecular modeling can address many of the key tasks of structural biology, but its practical applications are still limited. This is a major reason why the development of coarse-grained protein modeling methods is needed. Coarse-grained models are computationally more effective and enable simulations of much longer time-scales and/or larger sizes of the systems studied. Multiscale methods that allow the transfer of information between various levels of granularity are more efficient and enable an analysis of larger systems on a longer time scale. Although successful multiscale modeling needs efficient and reliable algorithms for transferring information between calculations with different resolutions

CG MODELING OF PROTEIN DYNAMICS

Given the reports on the essential importance of protein dynamics for its biological function, we have studied the local flexibility of protein near the folded state. In the comprehensive study of 140 globular protein dynamics, we have applied various coarse-grained approaches: ENM-based modeling technique (DynOmics) and two representative simulation tools: medium-resolution CABS model and low-resolution SURPASS model. The proposed protocol succeeded in capturing the experimentally determined features (from NMR ensembles) of the investigated systems.

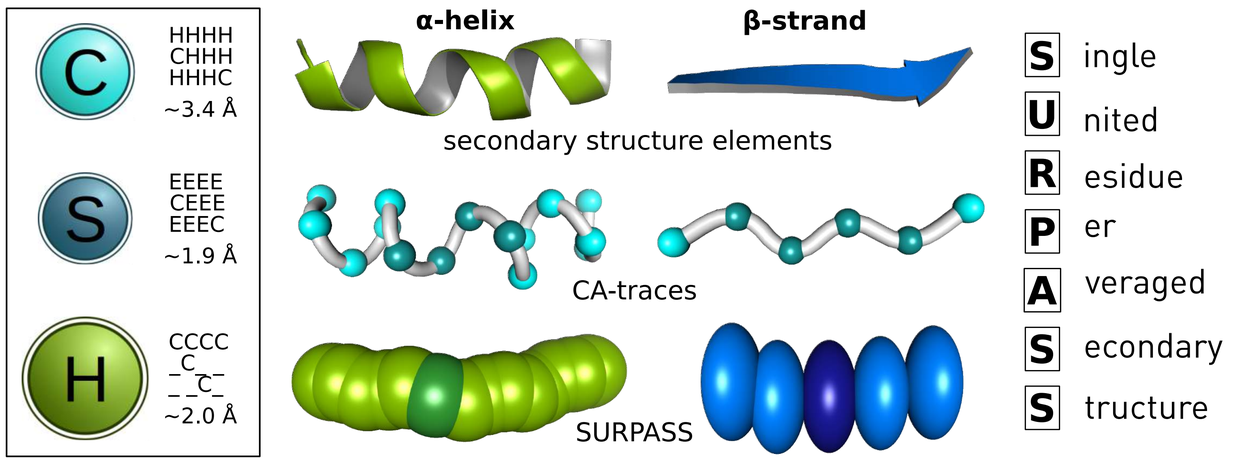


Due to its computational efficiency, SURPASS can be used for modeling long-time dynamics and large-scale structural transitions in protein systems that are significantly bigger than those tractable by the coarse-grained modeling tools of higher resolution. The models such as SURPASS can be useful as part of multiscale molecular modeling schemes. In such a scheme, SURPASS simulations can provide a collection of protein-like low-resolution starting structures, and these could be used for more accurate methods, e.g., as an input to replica-exchange simulations with a medium-resolution CG model (for example, CABS). Intermediate resolution structures can be finally subjected to all-atom reconstruction and MD refinement/scoring simulations.

SURPASS MODEL

LABORATORY of THEORY of BIOPOLYMERS http://biocomp.chem.uw.edu.pl/tools/surpass Andrzej Kolinski Research Group

RESEARCH PEOPLE PUBLICATIONS TOOLS CONTACT

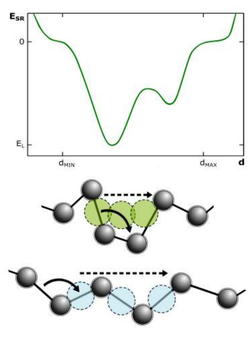


SURPASS is a low-resolution, deeply coarse-grained model of protein structure. The number of pseudo residues representing protein structure corresponds to the length of the protein sequence. The main idea behind the model is based on a unique generalization of the local geometry of a polypeptide chain. Namely, positions of pseudo atoms are defined by averaging the coordinates of the four consecutive α-carbons along the chain. These four-residue fragments are replaced by a single center of interactions. The choice of four-residue averaging is crucial for the geometry of the model. In contrast to other short fragments of different lengths, only the four-residue averaging leads to an almost linear shape of the SURPASS fragments representing helices or ß-strands. This feature of the model results in simple and effective sampling schemes. The SURPASS representation assumes three types of pseudo atoms depending on secondary structure assignment: H (helical), S (ß-strand), C (coil-like).

SURPASS FORCE FIELD

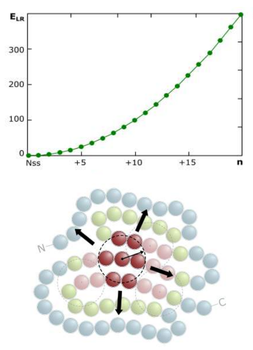
Designing and derivation the force field for a coarse-grained model is always a key point for its performance. A combination of the statistical potentials defines the knowledge-based SURPASS force field. They describe local structural regularities characteristic for most globular proteins. The generic terms are basically sequence-independent and are encoded non-directly via secondary structure assignment. The solvent is treated implicitly, and its effects (water with other small molecules or a membrane environment for transmembrane proteins) are included directly in the statistical potentials that describe interactions between the united residues.
The specific interaction model distinguishes the protein-like SURPASS chain from a random polymer.

SHORT RANGE R12, R13, R14, R15



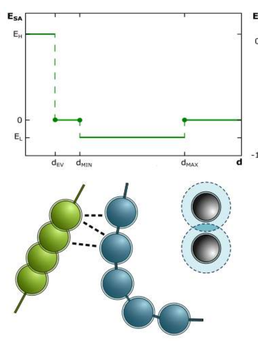
Distances and angles between atoms close along the sequence in polypeptide chains are highly restricted due to various short-range interactions, which provide the correct local geometry of the structure.

LOCAL REPULSIVE



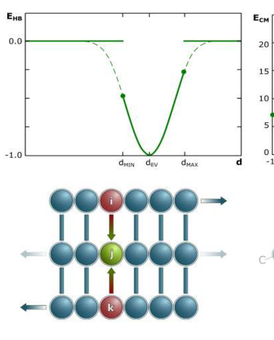
To prevent the excessive and non-physical collapse of a structure, generic local repulsions are needed. Using deeply buried elements derived from PDB, we estimated the number of neighboring SURPASS atoms.

SOFT-CORE ATTRACTION



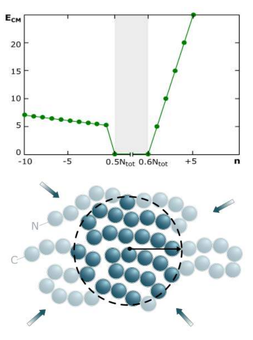
To avoid steric clashes between pseudo atoms distant in sequence but close in space, the excluded volume cut-off was derived. Contacts are rewarded only for specific distances between atoms of a given II-structure type.

HYDROGEN BOND



H-bonds between residues close to each other along the chain are treated implicitly. H-bonds between residues that are distant in the sequence (in the extended fragments) are modeled more directly.

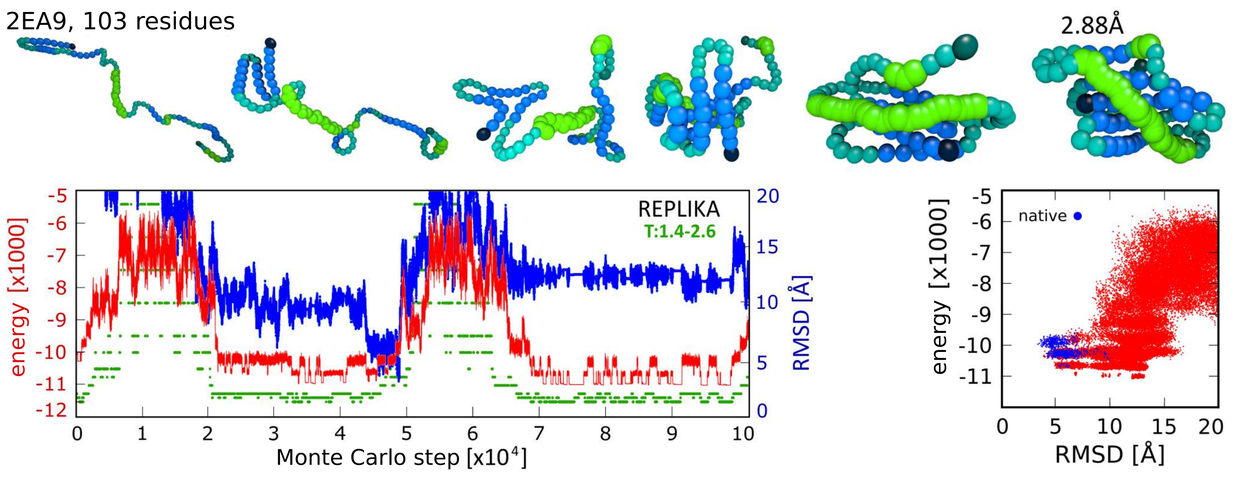
CENTRO-SYMMETRIC



To force the SURPASS chain to fold into globular topology, we used a simple centrosymmetric potential. Its purpose is to maintain a sufficiently high degree of packing of pseudo residues in the protein core.

SURPASS MODELING OF PROTEIN STRUCTURE

SURPASS model was used for replica-exchange Monte Carlo dynamics simulation of proteins, with secondary structure as the only sequence-dependent input data for the interaction model. The studied cases were a representative set of single-domain globular proteins. The set contained 7 helical proteins, 9 mostly ß-sheet, and 8 mixed alpha/beta proteins. In the test simulations presented here, the secondary structure assignments required by the model were taken directly from the PDB database. Replica exchange Monte Carlo simulations were performed with 12 replicas for each tested protein. The starting structures of all replicas had fully expanded the conformation of model chains.



The method efficiently samples the entire conformational space of polypeptide chains. Despite its deep simplification, the SURPASS model reproduces reasonably well the basic structural properties of proteins. Also, the accuracy of the resulting native-like models, measured by the RMSD between the generated chains and the SURPASS representation of experimental structures, is surprisingly good for such a level of coarse-graining. We demonstrated that different assignments and/or predictions of secondary structures are sufficient for enforcing cooperative formation of native-like folds of SURPASS chains for the majority of single-domain globular proteins. Simulations of globular protein structure assembly have shown that the accuracy of secondary structure data is usually not crucial for model performance.

S. Kmiecik, et.al., Chem. Rev. 2016, 116, 14, 7898–7936
A.E. Dawid, et.al., J. Chem. Theory Comput., 2017, 13(11), 5766–5779
A.E. Dawid, et.al., J. Chem. Theory Comput., 2018, 14(4), 2277–2287
S. Kmiecik, et.al., Int. J. Mol. Sci. 2018, 19(11), 3496
A.E. Badaczewska-Dawid, et.al., Methods Mol. Biol., 2020, vol 2165

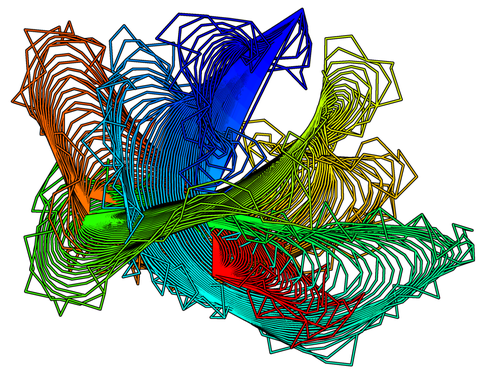
UNIVERSITY OF WARSAW

IOWA STATE UNIVERSITY

Algorithms and models for protein structure analysis

Aleksandra I Jarmolińska[1,2], Joanna I Sulkowska[1], Anna Gambin[2]

1Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097, Warsaw, Poland
2Faculty of Mathematics, Informatics, and Mechanics, Uniwersity of Warsaw , 02-097 Warsaw, Poland



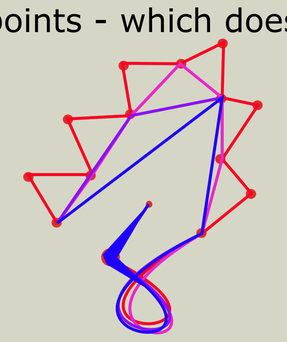
The last decade has seen a large increase in the number of studies related to protein topology. Currently, there are over 1500 known knotted or slipknotted protein chains and almost 10 000 protein links. Screening of available RNA structures has also found entanglements. Recent advances in the study of chromatin structure gave rise to new 3D models—many of which contain entanglements, including composite knots. Still, the subject of molecular entanglements remains relatively unknown to a lot of researchers, including those studying protein structures. One obvious reason is the steep learning curve for actually seeing the knots in a 3D structure visualization. Knot_pull (Jarmolinska et al, 2019) allows an easy analysis of topological intricacies by providing the user with a trajectory of smoothing steps—from the full structure, to the minimal number of coordinates preserving the original topology (with regard to fixed position of chain termini) — and the knot type (including separation of composite knots, and indication of any linking present) - without using the prevalent probabilistic approach.

Studying the sequences of entangled proteins also encounters problems - finding the most closely related protein family may require detecting the similarity based on sequence profiles, which are not easily (multiple-)aligned. To overcome this obstacle, I introduce two new heuristic for creating a multiple profile alignment, by using a modified Dijkstra's shortest path tree algorithm to find the maximum weight trace (Kececioglu, 1993) of a set of pairwise alignments. This allows for an easy, large scale comparison of loosely related protein groups.

Simplify

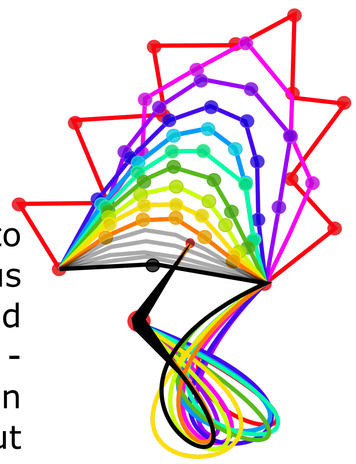
Previous approach

Structures are simplified by reducing the number of points - which doesn't necessarily extrude the ends.

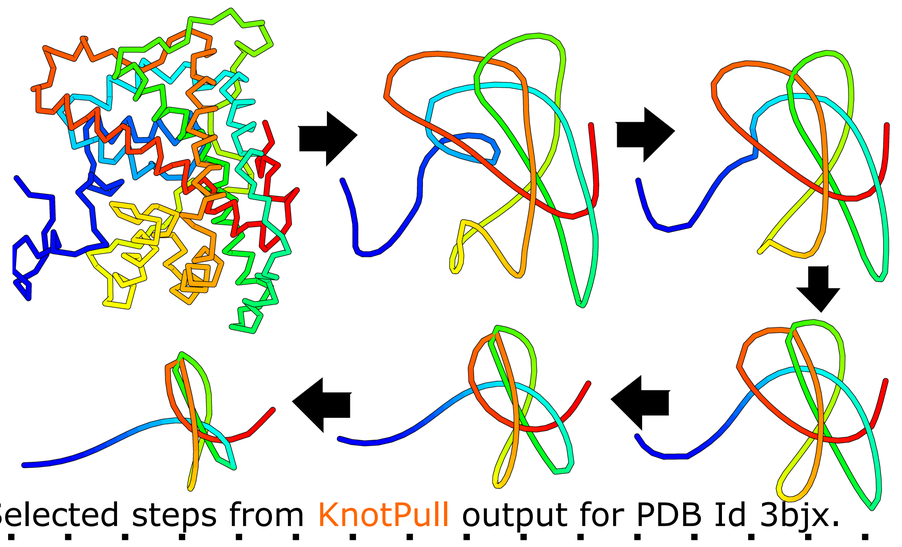


KMT

KnotPull allows all connections to be split or shortened thus tightening the entanglement - and making the ends stand out more - so that there exists a projection which could be closed without adding to the entanglement.

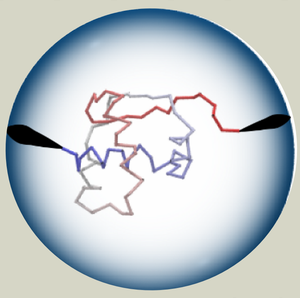


knot_pull



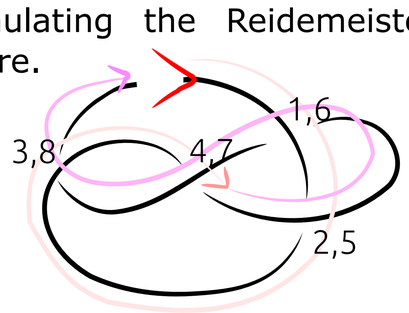
Selected steps from KnotPull output for PDB Id 3bjx.

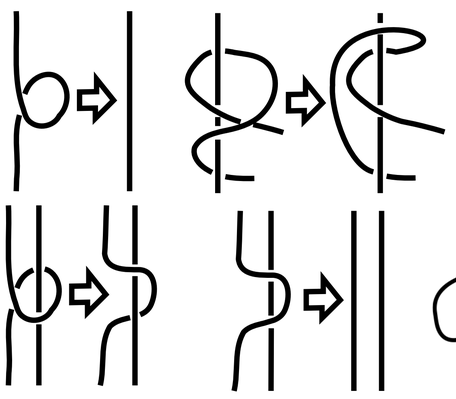
Recognize



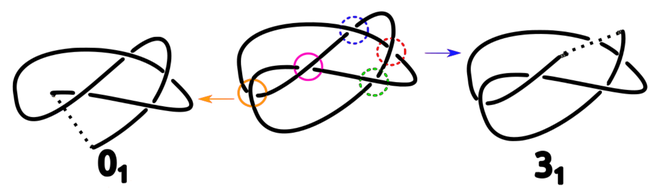
Knot type is calculated based on a polynomial, which requires a closed 3d curve projected on a plane. Open chain needs to be closed - randomly so as not to affect the results.

KnotPull uses the Dowker-Thistlethwaite code, which reads the structure as implicitly closed. This code is then simplified by calculations on the code itself simulating the Reidemeister moves on the structure.





Additinally, DT code recognizes different realizations (based on the location of the break - the temini) of the same knot type.



HHaligner

aligns multiple sequence profiles

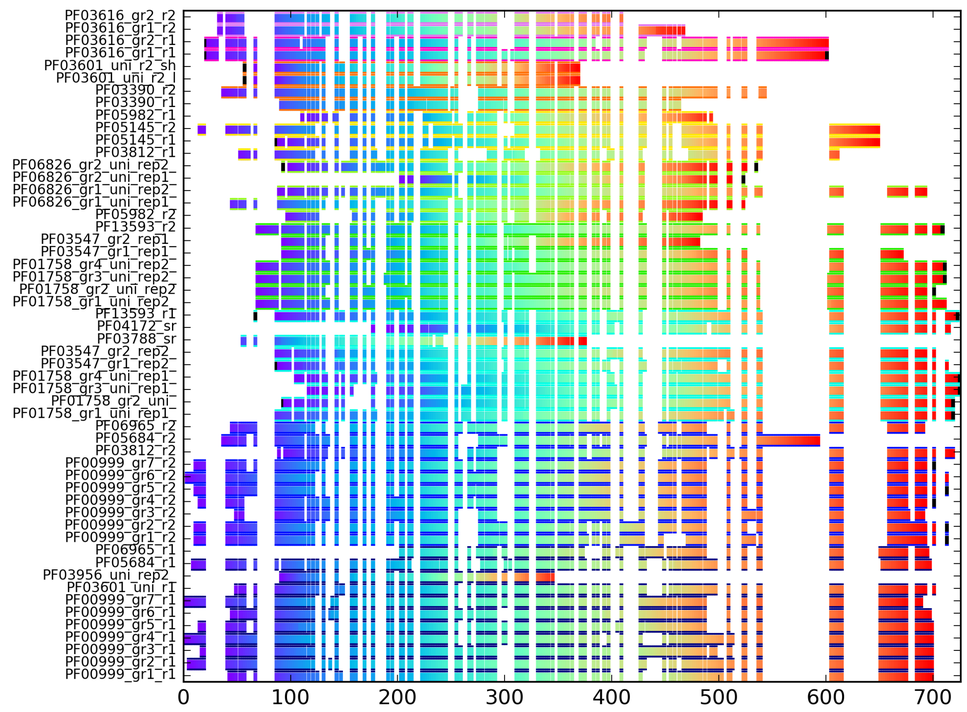
https://github.com/dzarmola/HHsearch-resuits-aligner

KnotPull

finds entanglements in 3D structures

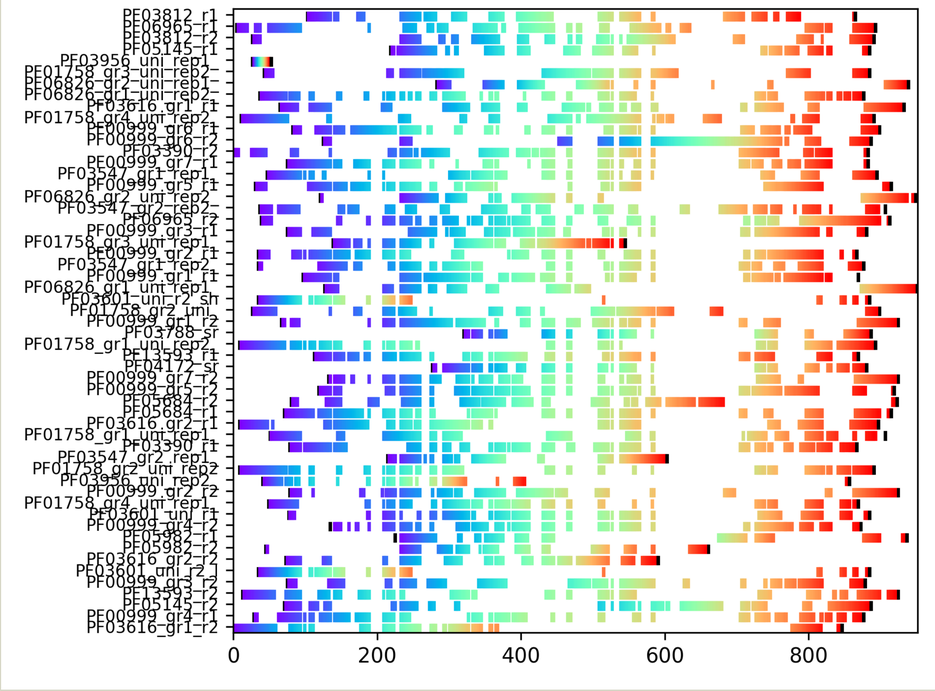
https://github.com/dzarmola/knot_pull

Multiple sequence alignment is a great way of analysing the similarities between related sequences - aligning their matching regions highlights the conserved characteristics. However, a single sequence is not enough to convey the diversity of a given protein - thus MSA cannot fully show the similarities between the entire families, like a multiple **profile** alignment could.



Right: MSA of the representatives of multiple distantly related families of transmembrane proteins.

Left: MPA of the profiles of the same



HHaligner creates a graph of positions in profiles connected based on their similiarity in pairwise profile alignments. Then, by heuristically optimizing the maximum weight trace, we can resolve the alignment.
Since the idea is to find the similarities, this approach additionally clears out any insertions specific to just one of the profiles.

Bibliography

Jarmolinska AI, Sulkowska JI, Gambin A, Bioinformatics (2020) - Knot_pull - python package for biopolymer smoothing and knot detection.

Kececioglu J, Annual Symposium on Combinatorial Pattern Matching (1993) - The maximum weight trace problem in multiple sequence alignment

Jarmolinska AI PhD thesis (2019) - Algorithms and models for protein structure analysis

CeNT CENTRE OF NEW TECHNOLOGIES

UW MM Wydział Matematyki, Informatyki i Mechaniki

INTEGRATIVE DATA ANALYSIS METHODS IN MULTI-OMICS MOLECULAR BIOLOGY STUDIES FOR DISEASE OF AFFLUENCE BIOMARKER RESEARCH

Anna Papież
Supervisor: Joanna Polańska
Data Mining Group, Institute of Automatic Control, Silesian University of Technology

Silesian University of Technology

Public Health England

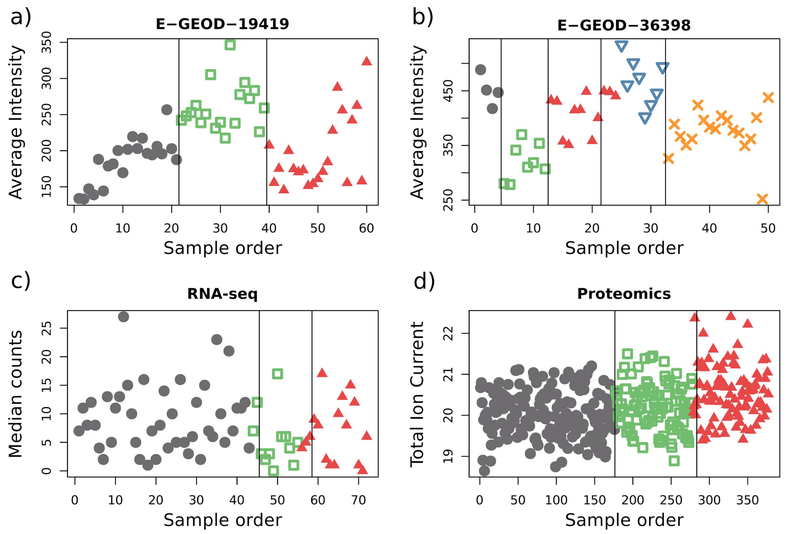
HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

AIM OF THE STUDY

The goal of this work was to investigate diverse approaches for high-throughput molecular biology integrative data analysis to enable the discovery of disease of affluence biomarkers. The research methodology comprises a thorough overview of existing approaches for data combination, merging, comparison, and joint analysis, as well as the development of new methods for handling multi-omics studies. The expected outcomes of this work include the establishment of novel tools and procedures tailored to the tasks of multi-platform and multi-omics data and result integration.

INTRA-EXPERIMENT INTEGRATION

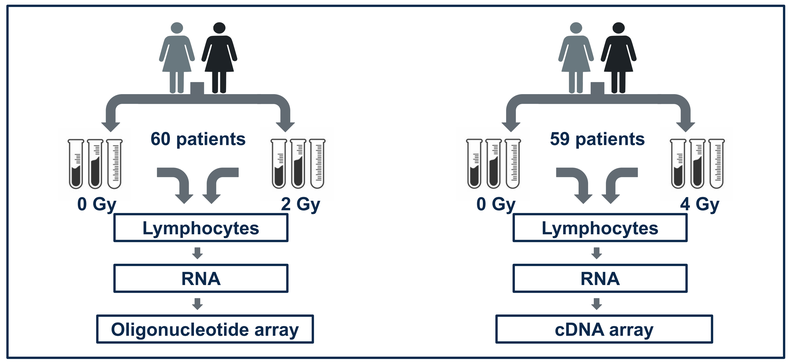
An original batch effect identification algorithm based on dynamic programming was proposed, as correcting for these effects constitutes a part of the intra-experiment data integration pipeline. The BatchI algorithm is based on partitioning a series of high-throughput experiment samples into sub-series corresponding to estimated batches. The dynamic programming method is used for splitting data with maximal dispersion between batches, while maintaining minimal within batch dispersion. The procedure has been tested on a number of available datasets with and without prior information about batch partitioning. Datasets with a priori identified batches have been split accordingly. Batch effect correction is justified by higher intra-group correlation. In the blank datasets, identified batch divisions lead to improvement of parameters and quality of biological information, shown by literature study and Information Content.



The BatchI algorithm's performance on identifying batch structure is proven to be highly efficient, and moreover, batch effect preprocessing entails potential new knowledge discovery in studied diseases and conditions. It is available to the scientific community as an R package.

INTER-PLATFORM INTEGRATION

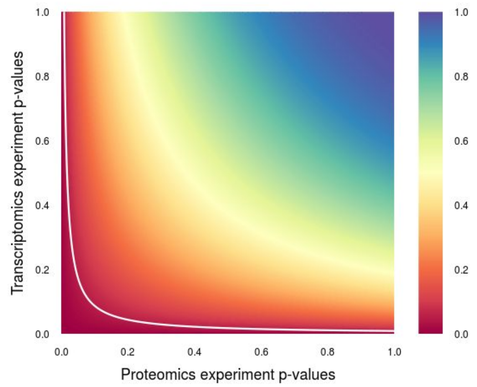
The analyzed data consist of two gene expression sets obtained in studies of radiosensitive and radioresistant breast cancer patients undergoing radiotherapy. The data sets were similar in principle; however, the treatment dose differed. It is shown that introducing mathematical adjustments in data preprocessing, differentiation and trend testing, and classification, coupled with current biological knowledge, allows efficient data analysis and obtaining accurate results. The tools used to customize the analysis workflow were batch effect filtration with empirical Bayes models, identifying gene trends through the Jonckheere–Terpstra test and linear interpolation adjustment according to specific gene profiles for multiple random validation.



The application of non-standard techniques enabled successful sample classification at the rate of 93.5% and the identification of potential biomarkers of radiation response in breast cancer, which were confirmed with an independent Monte Carlo feature selection approach and by literature references. This study shows that using customized analysis workflows is a necessary step towards novel discoveries in complex fields such as personalized individual therapy.

INTER-OMICS INTEGRATION

The goal of this part was to elucidate molecular mechanisms of radiation-induced IHD by integrating proteomics data with a transcriptomics study on post mortem cardiac left ventricle samples from Mayak workers categorized in four radiation dose groups (0 Gy, < 100 mGy, 100–500 mGy, > 500 mGy). The proteomics data originated from a label-free analysis of cardiac samples. The transcriptomics analysis was performed on a subset of these samples. Stepwise linear regression analyses were used to correct the age-dependent changes in protein expression, enabling the separation of proteins, the expression of which was dependent only on the radiation dose, age or both of these factors. Importantly, the majority of the proteins showed only dose-dependent expression changes. Hierarchical clustering of the proteome and transcriptome profiles confirmed the separation of control and high-dose samples. Restrictive (separate p-values) and integrative (combined p-value) approaches were used to investigate the enrichment of biological pathways.



Custom statistical integrative methods applied to a transcriptomics and proteomics data set on ischemic heart disease plutonium mine workers enabled discrimination of dose dependent protein expression changes from the age dependent changes and validation of pathways identified previously in the proteomic data.

This work was funded partially by:
- "DoktoRIS - Scholarship Program for Innovative Silesia", co-financed by the European Union under the European Social Fund
- POIG.02.03.01-24-099/13 grant funding and infrastructure:"GeCONiI - Upper Silesian Center for Computational Science and Engineering"
- National Science Centre Harmonia grant number DEC-2013/08/M/ST6/00924: "BioRadInt"
- National Science Centre OPUS grant number UMO-2015/19/B/ST6/01736: "BiTIMS"
- National Center for Research and Development PBS grantnumber PBS3/A7/29/2015/ID-247184 "MOLTEST BIS"