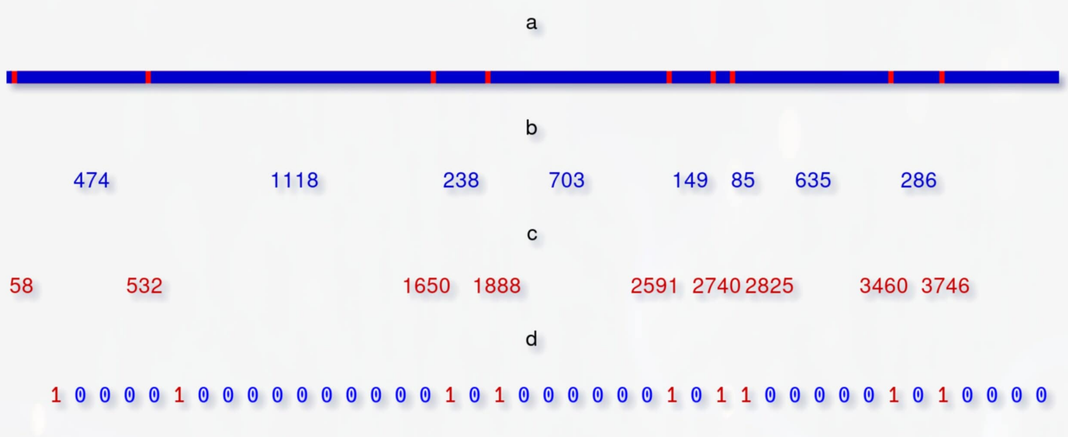
Binary genome maps assembly

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:

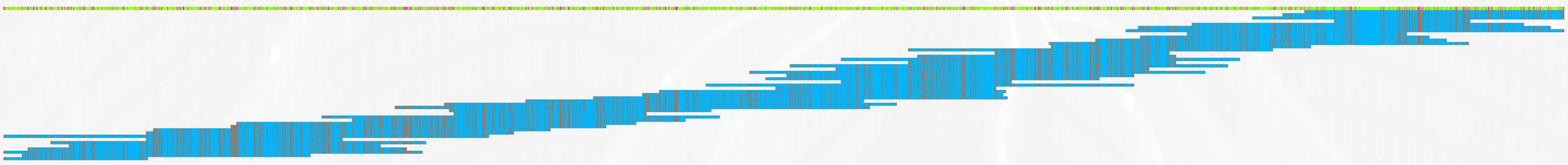
a. is restricted genome with red markers,
b. is distances between markers,
c. is set of positions,
d. is binary genome map

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.

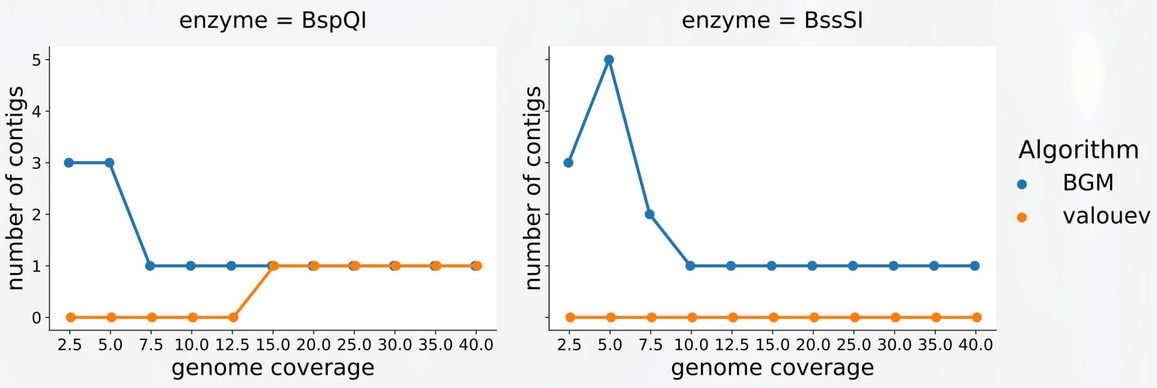
E.Coli maps visualisation



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

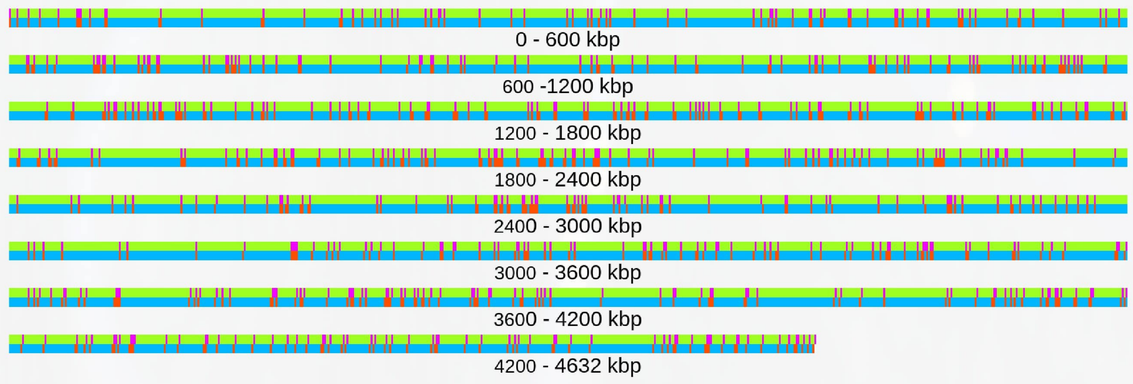
E.Coli experiments and comparison with *valouev et. al.* [1][2]

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.

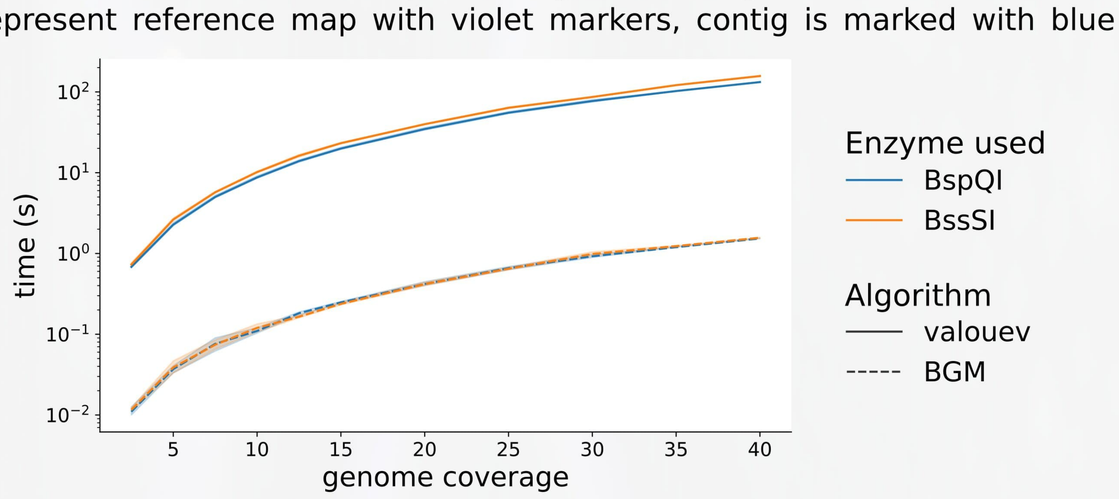


Using only 7.5x coverage of e.coli genome and *BspQI* enzyme we were able to obtain 1 contig. Larger contig was containing information about whole genome. The exact accuracy is not measurable due to nature of quantisation process as discussed above but very restriction site was restored with some artifacts 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]

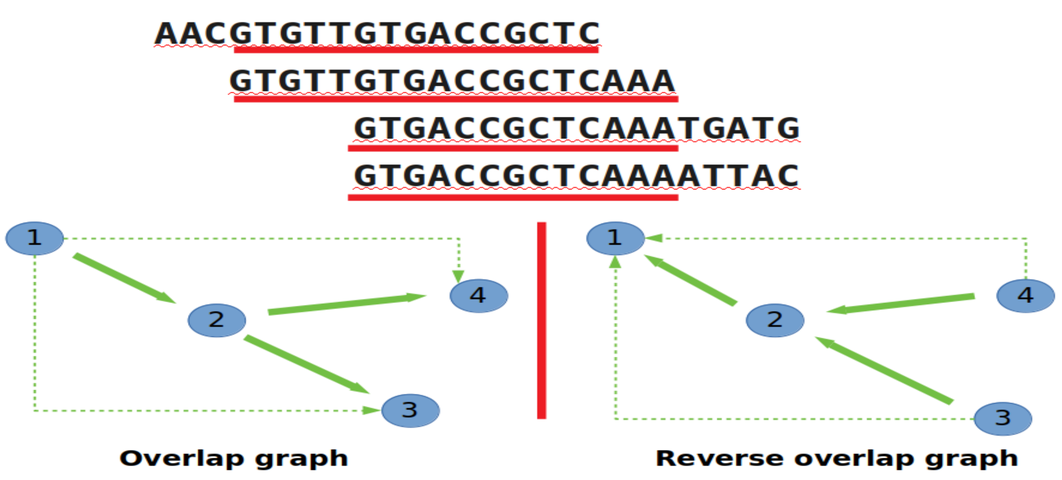
[1] Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[2] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

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.

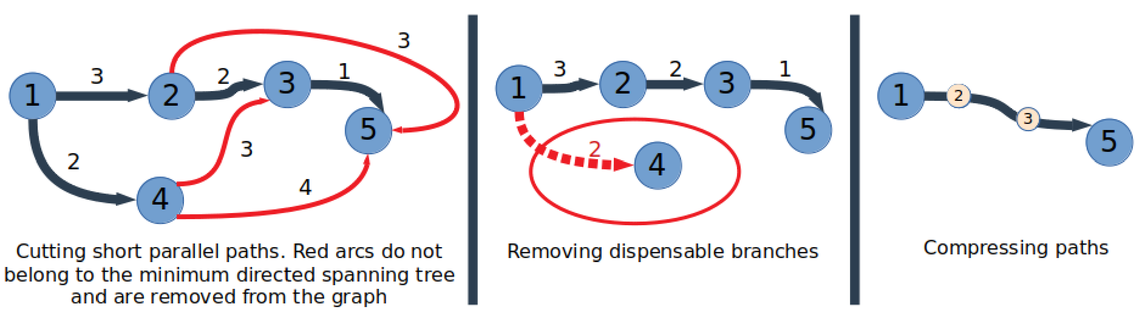
Overlap graph construction

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.



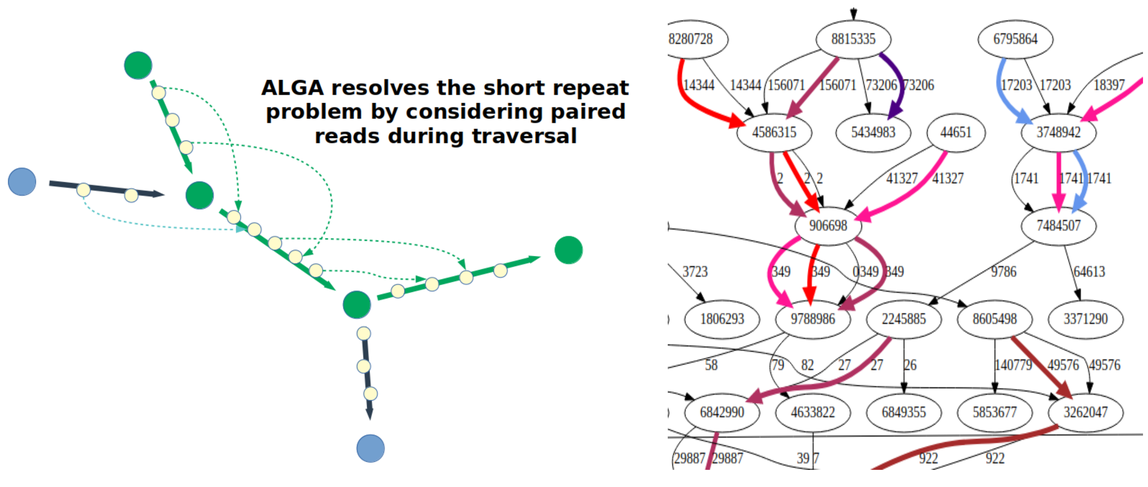
Graph transformation

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.

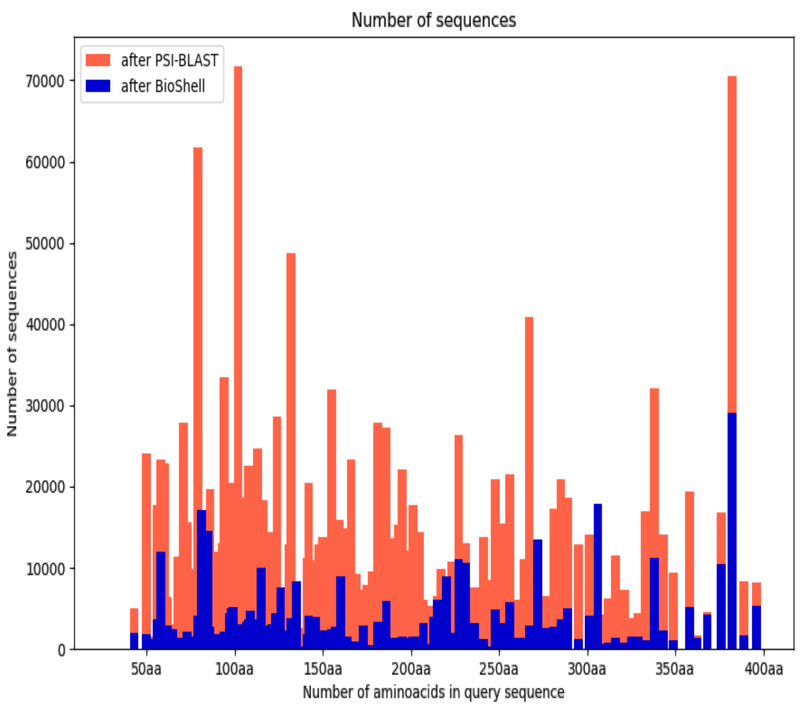
This work was supported by the European Regional Development Fund [grant POIR.04.02.00-30-A004/16] and carried out in the European Center for Bioinformatics and Genomics, Poznan University of Technology.

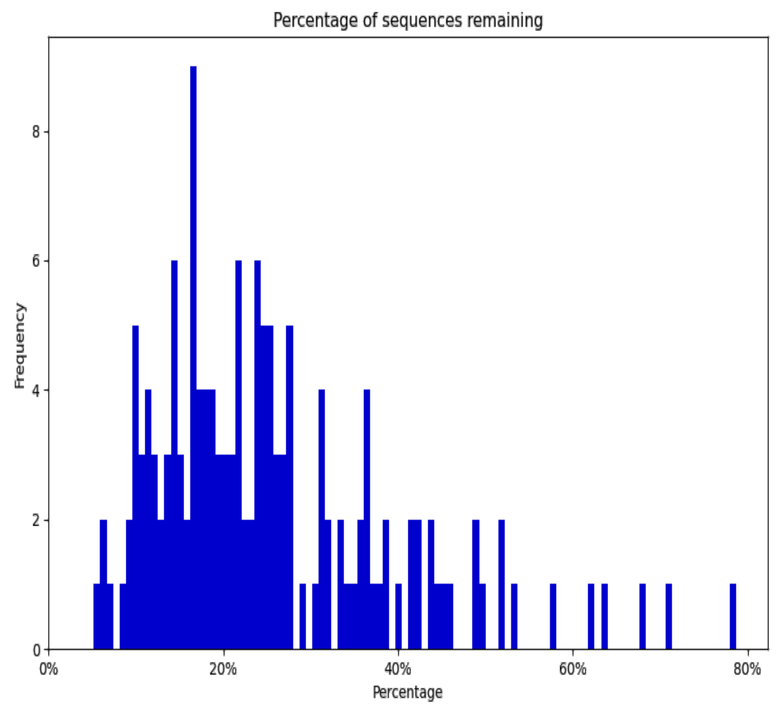
BioShell software reduce an overrepresentation of sequences, **increase quality** of a **MSA,** build **better sequence profile.**

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.

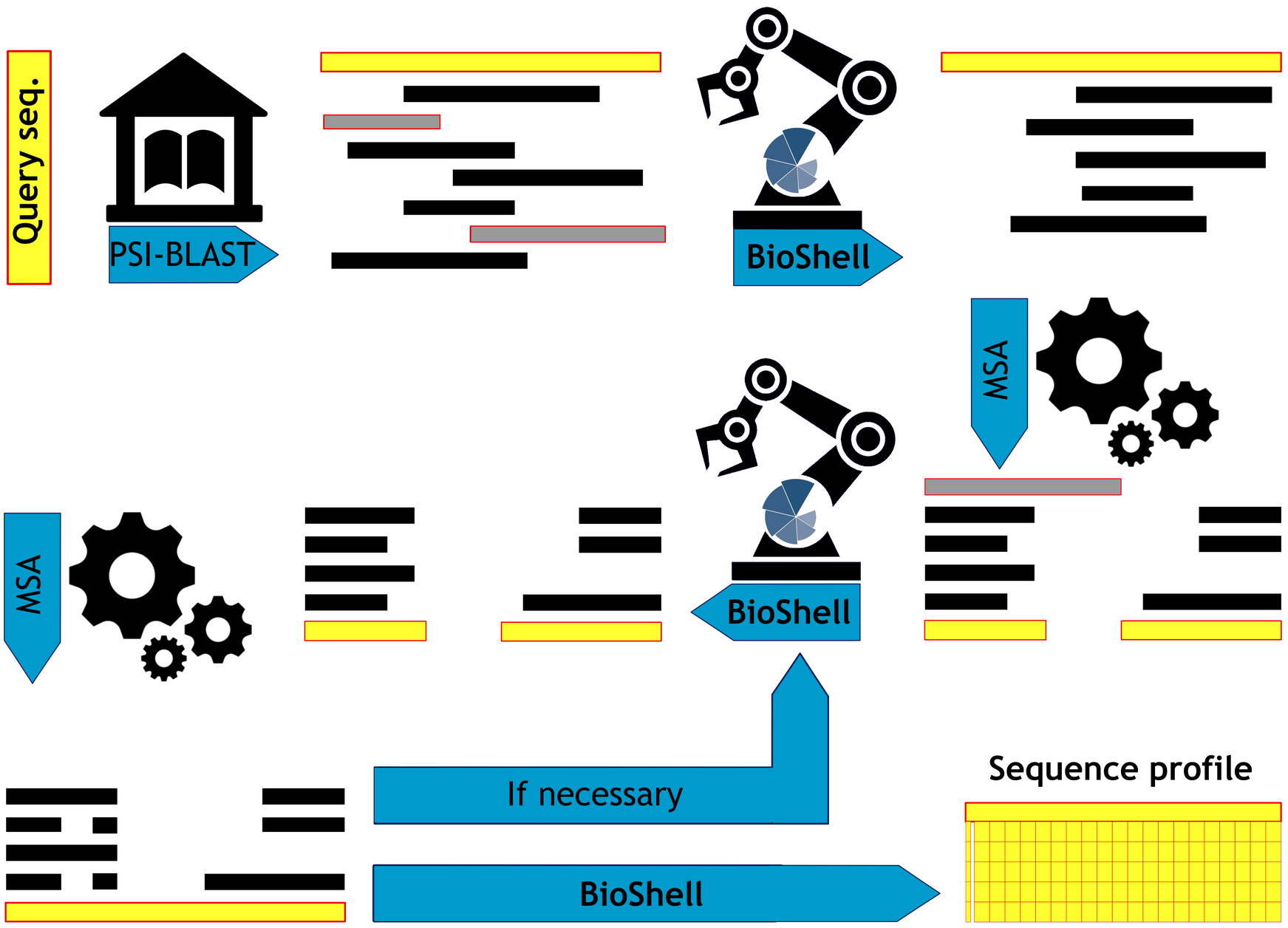
Results

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.

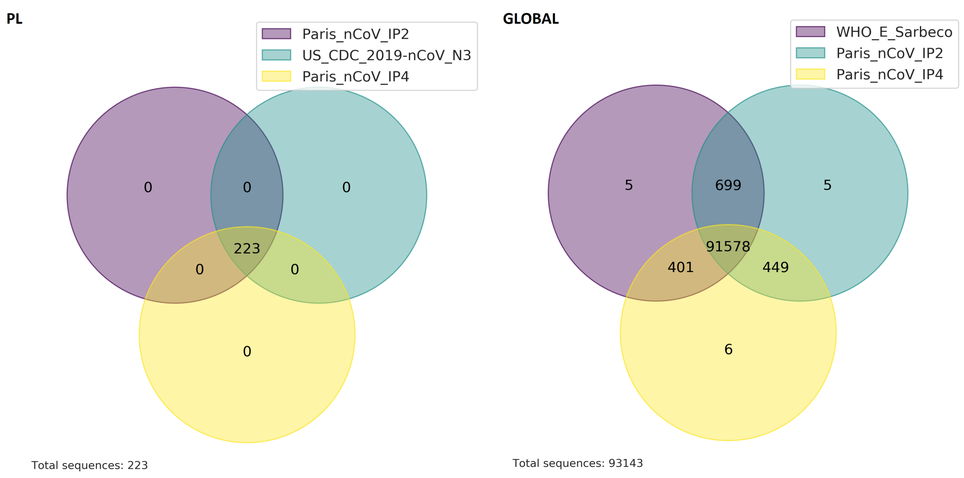


Figure 1: Venn diagram of three best primer pairs for diagnostic purposes of SARS-CoV-2 identification. On the left for Polish Sequences, on the right for global sequences downloaded from *GISAID*.

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

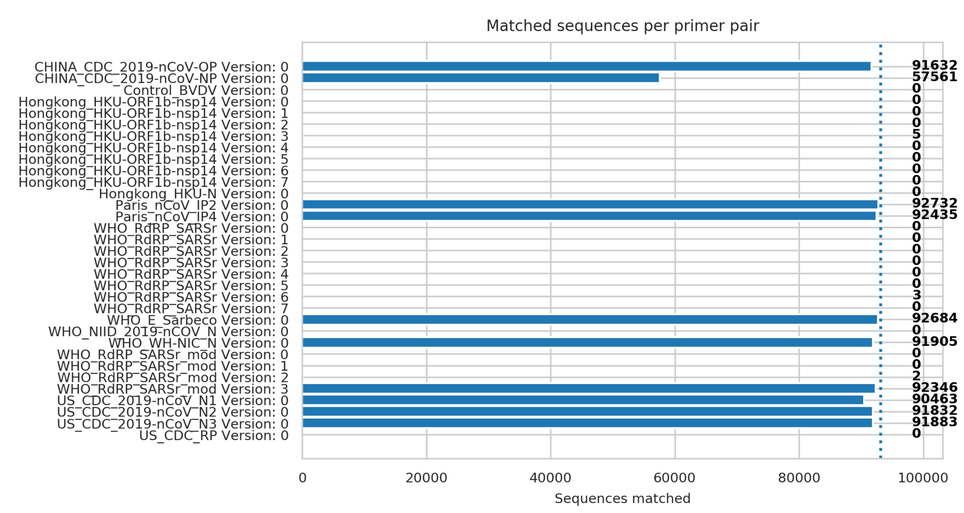


Figure 2: Horizontal bar plot that shows the overall performance of all primer sets. Length of the bars is determined by how much sequences given pair of primers had been able to match conservatively.

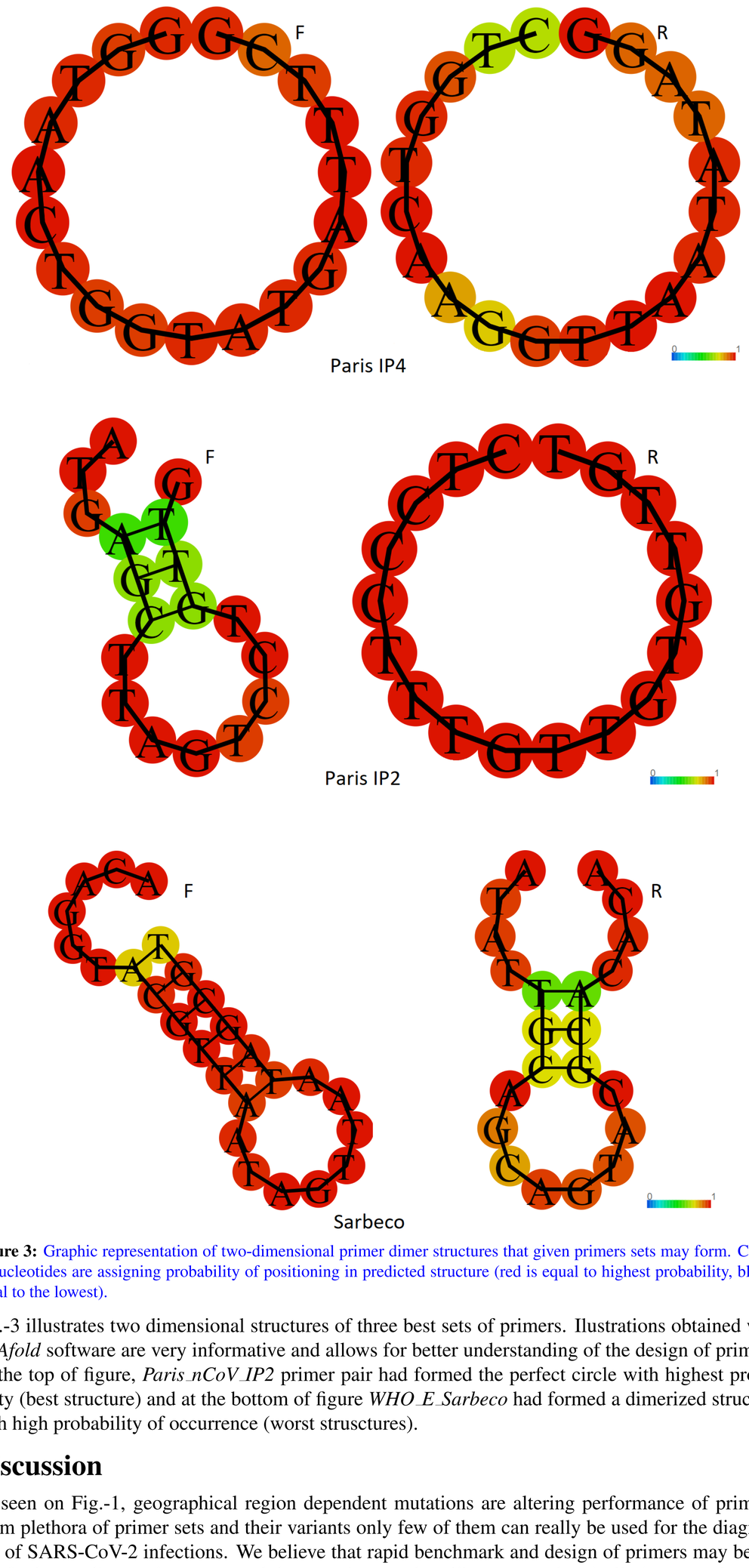


Figure 3: Graphic representation of two-dimensional primer dimer structures that given primers sets may form. Colors of nucleotides are assigning probability of positioning in predicted structure (red is equal to highest probability, blue is equal to the lowest).

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.