Csaba Kerepesi

DATA MINING IN GENOMICS, METAGENOMICS AND CONNECTOMICS

PhD Thesis

Supervisor: Dr. Vince Grolmusz

Department of Computer Science

Eötvös Loránd University, Hungary

PhD School of Computer Science

Eötvös Loránd University, Hungary

Dr. Erzsébet Csuhaj-Varjú

PhD Program of Information Systems

Dr. András Benczúr

Budapest, 2017

Acknowledgements

I would like to thank my supervisor Dr. Vince Grolmusz for his tireless support with which he started my scientific career.

I am very grateful to the PhD School of Computer Science and the Faculty of Informatics, ELTE for their inexhaustible support.

I would like to thank my co-authors for their precious work and I also would like to thank all my colleagues, who helped me anything in my researches.

Finally, I would like to thank my family for their everlasting support.

Contents

Chapter 1

Introduction

Biological data and the number of studies generating them grows exponentially [11], [12]. Altough most of such data are deposited and freely available in biological databases, only a small of them re-analyzed by researchers independent from the source research group. In turn complexity of life implies biological data have huge potential for discoveries. People comes other disciplines can look data in other aspects and using other methods resulted new explorations. Moreover as science are developing new questions emerge and the answers for them may be in data deposited earlier by a study group who searced answers for completely different questions. Data mining is a discipline specialized extracting knowledge from large data sets [13] In the dissertation we collect our explorations in various biological data sets (namely genomes, metagenomes and connectomes) and biological data mining tools developed by us. In the further part of this chapter we sortly summaries these results and methods.

DNA sequencing technologies are applied widely and frequently today to describe metagenomes, i.e., microbial communities in environmental or clinical samples, without the need for culturing them. Phylogenetic analysis of

the metagenomic data presents significant challenges for the biologist and the bioinformatician. The program suite AMPHORA and its workflow version are examples of publicly available software that yield reliable phylogenetic results for metagenomic data.

We have developed AmphoraNet, an easy-to-use webserver that is capable of assigning a probability-weighted taxonomic group for each phylogenetic marker gene found in the input metagenomic sample; the webserver is based on the AMPHORA2 workflow. We believe that the occasional user may find it comfortable that, in this version, no time-consuming installation of every component of the AMPHORA2 suite or expertise in Linux environment are required. The webserver is freely available at `http://amphoranet.pitgroup.org`.

The results mentioned above are detailed in the subsection 2.1 based on the paper [1].

We have developed a visual analysis tool that is capable of demonstrating the quantitative relations gained from the output of the AMPHORA2 program or the easy–to–use AmphoraNet webserver. Our web-based tool, the AmphoraVizu webserver, makes the phylogenetic distribution of the metagenomic sample clearly visible by using the native output format of AMPHORA2 or AmphoraNet. The user may set the phylogenetic resolution (i.e., superkingdom, phylum, class, order, family, genus, species) along with the chart type, and will receive the distribution data, detailed for all relevant marker genes in the sample. The visualization webserver is available at the address `http://amphoravizu.pitgroup.org`. The source code of the AmphoraVizu program is available at `http://pitgroup.org/apps/amphoravizu/AmphoraVizu.pl`.

The results mentioned above are detailed in the subsection 2.2 based on

the paper [2].

DNA sequencing technologies usually return short (100-300 base-pair long) DNA reads, and these reads are processed by metagenomic analysis software that assign phylogenetic composition-information to the data set. We have evaluated three metagenomic analysis software (AmphoraNet, MG-RAST and MEGAN5) for their capabilities of assigning *quantitative* phylogenetic information for the data, describing the frequency of appearance of the microorganisms of the same taxa in the sample. The difficulties of the task arise from the fact that longer genomes produce more reads from the same organism than shorter genomes, and some software assigns higher frequencies to species with longer genomes than to those with shorter ones. This phenomenon is called the "genome length bias".

Dozens of complex artificial metagenome-benchmarks can be found in the literature. Because of the complexity of those benchmarks, it is usually difficult to judge the resistance of a metagenomic software to this "genome length bias". Therefore, we have made a simple benchmark for the evaluation of the "taxon-counting" in a metagenomic sample: we have taken the same number of copies of three full bacterial genomes of different lengths, break them up randomly to short reads of average length of 150 bp, and mixed the reads, creating our simple benchmark. Because of its simplicity, the benchmark is not supposed to serve as a mock metagenome, but if a software fails on that simple task it will surely fail on most real metagenomes.

We applied three software for the benchmark. The ideal quantitative solution would assign the same proportion to the three bacterial taxa. We have found that AMPHORA2/AmphoraNet gave the most accurate results and the other two software were under-performers: they counted quite reliably each short read to their respective taxon, producing the typical genome

length bias.

The results mentioned above are detailed in the subsection 2.3 based on the paper [3].

The first giant virus was identified in 2003 from a biofilm of an industrial water-cooling tower in England. Later, numerous new giant viruses were found in oceans and freshwater habitats, some of them having even 2,500 genes. We have developed a bioinformatics software called the "Giant Virus Finder" that is capable of discovering the very likely presence of the genomes of giant viruses in metagenomic shotgun-sequenced datasets. The new workflow is applied to numerous hot and cold desert soil samples as well as some tundra- and forest soils. We show that most of these samples contain giant viruses and especially many were found in the Antarctic dry valleys. The results imply that giant viruses could be frequent not only in aqueous habitats but in a wide spectrum of soils on our planet. The Giant Virus Finder software is available at the address `http://pitgroup.org/giant-virus-finder`.

The results mentioned above are detailed in the subsection 2.4 based on the paper [4].

The Kutch desert (Great Rann of Kutch, Gujarat, India) is a unique ecosystem: in the larger part of the year it is a hot, salty desert that is flooded regularly in the Indian monsoon season. In the dry season, the crystallized salt deposits form the "white desert" in large regions. The first metagenomic analysis of the soil samples of Kutch was published in 2013, and the data was deposited in the NCBI Sequence Read Archive. At the same time, the sequences were analyzed phylogenetically for prokaryotes, especially for bacterial taxa.

We have been searching for the DNA sequences of the recently discovered

giant viruses in the soil samples of the Kutch desert. Since most giant viruses were discovered in biofilms in industrial cooling towers, ocean water and freshwater ponds, we were surprised to find their DNA sequences in the soil samples of a seasonally very hot and arid, salty environment.

The results mentioned above are detailed in the subsection 2.5 based on the paper [5].

Fine-tuned regulation of the cellular nucleotide pools is indispensable for faithful replication of DNA. The genetic information is also safeguarded by DNA damage recognition and repair processes. Uracil is one of the most frequently occurring erroneous base in DNA; it can arise from cytosine deamination or thymine-replacing incorporation. Two enzyme families are primarily involved in keeping DNA uracil-free: dUTPases that prevent thymine-replacing incorporation and uracil-DNA glycosylases that excise uracil from DNA and initiate uracil-excision repair. Both dUTPase and the most efficient uracil-DNA glycosylase UNG is thought to be ubiquitous in free-living organisms.

We have systematically investigated the genotype of deposited fully sequenced bacterial and archaeal genomes. Surprisingly, we have found that in contrast to the generally held opinion, a wide number of bacterial and archaeal species lack the dUTPase gene(s). The dut- genotype is present in diverse bacterial phyla indicating that loss of this (or these) gene(s) has occurred multiple times during evolution. We have identified several survival strategies in the lack of dUTPases.

The results mentioned above are detailed in the subsection 2.6 based on the paper [6].

The human braingraph or the connectome is the object of an intensive research today. The advantage of the graph-approach to brain science is that

the rich structures, algorithms and definitions of graph theory can be applied to the anatomical networks of the connections of the human brain. In these graphs, the vertices correspond to the small (1-1.5 cm^2) areas of the gray matter, and two vertices are connected by an edge, if a diffusion-MRI based workflow finds fibers of axons, running between those small gray matter areas in the white matter of the brain.

The connectomes of different human brains are pairwise distinct: we cannot talk about an abstract "graph of the brain". Two typical connectomes, however, have quite a few common graph edges that may describe the same connections between the same cortical areas.

We have developed the Budapest Reference Connectome Server v2.0 which generates the common edges of the connectomes of 96 distinct cortexes, each with 1015 vertices, computed from 96 MRI data sets of the Human Connectome Project. The user may set numerous parameters for the identification and filtering of common edges, and the graphs are downloadable in both csv and GraphML formats; both formats carry the anatomical annotations of the vertices, generated by the FreeSurfer program. The resulting consensus graph is also automatically visualized in a 3D rotating brain model on the website.

The consensus graphs, generated with various parameter settings, can be used as reference connectomes based on different, independent MRI images, therefore they may serve as reduced-error, low-noise, robust graph representations of the human brain. The webserver is available at http://connectome.pitgroup.org.

The results mentioned above are detailed in the subsection 3.1 based on the paper [7].

We have constructed 1015-vertex graphs from the diffusion MRI brain

images of 395 human subjects and compared the individual graphs with respect to several different areas of the brain. The inter-individual variability of the graphs within different brain regions was discovered and described.

We have found that the frontal and the limbic lobes are more conservative, while the edges in the temporal and occipital lobes are more diverse. Interestingly, a "hybrid" conservative and diverse distribution was found in the paracentral lobule and the fusiform gyrus. Smaller cortical areas were also evaluated: precentral gyri were found to be more conservative, and the postcentral and the superior temporal gyri to be very diverse.

The results mentioned above are detailed in the subsection 3.2 based on the paper [8].

One main question of connectomics today is discovering the directions of the connections between the small gray matter areas. Our previous work, the Budapest Reference Connectome Server, generates the consensus braingraph of 96 subjects in Version 2, and of 418 subjects in Version 3, according to selectable parameters. After the Budapest Reference Connectome Server had been published, we recognized a surprising and unforeseen property of the server. The server can generate the braingraph of connections that are present in at least k graphs out of the 418, for any value of $k = 1, 2, \ldots, 418$. When the value of k is changed from $k = 418$ through 1 by moving a slider at the webserver from right to left, certainly more and more edges appear in the consensus graph. The astonishing observation is that the appearance of the new edges is not random: it is similar to a growing tree. We refer to this phenomenon as the dynamics of the consensus connectomes.

We hypothesize that this movement of the slider in the webserver may copy the development of the connections in the human brain in the following sense: the connections that are present in all subjects are the oldest ones,

and those that are present only in a decreasing fraction of the subjects are gradually the newer connections in the individual brain development.

Based on this observation and the related hypothesis, we can assign directions to the edges of the connectome as follows: Let G_{k+1} denote the consensus connectome where each edge is present in at least $k+1$ graphs, and let G_k denote the consensus connectome where each edge is present in at least k graphs. Suppose that vertex v is not connected to any other vertices in G_{k+1}, and becomes connected to a vertex u in G_k, where u was connected to other vertices already in G_{k+1}. Then we direct this (v, u) edge from v to u.

The results mentioned above are detailed in the subsection 3.3 based on the paper [9].

Chapter 2

Data Mining in Genomics and Metagenomics

2.1 AmphoraNet: The Webserver Implementation of the AMPHORA2 Metagenomic Workflow Suite

2.1.1 Introduction

Next generation sequencing technologies and the parallel development of high throughput short-read assembly methods make possible to view ourselves and our living environment quite differently than before [36, 43, 51, 56]. Metagenomics methods yield tools to discover and identify microorganisms in diverse clinical and environmental samples, without the need of culturing them [39]. These methods may shed light to the system of interactions between human and microbial cells that may lead to or prevent from diseases such as type 1 and type 2 diabetes [25, 31, 59, 68, 72, 85], oral- and colorectal can-

cers [28, 30, 34, 36, 40, 71, 77], autoimmune syndromes [29, 49, 52, 64, 76, 89] or obesity [37, 42, 50, 84, 88], just to list a few examples.

The results mentioned above from the last 2-3 years imply that in the next decade metagenomics will be an area of massive development both in biology and bioinformatics. Recognizing this trend, complex bioinformatical workflows were developed for analyzing metagenomics data, and assigning phylogenetic attributes to short nucleotide reads, coming from diverse species and environments [26, 62, 63, 66, 67, 92].

One of the successful approaches is the AMPHORA [90] suite of programs, together with its improved workflow version, called AMPHORA2 [87, 91].The AMPHORA2 workflow was already applied by numerous studies [41, 44, 78, 79]. Both AMPHORA and AMPHORA2 make use of several components of previously developed tools, such as getorf from EMBOSS [74], the HMMER sequence-search and alignment tool [45, 58], BioPerl components [80, 81] and RAxML [82, 83]. AMPHORA searches for phylogenetic marker genes with HMMER [45, 58], and makes suggestions for their phylogenetic placements using RAxML and a reference database. More detailed description of AMPHORA is in [90] and of AMPHORA2 is in [91].

2.1.2 Results and Discussion

While an installation script is supplied with AMPHORA2 [91], the proper installation of the numerous components of AMPHORA2 is not always an easy task, especially, if older versions of some components were previously installed, or the installation is done without superuser privileges. If the user is not familiar with Linux systems, the local installation is definitely a challenge. In order to make available the capabilities of this great suite of programs for more scientists, we prepared a web-server version of the

AMPHORA2 workflow under the address `http://amphoranet.pitgroup.org`. Phylogenetic analysis is a resource-hungry task, so if one needs to use AMPHORA2 on a daily basis with large amounts of data we suggest to install and use the programs locally. For occasional users, or just those who want to test the capabilities of AMPHORA2 quickly, our webserver can be a valuable tool.

The AmphoraNet webserver does not require any registration, and no e-mail address of the user is solicited. The user chooses between nucleotide and amino-acid sequences, specifies if bacterial or archaeal marker gene sequences are to be searched for, and then simply uploads the file in FASTA format, by clicking a button labeled "Check values". Next, the server verifies whether the file-size is under the limit allowed, and outputs basic characteristics of the file. Then the user may start the fully automatic workflow by clicking the "Schedule your job" button. Next, a unique web page is created, that will contain the results of the run. Since the processing may take more than 20 minutes, the user is advised to bookmark that unique web page, and return later to the output of the job. The identity of that unique web page is known only for the user; this feature allows moderate privacy for the users (since we do not require registration, we are not able to implement sophisticated access control measures).

Currently, there is a 50 MB upload limit for every single file into the AmphoraNet webserver. Larger jobs can be uploaded in several parts, as it is suggested in the support forum of the AmphoraNet user community: `https://groups.google.com/forum/#!forum/amphoranet`.

Sample Datasets

For the convenience of the users, we gave several sample input- and output files on the support forum of the server: `https://groups.google.com/forum/#!topic/amphoranet/SbsSWm6wVx8`

- For a complete bacterial genome, *Treponema pallidum subsp. pallidum DAL-1*, of 1.1 million base pairs (bp), AmphoraNet finishes in 30 minutes;
- For a complete archaeal genome, *Archaeoglobus profundus DSM 5631*, of 1.5 million bp, AmphoraNet gives a result in 15 minutes;
- For a sample from the Human Microbiome project (Buccal Mucosa sample (SRS050007)), of 0.7 million bp, AmphoraNet finishes in 20 minutes.

2.2 Visual Analysis of the Quantitative Composition of Metagenomic Communities: the AmphoraVizu Webserver

2.2.1 Introduction

Metagenomic communities contain numerous known and unknown bacterial and archaeal species. The DNA of most of the unknown species will probably not be sequenced in the next several years. Consequently, we can infer phylogenetic information on the metagenomes only from the highly inhomogeneous short DNA reads gained from next generation sequencing methods [43,51,56].

Clearly, with techniques available to date, detailed classification is not possible for samples containing hundreds of unknown species. The probabilistic inference of higher level taxa is, however, possible by comparing the nucleotide sequences of unknown species and already identified species from standard repositories.

One possible method is applying sequence alignment tools (e.g., BLAST and its clones) between the translated short reads found in the sample and the reference protein sequence databases (e.g., the MEGAN suite applies this approach, [53–55, 104]). An alternative way is looking for some predefined phylogenetic marker genes in the sample, and using these genes for phylotyping (e.g., AMPHORA in [90] and of AMPHORA2 in [91] or [1]).

The 31 phylotyping marker genes that were chosen in AMPHORA [90] are (i) universally present in bacteria, (ii) most of them are single copy genes in the known bacterial genomes, and (iii) they are housekeeping genes that are relatively recalcitrant to lateral gene transfer [57]. Because of they are mostly single-copy genes in genomes, one may infer quantitative relations by counting them for each taxa identified. Since, for bacteria, only 31 genes are considered, their alignment and HMM profile search is fast compared to the speed of the BLAST pre-processing needed for MEGAN: AMPHORA compares only these marker genes to the reference genomes, while similarity based methods compare every single contig to the reference genomes. The probability that the short reads in a metagenome contain several fragments from these 31 genes is much higher than for a smaller set of possible marker genes (e.g., where only 16S ribosomal RNA was used).

The AmphoraNet [1] is an easy–to–use webserver implementation of the AMPHORA2 suite of programs. It is capable of inferring phylogenetic information from metagenomic sets of short reads. Until now a graphical quan-

titative analysis tool for the textual output, generated by AMPHORA2 or AmphoraNet, was missing.

2.2.2 Results and Discussion

Here we present a graphical analysis webserver, called AmphoraVizu, that returns publication-quality charts with phylogenetic classifications according to marker genes identified in the sample.

The AmphoraVizu webserver does not require any registration, and no e-mail address of the user is solicited. The user needs to upload the AmphoraNet (`http://amphoranet.pitgroup.org`) or the AMPHORA2 output file to the visualization tool, and then has to specify the phylogenetic resolution of the chart by entering the lowest taxonomic rank requested; the following options are provided: superkingdom, phylum, class, order, family, genus, species. After choosing the chart type (i.e., bar chart or pie chart), by hitting the "Visualize" button, the graph is drawn.

The bar chart version (e.g., Figure 2.1) visualizes the phylogenetic distribution of the sample according to each marker gene in two modes: it computes either the relative frequencies or the absolute numbers of the identified genes.

Using the "Advanced Options" button, it is possible to filter the results according to minimum confidence [91] and minimum average, where the average height of the marker gene bars are computed for each identified phylogenetic unit separately. Therefore, the AmphoraVizu page is a visualization extension of the easy-to-use AmphoraNet webserver [1] that also facilitates to analyze the phylotyping distribution of the sample.

The source code of the AmphoraVizu program is available for download at `http://pitgroup.org/apps/amphoravizu/AmphoraVizu.pl`.

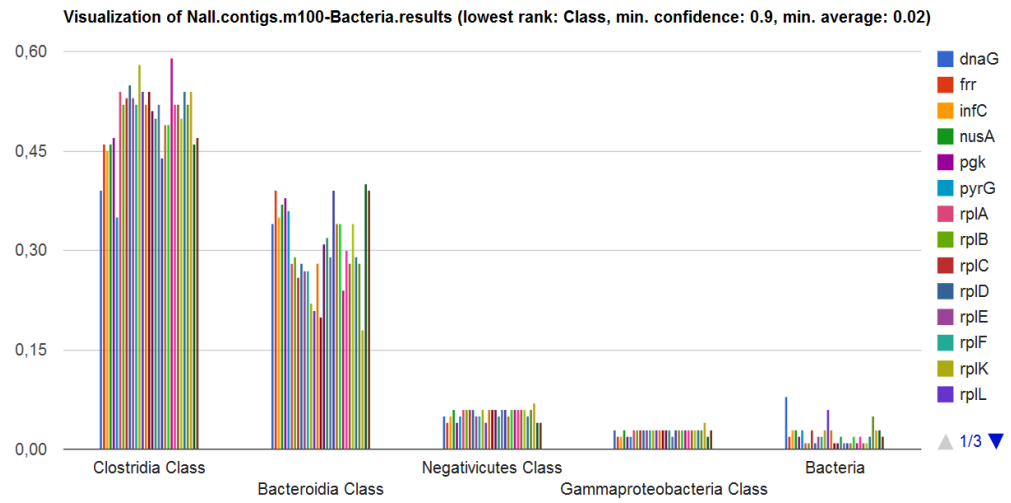

Figure 2.1: Screenshot of the chart generated by AmphoraVizu for the AMPHORA2 processing of the union of control (non-diabetic) gut metagenome datasets of the study [72]. On the right, the color codes of the marker genes are presented. The bar chart gives the distribution of bacteria into classes according to each marker gene; bacteria in unspecified classes are placed in the last group. The height of the bars represents the ratio of a given marker gene identified belonging to the labeled class. Only the taxa reaching the average height 0.02 are represented (this cut-off value can be modified by the advanced option "Minimum average to show"). Note that the sum of the heights of all bars of the same color is 1, except when some of them is missing due to the cut-off "Minimum average to show" value. The metagenome data was downloaded from `http://gigadb.org/dataset/100036`.

2.3 Evaluating the Quantitative Capabilities of Metagenomic Analysis Software

2.3.1 Introduction

Metagenomic analysis software, like MG-RAST [66], MEGAN [53,104], AMPHORA [90], AMPHORA2 [91], AmphoraNet [1], AmphoraVizu [2] are capable of inferring phylogenetic classification from raw metagenomic data. In the analysis of the metagenomes, we are interested in the detection and identification of the species or genera in the data, and very frequently, we need to know their phylogenetic distribution: that is, the fraction of bacterial cells in each taxon screened.

If the lengths of all genomes in the sample were the same, then this task would be relatively easy: one has to make the phylogenetic assignment to each read and evaluate the result. If the lengths of the genomes vary, then it is likely that short reads of the longer genomes are identified more frequently than short reads from the shorter ones, simply, because the reads from the long genome appear more frequently than those from the shorter ones. Consequently, for quantitative metagenomic analysis we need to use software that is capable of such tasks.

AMPHORA [90] and AMPHORA2 [91] applies marker genes for phylogenetic inference, and these marker genes are chosen to appear just once in the known bacterial genomes. Therefore, both short and long genomes will be counted just once. Naturally, the still unknown bacterial genomes could not be scanned for validating this property.

The MEGAN suite [53,104] applies BLAST search for the individual short reads, and attempts to identify those short reads phylogenetically. Therefore, MEGAN and similar methods will identify short reads from long genomes

more frequently than short reads from short genomes.

In the present work we demonstrate the hypothesis above on an example: we created a benchmark from three known bacterial genomes of different lengths, and found that AMPHORA2 [91], and its webserver implementation, AmphoraNet [1] worked very precisely in assigning quantitative phylogenetic information to the test data. Our benchmark is not intended to use as a general-purpose simulated metagenome, it was created *only* for the fast and straightforward analysis of the quantitative capabilities of the metagenomic annotation software.

Clearly, if a software fails on these straightforward and easy-to-evaluate tests, it will fail in numerous – but not necessarily all – real life scenarios as well.

However, the opposite implication is not necessarily true: if a workflow performs well in this simple benchmark, then, in the real life scenarios, where the number of taxa (both known and unknown) could be large and the coverage of the genomes by the short reads can fluctuate wildly from taxon to taxon, the workflow could fail to be quantitative. The reason for this phenomenon is that the marker genes, which are looked for in AMPHORA2, will not be found typically in genomes with very low coverage.

2.3.2 Results and discussion

Artificially, *in silico* created metagenome benchmarks are frequently used for testing metagenomic analysis software [65, 75]. We prepared a simple benchmark to evaluate the quantitativity of some software workflows in this work.

The four artificial metagenomes in the benchmark were constructed as described in the Methods section.

In Dataset 1, we have taken the same count of genomes of three species (*H. pylori, B. bacteriovorus, D. carboxyvorans* of different genome lengths, so the correct distribution expected is 33.3%-33.3%-33.3%. The results of the tests are demonstrated in Figure 2.2 and Table 2.1.

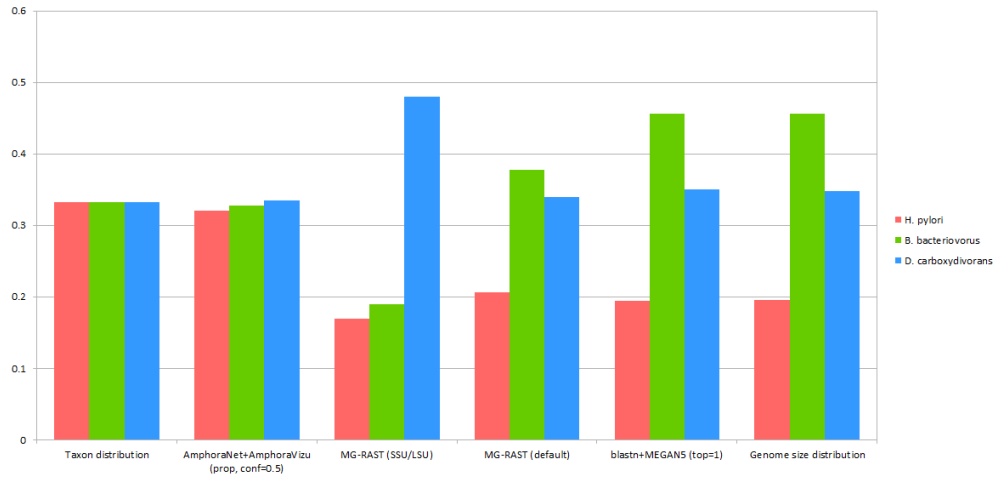

Figure 2.2: Summary of the results of the phylotyping software on the genus level: From left to right: The ideal annotation would give the same proportion for the three bacteria; the result of AmphoraNet with the AmphoraVizu evaluation; results of MG-RAST both with the rRNA option and the default settings; the result of the blastn annotation with MEGAN5; the short read distribution between taxa. Numerical results with more parameters are given in Table 2.1.

In Dataset 2, we have taken the same count of genomes of *H. pylori* and *B. bacteriovorus*, and twice as many from *D. carboxyvorans*; therefore, the correct distribution expected is 25%-25%-50%. The results of the tests are given in Figure 2.3.

In Datasets 3 and 4 we have taken genome-count distributions of the species to be 25%-50%-25% and 50%-25%-25%, resp. The results of the tests

	H. pylori	B. bacter.	D. carb.	SQ
AmphoraVizu				
Species (prop, conf=0.5)	0.312	0.328	0.201	0.01569
Species (prop, conf=0.1)	0.315	0.328	0.307	0.00106
Genus (prop, conf=0.9)	0.319	0.328	0.321	0.00039
Genus (prop, conf=0.5)	0.320	0.328	0.335	0.00021
Species (amount, conf=0.9)	0.169	0.330	0.153	0.05954
Species (amount, conf=0.1)	0.267	0.330	0.289	0.00638
Genus (amount, conf=0.9)	0.285	0.330	0.323	0.00245
Genus (amount, conf=0.5)	0.288	0.330	0.343	0.00216
MG-RAST				
Genus	0.206	0.368	0.285	0.01975
Species	0.207	0.378	0.339	0.01799
Genus (SSU/LSU)	0.170	0.190	0.480	0.06590
MEGAN5				
Species (top=1)	0.195	0.456	0.349	0.03443
Species (top=10)	0.193	0.458	0.348	0.03545
Genus (top=1)	0.194	0.456	0.350	0.03474
Genome size distribution	0.196	0.456	0.348	0.03412
Taxon distribution	0.333	0.333	0.333	0

Table 2.1: The frequencies detected by different software for different phylotypes. For AmphoraNet, different confidence settings (conf) and two quantifying methods are applied in the AmphoraVizu evaluation of the results: "proportion" (prop) and "amount". The last column gives the value $\sum_{i=1}^{3}(x_i - \frac{1}{3})^2$, where x_i, $i = 1, 2, 3$, are the numbers of the row. Clearly, SQ is zero if and only if $x_i = \frac{1}{3}$, $i = 1, 2, 3$. For the taxon distribution (in the last row) SQ is zero; the closer is SQ to zero, the better is the result computed.

are given in Figures 2.4 and 2.5.

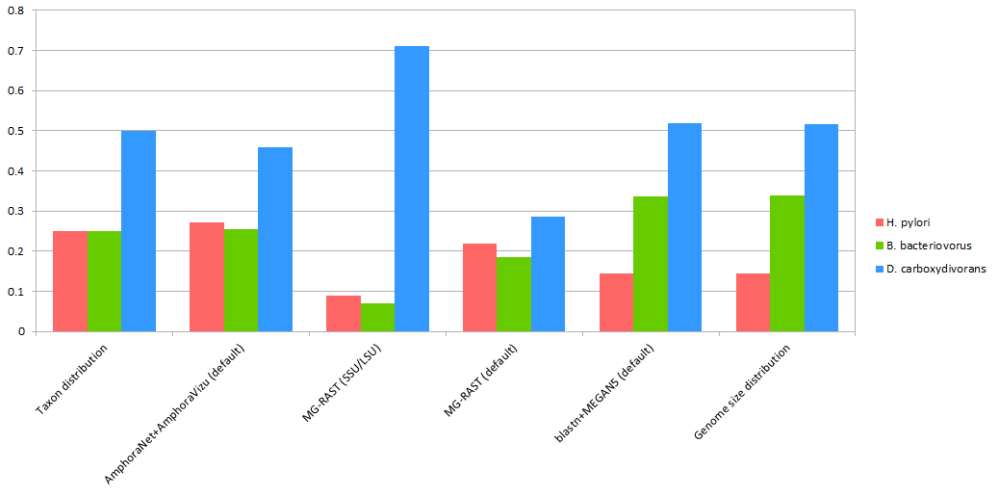

Figure 2.3: The results of the metagenomic analysis software for the Dataset 2 on genus level. Here we have taken the same count of genomes of H. pylori and B. bacteriovorus, and twice as many from D. carboxyvorans; therefore, the correct distribution expected is 25%-25%-50%. (the Taxon distribution). The dataset is available at `http://pitgroup.org/static/2D100kavg150bps.fna`.

The software examined were the webserver implementation of AMPHORA2 [91]: the AmphoraNet [1]; MG-RAST [66] and MEGAN5 [53,104].

2.3.3 Methods

Design of the benchmark. Three bacterial genomes were chosen randomly from the list of full bacterial genomes maintained at the European Bioinformatics Institute Genomes Pages

`http://www.ebi.ac.uk/genomes/bacteria.html`, namely *Bdellovibrio bacteriovorus* HD100, *Desulfotomaculum carboxydivorans* CO-1-SRB and *Helicobacter pylori* Puno120. The genomes have lengths 3,782,950 bp, 2,892,255

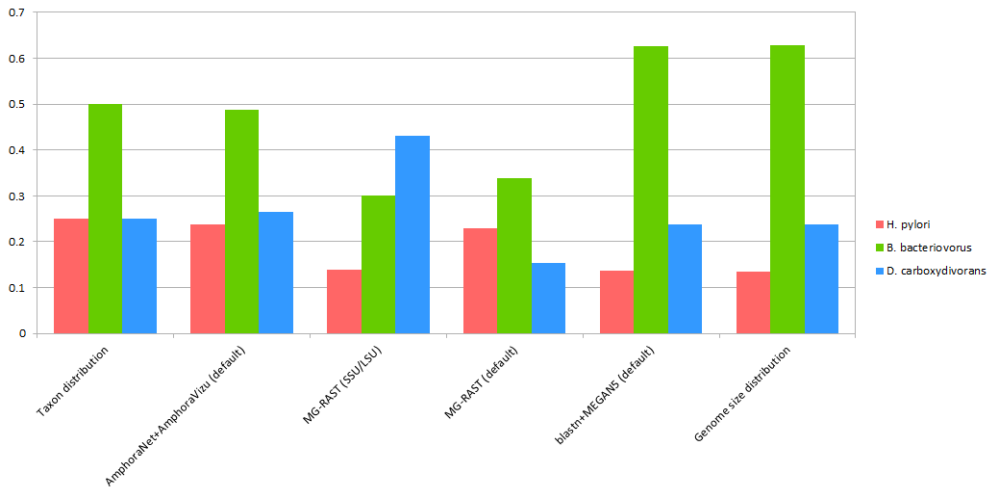

Figure 2.4: The results of the metagenomic analysis software for the Dataset 3 on genus level. Here we have taken the same count of genomes of H. pylori and D. carboxyvorans, and twice as many from B. bacteriovorus; therefore, the correct distribution expected is 25%-50%-25%. (the Taxon distribution). The dataset is available at `http://pitgroup.org/static/2B1D1H100kavg150bps.fna`.

bp, 1,624,979 bp respectively.

Next, MetaSim [75], a shotgun sequencing simulator, was applied to these genomes. 100,000 simulated reads were chosen by MetaSim, each with an expected length of 150 bp and standard deviation of 10.

In Dataset 1, the probability of a read chosen from a given genome was proportional to the length of that genome: this distribution simulates the case when we have the same number of cells, or in other words, the same number of genomes from the three species; for example, simulated reads from *Desulfotomaculum carboxydivorans* were chosen by more than twice more frequently than reads from *Bdellovibrio bacteriovorus*. The exact values are as follows: *B. bacteriovorus* is represented by 45,516 reads, *D. carboxydivoran*

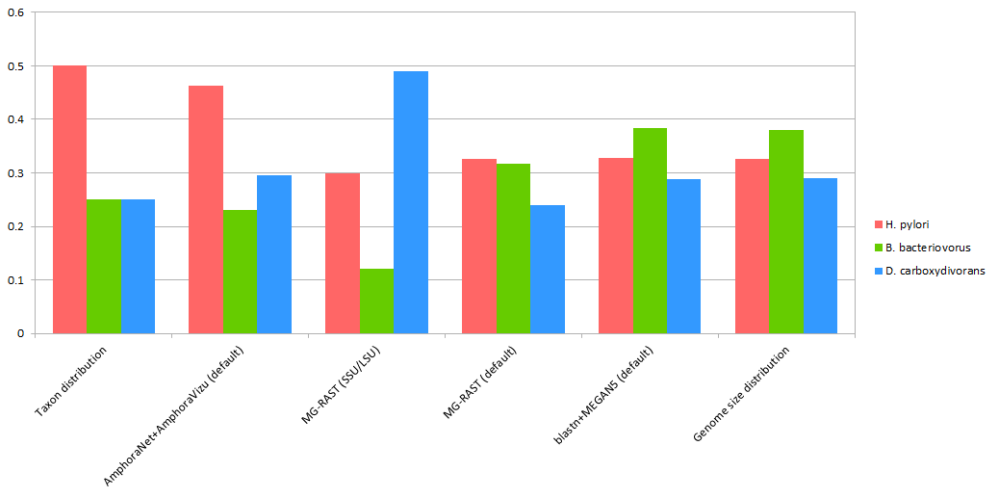

Figure 2.5: The results of the metagenomic analysis software for the Dataset 4 on genus level. Here we have taken the same count of genomes of B. bacteriovorus and D. carboxyvorans, and twice as many from H. pylori; therefore, the correct distribution expected is 50%-25%-25%. (the Taxon distribution). The dataset is available at `http://pitgroup.org/static/2H100kavg150bps.fna`.

by 35,001 reads and *H. pylori* by 19,483 reads. Note that while the bacteria with longer genomes are represented by more reads than those with a shorter genome, the artificial metagenome created describes a community with the same number of each of the three bacteria.

The simulated reads are available for downloading at `http://pitgroup.org/static/3RandomGenome-100kavg150bps.fna`.

In Datasets 2, 3 and 4 we have simulated the scenario when one of the species has twice as many cells as each of the other two.

In Dataset 2 twice as many *Desulfotomaculum carboxydivorans* cells are simulated as each of the other two, in Dataset 3 twice as many *Bdellovibrio bacteriovorus* cells as each of the other two, and in Dataset 4 twice as many

Helicobacter pylori cells are simulated as each of the other two. The links to those sets are given in the "Availability" section, and the results in the Figures 2.3, 2.4 and 2.5.

The application of the benchmark. The webserver implementation of AMPHORA2 [91], AmphoraNet [1], with the evaluation/visualization component AmphoraVizu [2]; MG-RAST [66] and MEGAN5 [53, 104] were applied to the datasets of the benchmark.

The AmphoraNet webserver's running times were between 48m and 52m. The result file was processed by the AmphoraVizu webserver [2]. The results are given in Table 2.1 for Dataset 1.

The MG-RAST webserver [66] was run on the benchmarks both with the default settings (on the "Metagenome Overview" page the "Taxonomic Hits Distribution" section) and with the rRNA-based marker genes (in the "Metagenome Analysis" section, choosing both SSU RNA and LSU RNA databases; the results are denoted by "MG-RAST(SSU/LSU)" on the figures). The default settings were applied (Max. e-Value Cutoff 1e-5; Min. % Identity Cutoff 90%; Min. Alignment Length Cutoff 60); the running time – with the "Data will be publicly accessible immediately after processing completion - Highest Priority" option chosen – was between 1 hours and 1 h 23 m for the datasets. Table 2.1 summarizes the data from the pie chart of the "Taxonomic Distribution" section of the MG-RAST results page for Dataset 1. For other datasets, the results are visualized in the Figures 2.3, 2.4 and 2.5.

MEGAN5 [53,104] was applied as follows: firstly `blastn` was run against the nt nucleotide sequence database of the NCBI, on our local server the processing times were between 1886 m and 2085 m. Next, MEGAN5 was applied for the evaluation of the raw blast file, it was completed in around

10 m. Table 2.1 shows the read distribution among distinct phylotypes, predicted by MEGAN5 for Dataset 1. For other datasets, the results are visualized in the Figures 2.3, 2.4 and 2.5.

In summary, we have constructed a simple artificial benchmark for examining the quantitative capabilities of metagenomic phylotyping software, consisting of four datasets. Our results show that the marker-gene detecting AMPHORA2 pipeline highly outperforms the other software examined. MEGAN5 detected very reliably the phylotypes of the short reads (as seen by comparing the last two rows of Table 2.1), but the percentages there returns the correct proportions of the short-read distribution between the genomes, but unfortunately, not the genome-distribution within the sample.

2.3.4 Availability

The benchmarks, with the marked genome compositions, are available at the following addresses:

Dataset 1 with distribution 1H-1B-1D (results shown in Figure 2.2 above):
`http://pitgroup.org/static/3RandomGenome-100kavg150bps.fna`,

Dataset 2 with distribution 1H-1B-2D (results shown in Figure 2.3):
`http://pitgroup.org/static/2D100kavg150bps.fna`,

Dataset 3 with distribution 1H-2B-1D (results shown in Figure 2.4):
`http://pitgroup.org/static/2B1D1H100kavg150bps.fna`,

Dataset 4 with distribution 2H-1B-1D (results shown in Figure 2.5):
`http://pitgroup.org/static/2H100kavg150bps.fna`.

AmphoraVizu [2] is available at `http://pitgroup.org/amphoravizu/`, its source code at `http://pitgroup.org/apps/amphoravizu/AmphoraVizu.pl`. AmphoraNet [1] is available at `http://amphoranet.pitgroup.org`.

2.4 The "Giant Virus Finder" Discovers an Abundance of Giant Viruses in the Antarctic Dry Valleys

2.4.1 Introduction

The discovery of new giant viruses caused a considerable turmoil in virology in the last decade: these viruses are larger than numerous bacteria and may have even more than 2,500 genes [32,38,73,93]. They are parasitic to amoeba cells living in freshwater reservoirs or seawater habitats. Until now, they were not reported to be found in soil samples or arid environment.

The *Acanthamoeba polyphaga mimivirus* was first found in a cooling tower of Bradford, England in 1992, and was later identified as the first giant virus in 2003 [60]. Its genome consists of 800,000 basis pairs (bp).

Marseillevirus was found in the biofilm of a cooling tower near Paris [33]; its genome contains 368,000 bp.

The *Cafeteria roenbergensis virus (CroV)* was discovered in the seawater off the Texas coast in the early 1990s [46, 47]; its genome contains 730,000 bp.

The *Megavirus chilensis* [27] was discovered in 2010 in a seawater sample off-coast Chile; it has a 1.2 million bp DNA that encodes 1,100 proteins.

Pandoraviruses [70] were discovered in 2013 and they have the largest genome of any viruses known. Their diameter is close to 1 μm. *Pandoravirus salinus* was found in seawater off-coast Chile, and has a 2.5 million bp genome that encodes around 2,500 proteins. *Pandoravirus dulcis* was found in a garden pond in Latrobe University, Melbourne, Australia, has a 1.9 million bp genome.

The Samba virus [35] was found in surface water samples of the Amazon river system in Brazil. Its 1,200,000 bp long DNA encodes 938 proteins.

The *Pithovirus sibericum* was identified in a thirty-thousand year old frozen Siberian sample [61]. Its 610 kbp long genome encodes 467 proteins.

The *Mollivirus sibericum* was also identified from the same sample as the *Pithovirus sibericum* [23]. Its genome-size is 651 kbp, and it has 523 protein-coding genes.

It is reported in [48] that DNA strands similar to that of the Mimivirus can be found in the Sargasso sea environmental sequences database [86].

In the present section we re-analyze a dataset published with the article [102], describing the soil microbiota of 16 samples of diverse geographic locations, including the North-American prairie, the Chihuahuan- and the Mojave deserts in New Mexico and California, the Antarctic dry valleys, the Alaskan tundra, and several forests in tropical and temperate regions. The focus of the work of [102] was the thorough metagenomic analysis of 16 environmental samples for bacteria and archaea, enlightening phylogenetic- and functional annotation of the nucleotide sequences found. No detailed analysis was performed for viruses and viral genes.

Applying our new Giant Virus Finder workflow, we have found DNA segments of giant viruses in the samples, implying the very probable presence of giant viruses in these diverse soils.

The Giant Virus Finder

The "Giant Virus Finder" is a general workflow that we have developed for the task of finding giant virus nucleotide sequences in metagenomic samples. The workflow is a collection of scripts with carefully set parameters for BLAST-based searches [103] of short-read metagenomic data sets. The

"Giant Virus Finder" is available at the address `http://pitgroup.org/giant-virus-finder`.

The workflow is presented in detail in the "Methods" section in Figure 2.7. We emphasize here three important features:

(i) We have prepared a list of giant viruses that takes into account only the genome or (if there is no complete genome deposited) sequence size: viruses with 300 kbp or longer genomes or sequences are the members of the list. Clearly, all of the known giant viruses are on the list, but some large viruses, usually not listed as "giants", are also there; e.g., the Canarypox virus, or some large bacteriophages. We note that the user of the method can easily adjust this 300 kbp threshold to arbitrary other value.

(ii) Our method searches for the whole short read (and not only the best-aligned subsequence of the short read), taken from the metagenomic dataset, in the NCBI Nucleotide Collection (nt). This is an important point: if a giant virus is present in the sample, then some short reads come entirely from its genome.

(iii) The word size in the BLAST searches [103] are set cautiously: Too long word size in BLAST searches would not find highly scored non-giant virus sequences in the specificity validation step. Short word sizes, however, increase the precision and also the computational time considerably. We have used $w = 7$ word size in blastn search [103] (instead of the default $w = 28$ word size in Megablast or the $w = 11$ word size in blastn.) In a 16-core server, the running time was a little over four days.

2.4.2 Results and discussion

We have examined the metagenomes collected and deposited with the article [102] for the presence of nucleotide sequences characteristic of giant viruses.

The summary of our results is given in Figure 2.6. A detailed list of the best hits with extremely good E-values are given in Table 2.2.

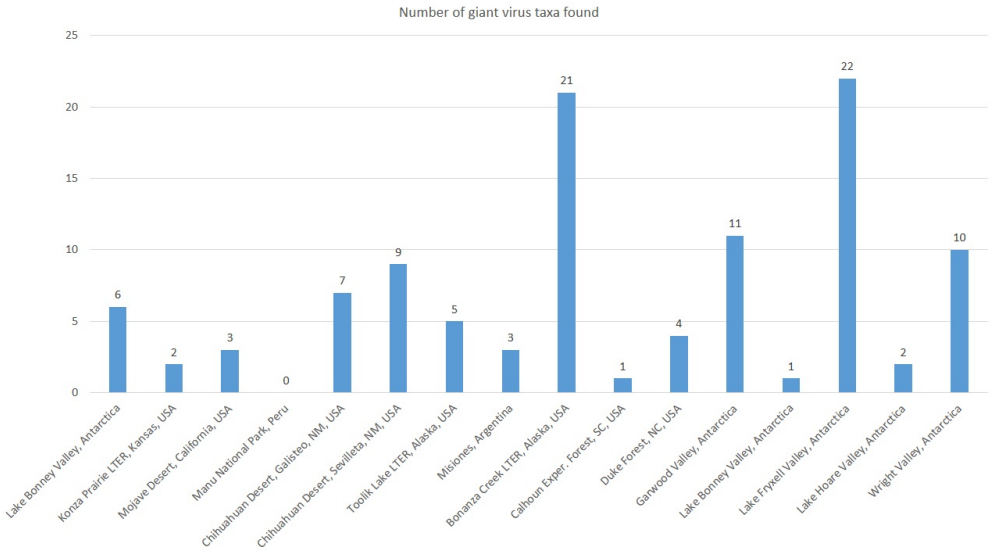

Figure 2.6: Summary of the results of the application of the Giant Virus Finder for the 16 metagenomes of [102]: Each metagenome is denoted on axis x by its geographic location, and the bars visualize the number of the giant virus reads found in the dataset. Detailed results can be found at `http://pitgroup.org/public/giant-virus-finder/Giants-in-16Soil-metagenomes`.

While the "Giant Virus Toplist", defined in the "methods" section, contains large phages and a few other viruses that are usually not considered to be Giant viruses, our top results — measured by E-values and given in Table 2.2 — contains mostly giant viruses when applied to the metagenomes of [102]. For the criterion of assigning a short read to Giant viruses we use

Read identifier	MTG	Location	E-value	Identity	Putative taxa
6:88:18701:16918	803	Lake Bonney Valley	4e-30	91/100	O.Lake phycodnavirus 1
6:47:2094:15918	902	Lake Fryxell Valley	8e-26	87/99	Mimiviridie [family]
7:99:13938:20909	904	Wright Valley	3e-25	86/98	P.bursaria Chlor.virus
4:84:16596:9047	876	Bonanza Creek	1e-24	87/100	Mimiviridie [family]
4:2:19051:10732	876	Bonanza Creek	1e-23	86/100	Mimiviridie [family]
4:114:18824:12821	876	Bonanza Creek	1e-23	86/100	Mimiviridie [family]
4:46:3341:11752	876	Bonanza Creek	1e-22	84/98	Mimiviridie [family]
6:81:6130:14704	803	Lake Bonney Valley	2e-20	83/99	Mimiviridie [family]
6:114:9759:15200	902	Lake Fryxell Valley	3e-19	71/80	Pandoravirus dulc./sal.
4:22:15009:3518	876	Bonanza Creek	3e-19	80/95	Enterobact.[fam.]phage
4:104:7691:17992	901	Lake Bonney Valley	3e-19	83/100	Enterobact.[ord.]phage
6:73:2193:17269	902	Lake Fryxell Valley	4e-18	82/100	Mimiviridie [family]
6:62:15221:2441	803	Lake Bonney Valley	1e-17	72/84	Mimiviridie [family]
6:66:10892:20320	902	Lake Fryxell Valley	4e-17	76/91	Mimiviridie [family]
6:89:6245:20070	900	Garwood Valley	1e-16	81/98	Mimiviridie [family]
6:114:12016:8378	902	Lake Fryxell Valley	5e-16	74/89	Mimiviridie [family]
4:22:17523:8570	876	Bonanza Creek	5e-16	75/91	Mimiviridie [family]
6:79:15305:6160	872	Chihuahuan Desert	5e-16	80/99	Mimiviridie [family]
6:39:10664:8341	900	Garwood Valley	6e-15	72/86	Mimiviridie [family]
7:52:4423:10207	904	Wright Valley	6e-15	60/67	Mimiviridie [family]
7:16:9740:9012	904	Wright Valley	6e-15	73/89	Mimiviridie [family]
4:7:2721:12270	873	Chihuahuan Desert	2e-15	66/75	P.bursaria Chlor.virus
5:83:4473:7350	874	Toolik Lake	2e-15	65/75	Mimiviridie [family]
5:42:4010:17638	899	Duke Forest	2e-15	81/99	C.roenbergensis virus
7:31:3572:1747	904	Wright Valley	2e-15	74/90	Moumovirus

Table 2.2: Best hits, ordered by the E-value, found by applying the Giant Virus Finder for the 16 metagenomes of [102]. **Read identifier:** identifies the read. **MTG:** relevant digits that identify the metagenome. **Location:** Geographic name of the source sample. **E-value:** in Phase 2, the smallest (i.e., best) E-value of the hits found. Identity: the number of identical nucleotides in the best-aligned hit. **Putative taxa:** Assigned taxon using the top 20% rule similarly to the MEGAN LCA algorithm [104].

a MEGAN5-like approach [104]: if every taxon in the top-scored 20% of the Phase 2 alignments are listed in the "Giant Virus Toplist", then we accepted

the read as a giant virus hit.

Samples from Lake Fryxel Valley, Garwood Valley and the Wright Valley, Antarctica, and from Bonanza Creek Forest LTER, Alaska contained the most giant virus taxa. No positive evidence (in the sense described in the "Methods" section) was found for the presence of giant virus DNA fragments in the sample originated from the Manu National Park, Peru.

It is surprising that both hot and cold desert soils contain giant viruses; this finding is in line with our previous result concerning the presence of the giant viruses in the soil samples of the Indian Kutch saline desert [5].

It is worth mentioning that the independent validation of the results presented is easy with the NCBI blastn webserver: one needs to choose a result file which has "GiantVirusFinder-0.2.fasta" filename ending and then needs to feed it into the NCBI blastn webserver selecting the "Somewhat similar sequences (blastn)" program option and setting the word size 7 at the "Algorithm parameters setting" option.

2.4.3 Methods

We believe that the method, presented here, is a general workflow: it could also be applied for identifying other sets of taxa, not only giant viruses. The steps of the general workflow:

(i) Identify the set X of genomes to be searched for (in our application example X is the set of genomes of the giant viruses);

(ii) Apply subsequence-search for the sequences in X in the target metagenomic shotgun sequence database Y (in our example Y is one of the 16 metagenomes of [102]);

(iii) Verify the specificity of the hits: the whole fragments in the metagenomic dataset, containing the highest-scored alignments, are aligned to the sequences of a large nucleotide database. Suppose that the top scored hit has score z. If all the hits with scores greater than $0.8 \times z$ are from the set X, ACCEPT, otherwise REJECT the hit (in our example, the hits are aligned to the sequences of the Nucleotide Collection (nt) of the NCBI; and a hit is accepted only if every sequence in the top-scored 20% belong to set X that is, to the giant virus list).

10% cut-off is applied as a default value in the MEGAN phylogenetic analysis tool [104] for a similar decision. We have found this number is too low for our purpose so we set a more stringent value of 20%. Users can simply change this threshold.

The steps of the method are summarized in Figure 2.7, and in the README file of the GiantVirusFinder-1.1.zip archive on `http://pitgroup.org/giant-virus-finder/latest`.

The Giant Virus Toplist

In the workflow described above, we need a list X of the genomes and sequences of the organisms we are searching for. Defining what is a giant virus and what is not, is a difficult question. We would not like to use potentially questionable and much disputed phylogenetic information in this definition: we simply have constructed the list of viruses with viral genomes or partial genomes (if there is no complete genome deposited) larger than 300 kbp as it is detailed in `http://pitgroup.org/giant-virus-toplist/`. Reference genome data are taken from the `ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/all.fna.tar.gz` file from the NCBI Genome FTP. Note that the length of distinct genome sequences (segments) belonged to a sin-

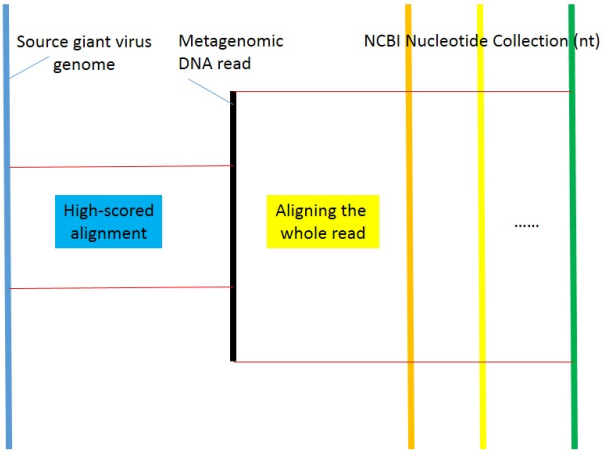

Figure 2.7: Summary of the Giant Virus Finder workflow. First, the giant virus genomes are selected: the selection criterion is a viral genome of a size of at least 300,000 bp (if only a partial genome is deposited, its size needs to be at least 300,000 bp). Next, all genomes of giant viruses are aligned to all DNA short reads in the metagenomic dataset. If a high-scored alignment is found, then the *whole read* that contains the aligned subsequence (and not only the subsequence of the high-scored alignment) is blasted to the whole NCBI Nucleotide Collection (nt). The short read is accepted as a DNA short read from a giant virus if *every sequence* from the top 20% scored hits, found in the NCBI Nucleotide Collection, corresponds to giant viruses.

gle genome are summarized. Other sequences are added from the NCBI Nucleotide database using the search term: *"Viruses"[Organism] AND 300000:10000000[Sequence Length] NOT "Bacteria"[Organism] NOT "Archaea"[Organism].* The list of the viruses found is also given in Table S2 in the supporting material, together with the sequence accession numbers applied in this work.

The inspiration for the Giant Virus Toplist came from `http://www.giantvirus.org/top.html`. Our toplist is more up-to-date and contains not only the complete, but also partial genomes.

Sequence alignments

The metagenomic data of the article [102] is deposited in the MG-RAST archive: `http://metagenomics.anl.gov/metagenomics.cgi?page=MetagenomeProject&project=2997`. We downloaded and converted the files into fastq formats. Next, with the stand-alone BLAST distribution [103] downloadable `makeblastdb` program we created 16 BLAST databases for each of the 16 metagenomes.

In Phase 1 (Figure 2.7) we used the stand-alone UNIX blastn program with the default Megablast algorithm changed the word-size from 28 to 16 and e-value from 10 to 0.01, all the other parameters and the scores and penalties were the default for blastn.

Next, in Phase 2, the hits with better E-value than 0.01 were collected from each alignment, and were aligned using `blastn` with word-size of 7 against the whole Nucleotide Collection (nt) of the NCBI. Suppose that the top scored hit has score z. If all the hits with scores greater than $0.8 \times z$ are from the Giant Virus Toplist, we accepted the hit, otherwise rejected it.

The summary of the results of the two-phase search process with the highest scored giant viruses is given in Figure 2.6. All the files created by the workflow are given at `http://pitgroup.org/public/giant-virus-finder/Giants-in-16Soil-metagenomes/`.

The advantage of the two-phase method

Using a straightforward one phase method (simply blastn all reads against the nt database with the word-size=7 option) would require about 1080 years (about 0,084 h/read) in a machine using a single CPU core. Selecting 9,829 candidate reads from the whole 112,674,624 reads of the 16 metagenomes in Phase 1 reduced the running time to about 34 days in a single-core machine.

Data availability: The metagenomes of the article [102] can be downloaded from `http://metagenomics.anl.gov/metagenomics.cgi?page=MetagenomeProject&project=2997`. The Giant Virus Finder is downloadable from `http://pitgroup.org/public/giant-virus-finder/latest`. The detailed alignment results in both phases of the search are found in `http://pitgroup.org/public/giant-virus-finder/Giants-in-16Soil-metagenomes`.

2.5 Giant Viruses of the Kutch Desert

2.5.1 Introduction

In the present section we analyze the Kutch desert metagenome [69], collected from soil samples with high salinity levels, by "The Giant Virus Finder" workflow detailed in the previous section.

2.5.2 Results and discussion

The results of the two-phase method are given in

`http://pitgroup.org/public/giant-virus-finder/Giants-in-Kutch-metagenomes/phase2-results/`. The top 20 hits are listed in Table 2.3, the numbers of giant viruses found in each metagenome

are visualized in Figure 2.8.

The geographic locations of the metagenomes are given in Figure 2.9.

Read identifier	**Length**	**E-value**	**Identities**	**Putative taxa**
SRR1245949.120967	233	2e-77	209/230	Organic Lake phycodnavirus 1
SRR1245949.1849224	204	1e-59	171/197	Cafeteria roenbergensis virus BV-PW1
SRR1245949.597441	204	1e-59	171/197	Cafeteria roenbergensis virus BV-PW1
SRR1245949.1759145	160	6e-56	145/158	Organic Lake phycodnavirus 1
SRR1245949.1643015	215	2e-46	176/216	Organic Lake phycodnavirus 1
SRR1246239.1961729	241	2e-45	186/239	Moumouvirus Monve isolate Mv13-mv
SRR1245949.1773289	255	1e-42	195/255	Organic Lake phycodnavirus 1
SRR901749.176694	201	7e-38	158/198	Mimiviridae [family]
SRR901747.2102813	234	1e-35	175/229	Bacillus phage G
SRR901747.2102154	218	1e-34	149/188	Phaeocystis globosa virus strain 16T
SRR901749.48809	211	5e-34	164/213	Mimiviridae [family]
SRR901747.677984	127	4e-31	112/130	Phaeocystis globosa virus strain 16T
SRR901749.794757	215	8e-31	155/201	Phaeocystis globosa virus strain 16T
SRR901749.784789	219	4e-29	127/161	Enterobacteriaceae [Family] phage
SRR901749.92543	225	1e-28	142/184	Choristoneura biennis entomopoxvirus 'L'
SRR901747.2414554	131	1e-25	109/133	dsDNA viruses, no RNA stage
SRR1246238.1084710	146	4e-26	117/144	Mimiviridae [family]
SRR901747.1262471	184	2e-24	136/182	Bacillus phage G
SRR901749.1594958	207	3e-24	111/139	Bacillus phage G
SRR901747.197000	159	2e-23	114/145	Bacillus phage G

Table 2.3: The top 20 hits of Giant Virus Finder in the Kutch metagenomes. Note the extremely good E-values of the hits.

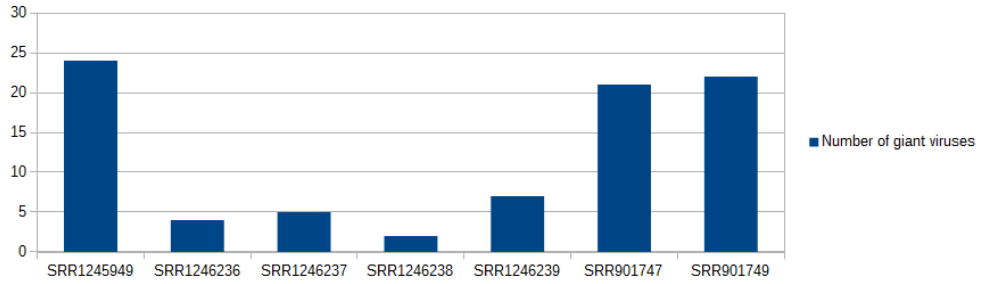

Figure 2.8: The number of giant virus sequences found in the metagenomes of the Kutch desert.

Figure 2.9: Locations of the sample sources. Left to right: red: S5, blue: S4, S6, S7, black: S3, green: S1, orange: S2. (Made with `http://www.copypastemap.com/` and Google Maps).

2.5.3 Materials and Methods

The "Giant Virus Finder" workflow was applied for the search in the metagenomes published in [69]. The "Giant Virus Finder" workflow is described in detail on its web page `http://pitgroup.org/`

`giant-virus-finder`, in [4], and the previous section. Here we give a short overview of the method.

Step 1 A list of the giant viruses was generated, containing virus genomes of size 300 kbp or more; called the "Giant Virus Toplist".

Step 2 Sequential similarities are searched for between the Giant Virus Toplist and the Kutch metagenomes [69] by a blastn search.

Step 3 The best hits of Step 2 were identified, and the metagenomic short read (let us denote it with R), which contained the best hit, is aligned to the whole NCBI Nucleotide Collection (nt) with blastn with a wordsize $w = 7$ instead of the default $w = 28$.

Step 4 Suppose that when we aligned R against the nt database, and the score of the top scored hit from nt was z. Now, we say that R is ACCEPTED as a giant virus DNA segment, if all hits from nt of scores greater than $0.8z$ are from Giant Virus Toplist, otherwise we REJECT.

We note that Step 1 is needed for the speed-up of the process: without Step 1, all the short reads from the metagenome could have aligned to the whole nt database, and the acceptance could have been defined exactly as in Step 4. In this case, however, the running time was several processor hundred years instead of several CPU days with the Giant Virus Finder. More exactly, on a single CPU core, the Giant Virus Finder's phase 2 runs for 17 days, for the 2517 candidate reads. Without phase 1, however, with the $w = 7$ word size, for the 17,401,054 reads, the running time were 319 years.

We also note that setting the word size to $w = 7$ is crucial. From our top 20 hits, with the default word size, 16 give negative results (No significant similarity found). On the other hand, when the specificity is verified in Step

4, we require that the top 20% of the hits are giant viruses. Unfortunately, blast with the default word size would not find numerous non-giant virus hits, and, consequently, would yield false positive results.

For a graphic description of the process, we refer to Figure 2 in [4].

It is worth mentioning that the independent validation of the results presented are easy with the NCBI blastn webserver: choose a result file from here `http://pitgroup.org/public/giant-virus-finder/Giants-in-Kutch-metagenomes/phase2-results/` which has "GiantVirusFinder-0.2-with_hits.txt" filename ending and feed it into the NCBI blastn webserver, choose the "Somewhat similar sequences (blastn)" program option and set the word size $w = 7$ at the "Algorithm parameters" setting and uncheck "Low complexity regions".

Data availability

The metagenomes of the article [69] can be downloaded from the NCBI Short Read Archive, the download links are given in our supplementary Table S1. The Giant Virus Finder is downloadable from `http://pitgroup.org/public/giant-virus-finder/latest`. The detailed alignment results in both phases of the search are found in `http://pitgroup.org/public/giant-virus-finder/Giants-in-Kutch-metagenomes/`.

2.5.4 Conclusions

In the last two section we have shown, by our knowledge at the first time, the very probable presence of giant viruses in diverse environmental soil samples by a two-phase search strategy in metagenomic samples and the NCBI Nucleotide Collection (nt). Our result implies that not only the oceans, biofilms

in cooling towers or small freshwater ponds, but in various non-aqueous environments as the Kutch Desert, the Antarctic dry valleys, the Mojave desert, the prairie and several forest-soil can also accommodate these newly discovered viruses.

2.6 Life without dUTPase

2.6.1 Introduction

The DNA macromolecule is the repository for genomic information in most organisms (with the notable exception of RNA viruses). Stable storage and faithful transmission of genomic information would optimally require a stable macromolecule for these roles. However, the inherent chemical reactivity of DNA and the presence of reactive metabolites and other molecular species within the cell leads to numerous chemical modifications within the DNA even under normal, physiological conditions [113–116]. Mutations arising from these modifications need to be kept under control, and numerous DNA damage recognition and repair processes evolved to deal with these problems [117]. It is also important to mention that mutations are important instruments in driving evolutionary changes and development, as well. Especially for single cell organisms, eminently for bacteria, increased mutational rates leading to new phenotypes may be even advantageous for the species – appearance of antibiotic resistant strains may be a prominent example in this respect [118, 119]. Meanwhile, cells that acquired mutations deleterious for the phenotype will be overgrown by cells with advantageous mutations. In multicellular eukaryotes, such evolutionary changes are more complex since, in these organisms, the viable phenotype is more restricted due to the highly increased interactions within the cellular environment and also with the other

cells/organs.

In response to the need of conserving the DNA-encoded information, a number of specific and highly efficient DNA repair pathways have evolved, such as base-excision repair, nucleotide excision repair, mismatch repair and double-strand break repair [120]. These are strongly conserved from bacteria to man, and the protein factors responsible for these processes are usually ubiquitous, although the cognate protein families and isoforms may differ among organisms of different evolutionary branches. For pathways of key significance, it is also frequently observed that multiple protein families with similar functions are present in one organism to safeguard DNA-encoded information [121]. In addition to the dedicated DNA damage recognition and repair pathways, sanitization and proper balance of the nucleotide pools are also of high importance [122]. Hence, regulation of nucleotide de novo biosynthesis and salvage pathways need to be fine-tuned, and unwanted dNTPs, such as dUTP and dITP have to be removed. Sanitizing enzymes are usually dNTPases catalyzing pyrophosphorolysis of the specific un-orthodox dNTPs [123]. A prominent example in this regard is the dUTPase enzyme family, representatives of which are considered to be ubiquitous and essential for viability in all free-living organisms [115, 124, 125]. There is an intimate cross-talk between enzymes responsible for sanitizing of nucleotide pools and the respective base-excision repair DNA N-glycosylases that act hand in hand first to prevent incorporation of the unwanted nucleotide building block containing modified bases into newly synthesizing DNA and second, to excise those moieties that escaped the preventive measure or got produced within the DNA in situ. For the uracil moiety, the preventive/excising enzyme activities are presented by the dUTPase and the uracil-DNA glycosylase enzyme families, respectively [124–128].

The crosstalk between preventive and excising activities constitutes joint functional efforts with the aim to guard genome integrity. For the dUTPase/UNG enzyme pair, knock-out of the preventive activity of dUTPase is highly dangerous for the cell because it induces numerous uracil-incorporation events that will overload the base excision repair mechanism and transforms it into a hyperactive futile cycle [124,125,129,130]. Knock-out of UNG, however, can be tolerated [131]. In an ung- background, complementing enzyme families with uracil-DNA excising activities (TDG/MUG, SMUG, MPD4 enzyme families) are still functional, although less effective [121, 132]. Also, organisms with uracil-substituted DNA are still viable in lack of UNG, the most efficient uracil-excising enzyme [125, 133].

In a dUTPase knock-out background, viability can be still restored in some cases by simultaneous UNG knock-out [126, 127, 134], or by inhibiting the UNG enzyme with its specific and highly efficient protein inhibitor, UGI. In the double mutant organisms, the uracil content within DNA is highly elevated, however, the cells can survive, most probably since the majority of uracil moieties under these conditions are present as thymine-replacements, i.e., with the same Watson-Crick coding characteristics. Such circumstances have been observed in artificially engineered bacteria (E. coli), or similar situations are also found in specific life stages of wild type Drosophila melanogaster where dUTPase is down-regulated during development and the ung gene is absent from the genome [125, 133].

However, to our knowledge, there is no report published on any free-living organism where the gene for dUTPase is not present within the genome. Our recent observations in several Staphylococcus strains shed light on circumstances where the dUTPase gene on the bacterial chromosome is present only due to insertion of a phage-encoded gene (in prophage form) [128]. A

wide survey of Staphylococcal strains also revealed several occasions where strains are viable and infectious in the absence of dUTPase gene(s) present in the genome, still, these strains are viable [135, 136]. This intriguing situation prompted us to investigate in details the genotypes of prokaryotes and Archaea with respect to the existence of genes primarily involved in uracil-DNA metabolism. Towards this aim, we have analyzed all fully-sequenced bacterial and archaeal genomes deposited in NCBI, that is, 2261 bacterial and 151 archaeal genomic sequence sets. In these investigations, we have specifically looked for the existence or lack of the genes of the dUTPase enzyme families, UNG the most proficient uracil-DNA glycosylase, as well as the genes for the proteins, described up to date as inhibitors of either dUTPase or UNG. Results clearly showed that numerous investigated microbes do not possess dUTPase genes, and this genotype can be paired with different patterns of presence/absence of UNG and inhibitor proteins. We conclude that the genetic distribution of proteins involved in uracil-DNA metabolism is unexpectedly diverse, and these conditions may have physiological consequences.

2.6.2 Materials and Methods

Here we describe the workflow that has generated the list of bacterial and archaeal genomes without dUTPase and from these genomes those with and without UNG, UGI, SAUGI and P56. The list, tables and the source of the in-house programs referred below, are available at the website `http://pitgroup.org/static/life_wo_dutpase/`.

Finding bacterial genomes that do not contain dUTPase

The source of the bacterial and archaeal genome sequences was downloaded from the NCBI FTP site: `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz`. For sequence search and alignment, the standalone UNIX blast program [137] was applied from the site `http://www.ncbi.nlm.nih.gov/books/NBK52640/` on our local servers. Next, with the makeblastdb program, databases were generated for the genomic sequences for processing with blast. We filtered out the DNA sequences corresponding to plasmids by applying our in-house scripts GenAllGenomesFileNames.sh and allgenomes_wo-plasmids.pl. Search for dUTPase sequences, the UNG sequence and the UNG inhibitor UGI-SAUGI-P56 sequences were directed by the run-blast.pl script that calls the program tblastn; the applied fasta files to search for in the database were: dUTPase-tri-di1-di2-arch.fasta,UNG.fasta, UGI-SAUGI-P56.fasta., all downloadable from `http://pitgroup.org/static/life_wo_dutpase/`. The dUTPase fasta file contains one trimeric (E. coli dUTPase, UniProt: P06968), two dimeric (C. jejuni and S. aureus phiEta phage dUTPases, UniProt: O15826 and Q9G011, respectively), as well as and one archaeal dUTPase-like sequence (the putative dCTP deaminase from Pyrococcus furiosus, Uniprot accession number Q8X251). The UNG fasta file contains the NCBI Reference Sequence WP_001262716.1 of Enterobacteriaceae uracil-DNA glycosylase. The fasta file for the UNG inhibitor proteins consists of the sequences corresponding to the UniProt accession numbers P14739, Q936H5 and Q38503.

The evaluation of the tblastn results was performed by the script findnohits.pl that returned a table of the bacterial/archaeal genomes without dUTPase genes where no alignments were found with smaller than 0.01 E-value for any of the three dUTPases we search for. The genomes without

dUTPase hits were also partitioned into classes (i) according to the containment of UNG genes with better than 0.01 E-value, and (ii) containment of any UNG inhibitors with sequence-similarities from the fasta file UGI-SAUGI-P56.fasta of 0.01 E-value or less. The genomes without dUTPase and with UNG are listed in Supplementary Table S1. The memberships in the partitions of (i) and (ii) are denoted in the first two columns of Table S1. The genomes without both dUTPase and UNG are listed in Supplementary Table S2. The supplementary material is downloadable from `http://uratim.com/Life_without/LWO_Supplementary.zip`.

The interested reader can easily reproduce the results in each row of Tables S1 and S2 by using the on-line webserver at NCBI at the site: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=tblastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome` by choosing the "Align two or more sequences" option, copying the content of the fasta file tri-di1-di2-arch-UNG-UGI-SAUGI-P56.fasta in the first and copying the NC number of the row of the table into the second input field, and setting "Expect threshold" value to 0.01 at the "Algorithm parameters" menu (see the Supplementary Figure S2 for a screenshot). The hits are colored black while the sequences without hits by gray color.

Generating the taxonomic distribution figure from the results Tables S1 and S2:

We have used the MEGAN5 [138] metagenomic analysis software in a creative way for generating the evolutionary distribution of the genomes with and without dUTPase and UNG. Certainly, we do not have metagenomes here, but we can exploit a particular capability of the MEGAN5 software as follows. MEGAN5 is capable of comparing the taxonomic distribution of

three metagenomes, and it can generate a phylogenetic tree to visualize the distribution. The membership in the three metagenomes can be described by a length-3 0-1 characteristic vector, the ith value is 0 if the taxon is not in the metagenome and 1 if it is in the metagenome, for $i = 1, 2, 3$. Here we substitute these "memberships in metagenomes" with the memberships of sets of genomes with and without dUTPase and UNG as follows: 1,0,0 is substituted if the genome contains dUTPase gene, 0,1,0 is written if the genome does not contain dUTPase but it contain UNG, and 0,0,1 is written if the genome does not contain dUTPase and UNG.

The more technical description of the workflow is as follows.

First, the file that maps the gi values the Taxonomy IDs was downloaded from the NCBI FTP site: `ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/gi_taxid_nucl.dmp.gz`. From this file, using the non-plasmid bacterial/archaeal genome-headers, with a script enclosed as Annot-w-TAXID.pl, NC-numbers were mapped to gi and Taxonomy IDs; the resulting file is NC-GI-TAXID-wo-plasmid.csv.

Next, the gen-megan.pl script of ours was applied to get life_wo_di1-di2-tri-arch_dUTPase_E001.megan file that was opened by the MEGAN5 software (downloadable from `http://ab.inf.uni-tuebingen.de/software/megan5/`. The evolutionary tree figures were created by setting the Rank, and in the Tree menu by setting the Show Number of Read Summarized and Show values on log scale options. The leaves, containing only few genomes can be filtered by setting the Tree/Hide Low Support Nodes option in MEGAN5.

2.6.3 Results and Discussion

Figure 2.10 describes how UNG and dUTPase collaborate to keep DNA uracil-free and also shows the inhibitory protein factors described so far in the

literature for either dUTPase or UNG. To date, only one dUTPase-inhibitory protein has been identified at the molecular level, namely, the repressor protein termed Stl. This protein is encoded within the S. aureus SaPIBov1 pathogenicity island. For UNG, three different proteins have been identified with significant inhibitory effectivity. Two of these (UGI and p56) are encoded by different bacteriophages (phages PBS1/PBS2 and phi29 of Bacillus subtilis ([139,140], respectively). The UGI function encoded in phages is either required to allow synthesis of uracil-enriched DNA (in the case of phages PBS1/PBS2) or protects against the cleavage of phage genome at uracil positions thereby facilitating viral DNA replication [141]. The third protein with UNG inhibitory activity was recently identified in S. aureus (SaUGI) and interestingly, this is the first such case where a UNG inhibitor is encoded in the cellular genome itself [142].

Both dUTPase and UNG are generally presumed to be ubiquitous in free-living organisms. It was, therefore, an unexpected finding that in S. aureus, the dUTPase gene is only found located on phages or prophages inserted into the cellular genome, while in strains cured of prophages and phages, the dUTPase gene is absent from the genome [128]. Such conditions where the dUTPase enzymatic activity is down-regulated or missing are highly deleterious but may be well tolerated if the uracil-DNA glycosylase activity is diminished. In light of the recent studies on dUTPase and UNG inhibitory proteins, we set out to investigate the genotypes of prokaryotes and Archaea and in these organisms, we describe the distribution of genes that act for or against of uracil occurrence in DNA.

In our studies, we investigated those prokaryote and Archaea genomes that are fully sequenced and deposited in the NCBI Genome database. For dUTPases, two protein families have been described to date, the all-beta

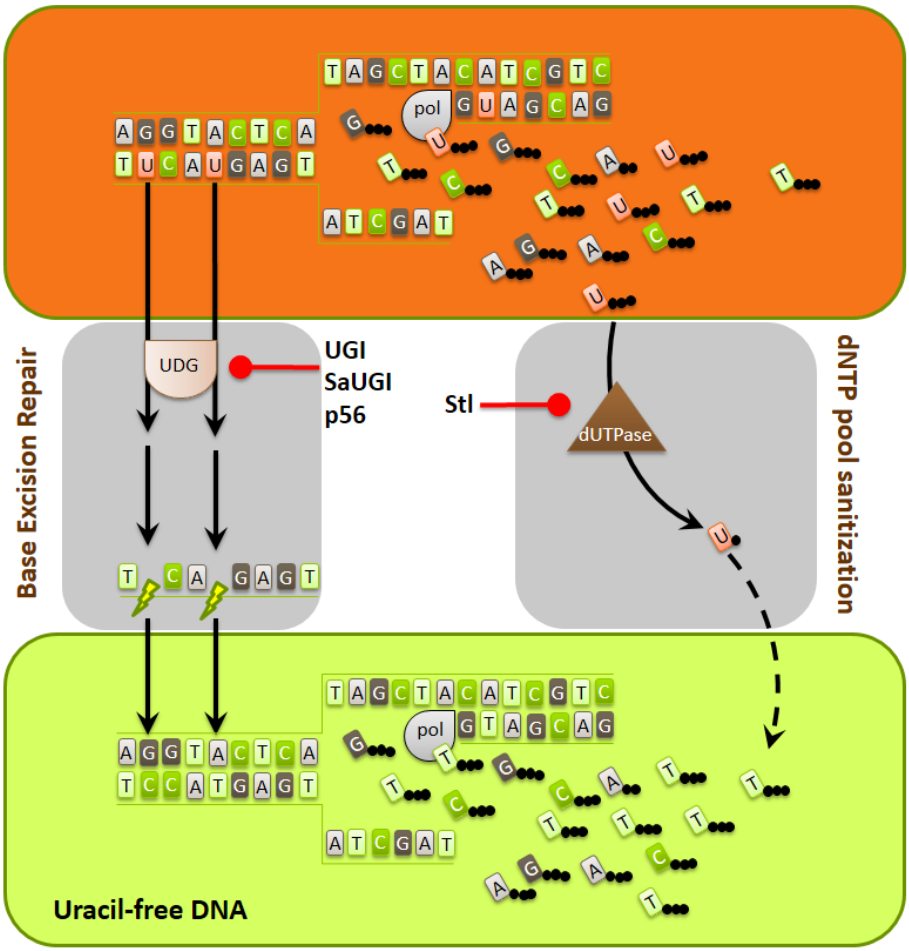

Figure 2.10: Pathways and protein factors involved in the metabolism of uracil-substituted DNA.The scheme illustrates that dUTPase and UDG are responsible for keeping uracil out of DNA by dNTP pool sanitization or uracil-excision, respectively. Inhibitor proteins against UDG (UGI, SaUGI and p56) and dUTPase (Stl) are also included on the figure, showing their point of inhibitory attack.

trimeric and the all-alpha dimeric dUTPases [123], hence we used representative sequences of these families in our search (dUTPases from E. coli and C. jejuni, respectively). Some Staphylococcal phages also encode a variety of

dimeric dUTPase, hence one such sequence was also inserted in the search. In addition, some dCTP deaminases, especially from Archaea, were shown to belong to the trimeric dUTPase fold and acting as bifunctional dCTP deaminase/ dUTPase enzymes. One such sequence was therefore also included (namely dCTP deaminase from P. furiosus). For uracil-DNA glycosylase, the sequence of the UNG enzyme from E. coli was used in our search, as this subfamily of uracil-DNA glycosylases is associated with the major uracil excising efficiency.

The result of screening the bacterial and archaeal genomes for the presence/absence of dUTPase and UNG genes is shown in Figure 2.11. Interestingly, this systematic approach revealed that the lack of dUTPase genes is far more frequent than usually thought. Numerous evolutionary branches showed up where a few or more species do not encode dUTPase protein (note the colored segments in Figure 2.11). In fact, most of the phyla contained some species where the dUTPase genes were not found. These instances are widely occurring on the bacterial evolutionary tree, and also among Euryarchaeota. These cases were further distributed into two groups depending on the simultaneous absence or presence of UNG gene (cf blue and pink segments in Figure 2.11, respectively). These two groups are expected to constitute highly different physiological conditions. Dual lack of both dUTPase and UNG possibly results in a viable phenotype with uracil enrichment in the DNA while the lack of dUTPase and presence of UNG is expected to result in genomic instability, and in many cases, cell death.

A more detailed analysis of the evolutionary distribution of species that do not have dUTPase genes is shown in Figure S1 (cf also Table S1 and S2). Table 2.4 summarizes those evolutionary groups where the occurrence of dut-genotypes is detected in $> 5\%$ of all genomes within the given evolutionary

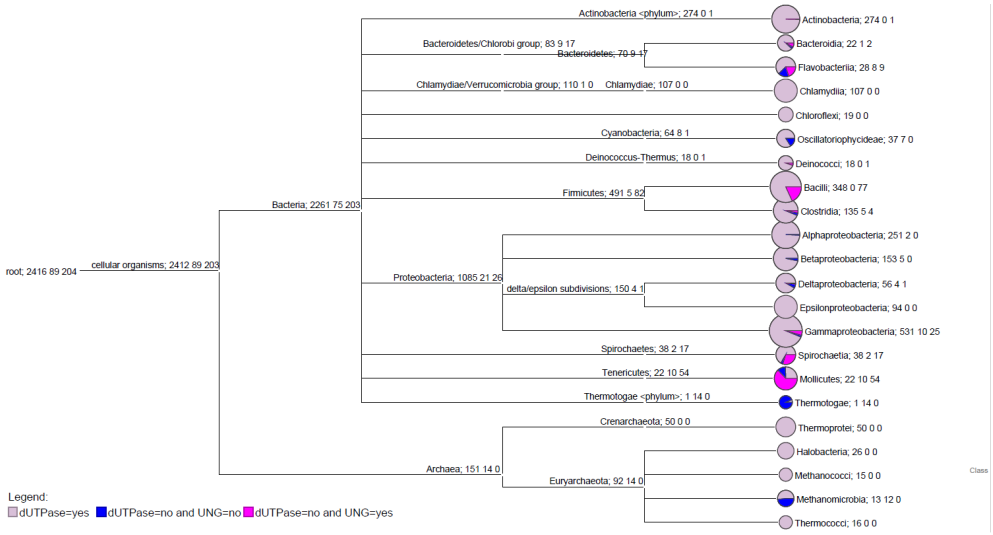

Figure 2.11: The distribution of bacterial/archaeal genomes without dUTPase. Only those classes are shown that have at least 15 genomes examined. Each node of the tree is labeled by three numbers: the first is the number of genomes with dUTPase under the node; the second is the number of genomes without dUTPase and UNG; the third is the number of genomes without dUTPase and with UNG. Since we show only the classes with at least 15 genomes at the right, the not shown classes account for the genomes, missing from the summation.

group and also indicates if the UNG gene is present or absent.

In summary, despite the usual textbook knowledge, we have clearly demonstrated that dUTPase is far from being ubiquitous in prokaryotes and Euryarcheota. It is of immediate further interest to understand how the different organisms may cope with this unexpected situation, especially when UNG is still present.

Inhibitory proteins of UNG may modify the physiological scenario, hence we investigated if any of the UNG inhibitory proteins may be encoded in those bacterial and archaeal genomes that showed up as dut-ung+ in our analysis.

Table 2.4: Distribution of dut genotypes among bacteria and Archaea. Evolutionary branches where the dut-ung+ or dut-ung- genotype occurs in $> 5\%$ of all genomes within the given evolutionary group

dut − ung+	**dut − ung −**
Staphylococcaceae	Oscillatoriophycideae
Flavobacteriaceae	Thermoanaerobacterales
Bacillaceae	Oceanospirillales
Enterococcaceae	Mycoplasmataceae
Vibrionaceae	Thermotogaceae
Spirochaetaceae	Methanomicrobia
Mycoplasmataceae	

We found that none of the phage-related UGI or p56 protein genes could be located on the genomes investigated. The gene for SaUGI, the S. aureus UNG inhibitory protein was located on the S. aureus genome, and a similar sequence was also found on the Butyrivibrio proteoclasticus genome but not elsewhere. Hence, uracil-DNA metabolism basically remains to be governed by the dUTPase and UNG enzymes, with only a very few exceptions, mostly S. aureus strains.

Survival strategies and possible physiological consequences

Since the dut-ung+ genotype is expected to result in genomic instability, it was of interest to investigate if any specific strategy may be employed by the species that are characterized by this unusual feature. First of all, it is important to mention that for S. aureus, numerous phages have been described that encode dUTPase (representatives from either the all-beta trimeric or the all-alpha dimeric dUTPase enzyme families). It has been also described that

in Salmonella enterica, the S. enterica Serovar Typhimurium Myophage Maynard also encodes a bona fide dUTPase gene [143]. Although fully genomic sequence information is limited for other Salmonella phages, this specific instance of phage-encoded dUTPase in the Myophage Maynard indicates the possibility that Salmonella strains also rely on phage-provided dUTPases.

Another strategy to supply some dUTPase-like enzymatic activity was found in Deinococcus radiodurans. This organism, known for its high resistance against ionizing radiation [144], encodes a MazG-like enzyme, with a rather promiscuous substrate specificity [145]. Among numerous dNTPs, the MazG-like D. radiodurans enzyme also cleaves dUTP [145]. Although less efficient and less specific, this supplementation of dUTPase enzymatic activity may ensure viability. In this respect, it is relevant to point out that in several systems, strong inhibition of dUTPase did not lead to lethality indicating that a residual dUTPase activity might be still enough for survival [124, 146]. Under these circumstances, the genomic DNA may contain a somewhat elevated level of incorporated deoxyuridine moieties.

For Thermatoga and Methanomicrobia, data from the literature indicate that the dut-ung- genotype found in our present work may be compensated for by including genes for a less specific MazG-like dNTPase together with an Archaea-like uracil-DNA glycosylase [147]. Lateral gene transfer between Archaea and bacteria has been suggested as the underlying mechanism that led to the appearance of Archaea-like uracil-DNA glycosylase in Thermatoga.

In conclusion, we have shown that the genes for the common dUTPase enzyme families are far from being ubiquitous in prokaryotes and Archaea. This unexpected genotype is observed in evolutionary well-separated branches suggesting that loss of the dut gene(s) might have occurred on multiple independent occasions during evolution.

Supplementary tables and figures

The supplementary material is downloadable from `http://uratim.com/Life_without/LWO_Supplementary.zip`

Figure S1 depicts the taxonomic distribution of bacterial/archaeal genomes without dUTPase on the family level. Only those families are shown that have at least 15 genomes examined. Each node of the tree is labeled by three numbers: the first is the number of genomes with dUTPase under the node; the second is the number of genomes without dUTPase and UNG; the third is the number of genomes without dUTPase and with UNG. Since we show only the families with at least 15 genomes at the right, the not shown classes account for the genomes, missing from the summation. Blue color denotes the proportion of genomes without dUTPase and UNG, while pink genomes without dUTPase and with UNG.

Figure S2 is a screenshot showing the proper settings for the verification of our results with the NCBI tblastn webserver.

Table S1 gives the list of the bacterial/archaeal genomes without dUTPase but with the UNG gene. The second column shows the presence of UNG inhibitors in the genome.

Table S2 gives the list of the bacterial/archaeal genomes without dUTPase and UNG.

Chapter 3

Data Mining in Connectomics

3.1 The Budapest Reference Connectome Server v2.0

3.1.1 Introduction

Several large-scale projects for brain–mapping are being executed [20, 105], but the neuron-scale graph of the human brain, where the nodes are the neurons, and two neurons are connected by an edge if they are joined through a synapse, is out of reach today [22]. The difficulties come from the number of the neurons to be mapped, and also from the lack of the high-throughput methods for mapping their connections. The neuron-scale graphs were constructed only for very simple organisms with a very small number of neurons [15,16,21] or for just small cortical areas of more complex organisms [14, 17].

The application of magnetic resonance imaging (MRI) offers numerous methods for mapping physical and functional connections between subdivided anatomical areas of the brain (called "Regions of Interests", ROIs),

each consisting of millions of neurons. The vertices are the ROIs, and two ROIs are connected by an edge if connections are detected between them by an MRI-based method. This method can either be diffusion MRI imaging, depicting the Brownian motion of water molecules in axons, consequently, mapping the axons between different cortical areas; or functional MRI (fMRI) imaging, depicting brain areas of elevated blood flow while the subject rests or performs different mental tasks.

In this note we present a web-server, which, starting from the diffusion MRI data published as a result of the Human Connectome Project [105], compiles differently parametrized reference graphs from the common edges of the graphs describing 96 different 1015-vertex graphs of 96 human subjects. Additionally, a default, single graph, the Budapest Reference Connectome v2.0 is also presented in two downloadable formats.

The resulting graphs may be used for identifying more robust, more error-free connections between the cortical areas, represented by ROIs: for example, in the default reference graph (i.e., the Budapest Reference Connectome v2.0), if an edge is present then it is present in at least 14 different source graphs. In general, one may set the "Minimum edge confidence" to value k anywhere between $k = 1$ (where an edge is included if it is present in at least one source graph) through $k = 96$ (where an edge is present in the reference graph if it can be found in all the 96 source graphs).

Therefore, the resulting graphs contain *common, consensus* edges (i.e., Fig. 3.1) originated from multiple graphs with user-specified parameters, computed from the diffusion MRI data of different subjects.

Version 2.0 of the Budapest Reference Connectome Server is described here in detail. Choosing Version 1.0 is also possible on the website: Version 1.0 applies the source data from the already classical article of [19], describing

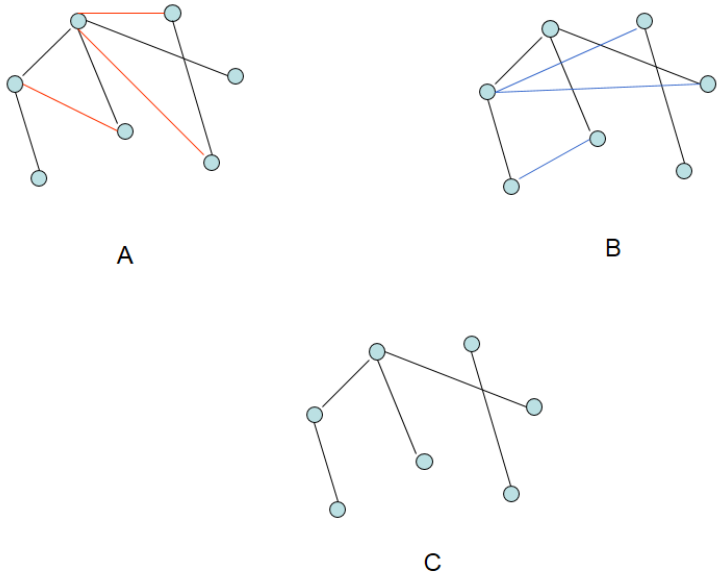

Figure 3.1: The black edges of graphs A and B are common edges; they form graph C, the consensus graph.

six connectomes of five subjects, each with 998 vertices. Version 1.0 of the webserver computes the consensus edges, with several parameter options, from those six graphs only.

By filtering edges with very few occurrences or those with small weights, one may get a connectome with more reliable edges and weights than in the case of any single dataset in the input. Therefore, we may get robust edges and weights in the consensus graphs generated by the server.

3.1.2 Results and Discussion

The Budapest Reference Connectome Server Version 2.0 is available at `http://pitgroup.org/connectome/?version=1`. The newest version, v3.0 (available at `http://pitgroup.org/connectome`) is not detailed in this section. The default, canonical "Budapest Reference Connectome v2.0" can be downloaded by simply hitting the "Download graph" button without changing the default options. This default graph has 1015 vertices, 8507 edges.

The following options can be set after choosing the "Show options" button:

(i) Version 1.0 or Version 2.0. The default choice is 2.0, using the graphs of 96 subjects, computed from the Human Connectome Project [105]. The user may alternatively choose Version 1.0, that applies only six graphs computed and described by the classical article of [19].

(ii) Minimum edge confidence: The graph to be constructed will contain all the edges that are present in at least k graphs, between the very same vertices in each graph. The valid choices for $k = 1, \ldots, 96$. The last choice means that each source graph needs to contain the edge in order to be presented in the resulting consensus graph.

For each edge $\{u, v\}$, the weight of that edge is a fraction n/L, where n is the number of fibers connecting u and v, and L is the average length of the fibers.

(iii) Minimum edge weight: One may set a slider to a value of minimum weight required. The returned graph will contain edges whose mean or median weights are larger than or equal to this value. The mode of computation (mean or median) can be set by the next option.

(iv) Weight calculation mode: There are two choices: Median or Mean. Choice "Median" means that from the list of weights appearing as the weights of the same edge in different graphs, we use the "central" element, that is, first we sort the weights, and next the element is chosen that separates the upper half of the weights from the lower half of the weights. "Mean" means the arithmetic average of the weights. The default choice is the median, since the median is more robust than the mean: typically extreme large or small strengths have less impact to the median than to the mean.

The resulting graph can be downloaded in CSV or GraphML formats, or can readily be visualized on the web page. The downloaded file-names inherit the parameter-settings as follows: e.g., the Budapest Reference Connectome Version 2.0 is given as the file "budapest_connectome_2.0_14_0_median.csv", that is, the csv file contains the graph generated by Version 2.0 of the server, with a minimum confidence of 14 (i.e., each edge of the graph is contained in at least 14 input graphs), and the minimum edge weight is 0 and the weights of the edges of the reference graph are computed as the median of the weights of the corresponding edges of the input graphs.

The format of the CSV file is demonstrated in Table 3.1.

The number of the common edges in at least n graphs ($n = 1, 2, \ldots, 96$) are given in Figure 3.2.

3.1.3 Methods

The main server, denoted as "v2.0", was created as follows:

The dataset is a subset of Human Connectome Project 500 Subjects Release (`http://www.humanconnectome.org/documentation/S500/`), con-

Label	Description
id_node1	the numerical ID of the first vertex of the edge
id_node2	the numerical ID of the second vertex of the edge
name_node1	the anatomical name of node 1
name_node2	the anatomical name of node 2
parent_id_node1	the ID of the parent region of node 1 on the 83-region atlas
parent_id_node2	the ID of the parent region of node 2 on the 83-region atlas
parent_name_node1	the name of the parent region of node 1 on the 83-region atlas
parent_name_node2	the name of the parent region of node 2 on the 83-region atlas
minimum_edge_confidence	the number of the graphs in which the edge is contained
median	the median of the weights of the same edge in different graphs
average	the average of the weights of the same edge in different graphs

Table 3.1: The column labels of the result file in csv format. The 83-region atlas refers to the atlas of the FreeSurfer tool.

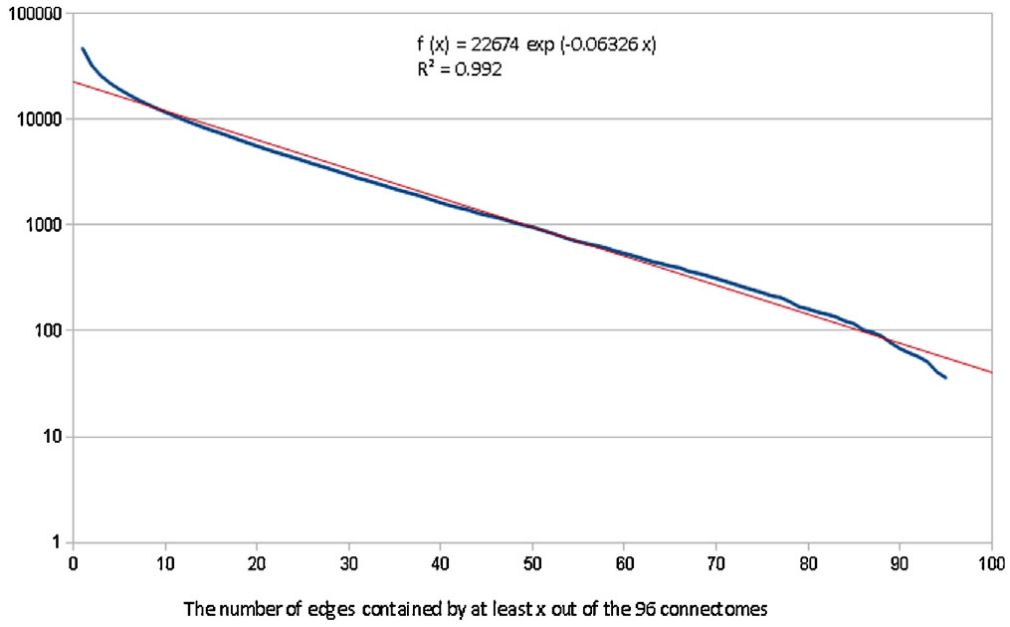

Figure 3.2: The plot of the number of common edges.

taining MRI images of healthy adult males and females between the ages of 22 and 35. The data was downloaded in October, 2014.

Partitioning, tractography, and graph construction were done by the Connectome Mapper Toolkit (`http://cmtk.org`).

Partitioning of the grey matter was done by the Lausanne2008 method [19] into 1015 ROIs of surface area of about 1.5 mm^2.

For tractography, the deterministic streamline method was applied.

The graphs were constructed as follows: Two ROIs were connected by an edge if there exists at least one fiber, determined by the tractography step, that connects these two ROIs. The number and the length of the fibers are taken care of by computing the weights of the edges: For each edge $\{u, v\}$, the weight of that edge is a fraction n/L, where n is the number of fibers connecting u and v, and L is the average length of the fibers.

After 96 graphs were computed, each with 1015 vertices, we identified the common edges, their confidence, and weights, computed according to their median and mean. The large, pre-computed tables were integrated into the webserver.

Version 1.0 of the webserver applies the six graphs that were described in [19]. The definition of weight (called strength) and its computation, and also the parcellation of the cortex used are described in [19]. The six connectomes were downloaded from `http://www.cmtk.org/datasets/homo_sapiens_01.cff` in September, 2014.

The visualization component applies a modified version of the WebGL Brain Viewer [18].

3.1.4 Data availability

The assembled graphs that were used to build the Budapest Reference Connectome Server can be downloaded at the site `http://braingraph.org/download-pit-group-connectomes/`.

Source codes and the workflow to reproduce our results are available at `https://github.com/kerepesi/Brain-Graph-Tools` (see the "BU-

DAPEST REFERENCE CONNECTOME WORKFLOW" section of the README file).

3.2 Comparative Connectomics: Mapping the Inter-Individual Variability of Connections within the Regions of the Human Brain

3.2.1 Introduction

Large co-operative research projects, such as the Human Connectome Project [105], produce high-quality MRI-imaging data of hundreds of healthy individuals. The comparison of the connections of the brains of the subjects is a challenging problem that may open numerous research directions. In the present work we map the variability of the connections within different brain areas in 395 human subjects, in order to discover brain areas with higher variability in their connections or other brain regions with more conservative connections.

The braingraphs or connectomes are the well-structured discretizations of the diffusion MRI imaging data that yield new possibilities for the comparison of the connections between distinct brain areas in different subjects [106, 107] or for finding common connections in distinct cerebra [7], forming a common, consensus human braingraph.

Here, by using the data of the Human Connectome Project [105], we describe, by their distribution functions, the inter-individual diversity of the braingraph connections in separate brain areas in 395 healthy subjects of

ages between 22 and 35 years.

Since every brain is unique, the workflow that produces the braingraphs consists of several steps, including a diffeomorphism [149] of the brain atlas to the brain-image processed. After the diffeomorphism, corresponding areas of different human brains are pairwise identified through the atlas and, consequently, can be compared with one another. The braingraphs, with nodes in the corresponded brain areas, are prepared from the diffusion MRI images of the individual cerebra through a workflow detailed in the "Methods" section. Every braingraph studied contains 1015 nodes (or vertices). The vertices correspond to the subdivision of anatomical gray matter areas in cortical and subcortical regions. For the list of the regions and the number of nodes in each region, we refer to Table 3.4 and Figure 3.6 in the Appendix.

Next, we describe the variability, or the distribution of the graph edges in each brain region, and also in each lobe. Note in this section we use the term lobe in a unique manner as a meaning with "larger area" (not restricted to only cortical areas).

Figure 3.3 contains a simplified example on three small graphs (1,2,3) each with only two regions (A & B). The example clarifies the method, the way the results are presented through a distribution function, and the diagrams describing these functions.

For any fixed brain area, and for any $x : 0 \leq x \leq 1$, let $F(x)$ denote the fraction of the edges[1] in the fixed area[2] that are present in at most the fraction x of all braingraphs, (for a more exact definition of $F(x)$ we refer to the "Methods" section). We note that $F(x)$ is a cumulative distribution function [148] of a random variable described in the "Methods" section.

[1] i.e., the number of the edges in question, divided by the number of all edges in the fixed area;

[2] i.e., with both vertices in the fixed area;

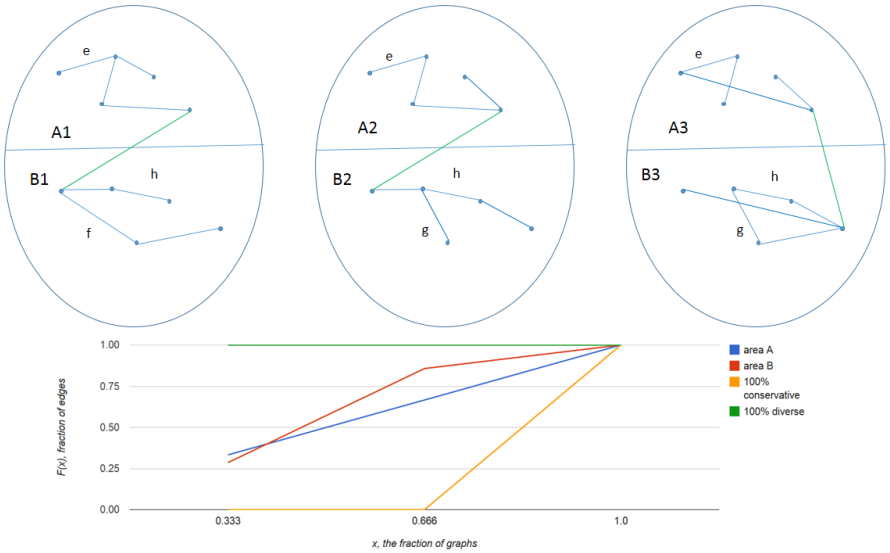

Figure 3.3: A simple example of computing the edge distribution between brain areas. In the example, there are three "braingraphs", each with two areas: A and B. We intend to count the edges that are present in all three graphs, only in two graphs and only in a single graph, respectively (between the same nodes, but in different graphs). For example, the copies of edge e are present in all three A areas, copies of edge h in all three B areas, copies of edge g in two B areas and edge f is present only in $B1$. The edges crossing the boundary of A and B (colored green) are ignored when counting the edge distribution within the areas A and B. In area A, two edges are present once, two edges twice and also two edges (including edge e) exactly three times. In area B, two edges (including f) are present once, four edges (including g) twice and one edge – h – three times. In the diagram on the bottom, we give the $F(x)$ distribution functions for both areas. On axis x, the fractions of the graphs are given, 1/3 correspond to one graph, 2/3 for two and 1.0 for all three graphs. $F(x)$ is defined as the fraction of the edges in the fixed area that are present in at most the fraction x of all braingraphs. Data points corresponding to area A are on the same blue line (1/3, 2/3, 1) and those, corresponding to area B are on the broken, red line (2/7, 6/7, 1). We remark that if all three graphs are the same, then the data points are (0,0,1) (the extremely conservative case, orange line). Similarly, if no two graphs have the same edges, the data points are (1,1,1) (that is the extremely diverse case, green line). This type of diagram is used for the presentation of the results of the distribution of the edges in separate areas of the brain: The faster the line reaches the top $F(x) = 1$ value, the more diverse is the edge set in the corresponding brain area. We also note that in the diagram the lines connect the data points corresponding to the discrete values on axis x, and *do not* describe the step-function $F(x)$ *between* the data points: we have chosen this visualization method because of its clarity even if a higher number of areas are shown (c.f. Figures 3.4 and 3.3 with numerous crossing lines).

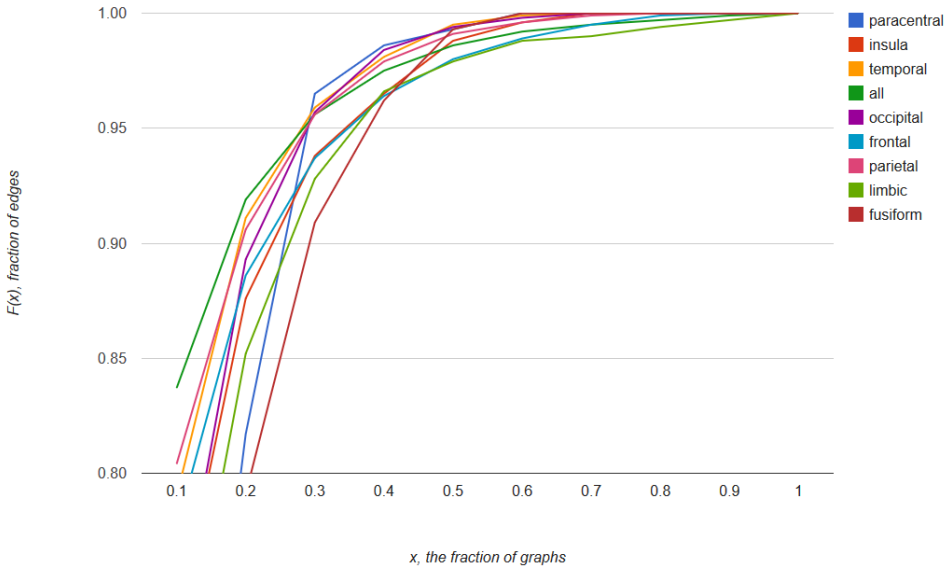

Figure 3.4: The diversity of the edges in different lobes, measured by the distribution function $F(x)$. Only the areas with more than 10 nodes and $F(x)$ values of more than 0.8 are visualized. The lobes, whose lines faster (i.e., with smaller x) reach value 1, have higher diversity. The fusiform gyrus and the paracentral lobule clearly moves from the bottom to the top of the diagram, relative to the other lines: this observation suggests that some of their edges are very conservative, and other areas have high diversity. An interactive version of this figure can be found at `http://uratim.com/diversity/Figure_2.html`

3.2.2 Results and Discussion

Table 3.2 summarizes the edge diversity results for the 395 graphs for the lobes of the brain, described by the distribution functions $F(x)$. The last column contains the data for the whole brain with 1015 nodes and 70,652 edges. The sum of the edges of the lobes in Table 3.2 is 30,326: these edges

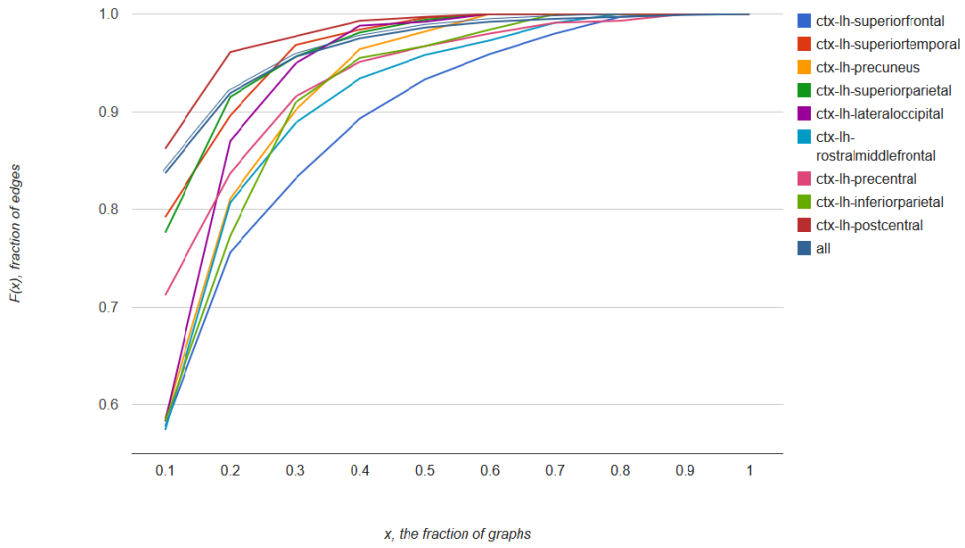

Figure 3.5: The diversity of the edges in different cortical areas of the left hemisphere, measured by the distribution function $F(x)$. The areas, whose lines faster (i.e., with smaller x) reach value 1, have higher diversity. An interactive version of this figure can be found at `http://uratim.com/diversity/Figure_3.html`

have both endpoints in the same lobe. More than forty thousand edges are present and accounted for only in the last column, because these edges connect nodes from different lobes. Therefore, the values in the last column cannot be derived from the other columns, since that column contains the contribution of edges that do not contribute to any other columns.

We want to find out which brain areas are more conservative and which are more diverse than the others. We suggest designating an area as "conservative" if for most x values, its $F(x)$ distribution function is less than the $F(x)$ of the all brain, given in the last column. We also suggest designating an area as "diverse" if for most x values, its $F(x)$ distribution function is greater than the $F(x)$ of all brain, given in the last column.

.	para-central	insula	temporal	thalamus	occipital	frontal	parietal	brain-stem	limbic	fusiform	basal-ganglia	all
nodes	23	33	148	2	95	335	268	1	67	35	8	1015
edges	142	258	3553	3	1900	14260	8873	1	1013	287	36	70652
x	F(x)	F(x)	F(x)	F(x)	F(x)	F(x)	F(x)	F(x)	F(x)	F(x)	F(x)	F(x)
0.1	0.578	0.733	0.790	0	0.730	0.776	0.804	0	0.693	0.537	0.194	0.837
0.2	0.817	0.876	0.911	0	0.893	0.886	0.906	0	0.852	0.791	0.222	0.919
0.3	0.965	0.938	0.959	0	0.957	0.937	0.956	0	0.928	0.909	0.361	0.956
0.4	0.986	0.965	0.981	0	0.984	0.964	0.979	0	0.966	0.962	0.444	0.975
0.5	0.993	0.988	0.995	0	0.994	0.980	0.991	0	0.979	0.993	0.472	0.986
0.6	1	0.996	0.999	0	0.998	0.989	0.996	0	0.988	1	0.500	0.992
0.7	1	1	1	0	1	0.995	0.999	0	0.990	1	0.611	0.995
0.8	1	1	1	0	1	0.999	1	0	0.994	1	0.667	0.997
0.9	1	1	1	0	1	1	1	1	0.997	1	0.694	0.999
1	1	1	1	1	1	1	1	1	1	1	1	1

Table 3.2: The number of nodes, the number of edges and the diversity of the edges in different lobes, measured by the distribution function $F(x)$. The list includes some brain areas that usually are not counted as lobes: like the fusiform gyrus, basal ganglia, and the paracentral lobule. The lobes, whose columns reach the value 1 faster (i.e. have more 1's at the bottom) have higher diversity. For example, the frontal and the limbic lobes are more conservative, while the temporal and the occipital lobes are more diverse. The distribution of the edges in the fusiform gyrus is particularly interesting: more than 10% of the graphs contain 46% of the edges which means this is a conservative brain area in that parameter domain, compared to the other lobes. The fusiform gyrus remains conservative for $x = 0.2$ and even for $x = 0.3$, but more than 50% of the graphs contain only 0.7% of the edges. Therefore, some edges of the fusiform gyrus are well conserved, and some other parts are very diverse. The paracentral lobule has a similar distribution. The data are also visualized in Figure 3.4 and an interactive figure `http://uratim.com/diversity/Figure_2.html`

The most conservative lobes are the smallest ones: the brainstem, the thalamus and the basal ganglia contain only 1, 2 and 8 nodes, resp., and most of the edges in those regions are present in almost all braingraphs. If

we take the average number of the braingraphs containing an edge from those regions, we get 316, 390 and 213 graphs, resp.

It is much more interesting to review the diversity of the connections in larger areas. The frontal and the limbic lobes are conservative for most values of x (i.e., their $F(x)$ values are less than that of the last column), while the temporal and the occipital lobes are diverse for larger x's. The distribution of the edges in the fusiform gyrus is particularly interesting: more than 10% of the graphs contain 46% of the edges which means this is a conservative brain area in that parameter domain, compared to the other lobes. The fusiform gyrus remains conservative for $x = 0.2$ and even for $x = 0.3$, but more than 50% of the graphs contain only 0.7% of the edges. That means that some edges of the fusiform gyrus are well conserved, and some parts are very diverse. The paracentral lobule has a very similar distribution.

Table 3.3 summarizes the diversity results for those cortical areas which have more than 222 edges (see Table 3.5 in the Appendix for the edge numbers).

3.2.3 Methods

We have worked with a subset of the anonymized 500 Subjects Release published by the Human Connectome Project [105]: (`http://www.humanconnectome.org/documentation/S500`) of healthy subjects between 22 and 35 years of age. Data were downloaded in October, 2014.

We have applied the Connectome Mapper Toolkit [151] (`http://cmtk.org`) for brain tissue segmentation, partitioning, tractography and the construction of the graphs. The fibers were identified in the tractography step. The program FreeSurfer was used to partition the images into 1015 cortical and sub-cortical structures (Regions of Interest, abbreviated: ROIs), and was

	ctx-lh-superior-frontal	ctx-rh-superior-frontal	ctx-lh-superior-temporal	ctx-rh-superior-temporal	ctx-lh-pre-cuneus	ctx-rh-pre-cuneus	ctx-lh-superior-parietal	ctx-rh-superior-parietal	ctx-lh-lateral-occipital	ctx-rh-lateral-occipital	ctx-lh-rostral-middle-frontal	ctx-rh-rostral-middle-frontal	ctx-lh-pre-central	ctx-rh-pre-central	ctx-lh-inferior-parietal	ctx-rh-inferior-parietal	ctx-lh-post-central	ctx-rh-post-central	all
Edge #	910	774	250	228	222	227	317	314	254	263	331	352	448	500	242	340	305	273	70652
x																			
0.1	0.578	0.601	0.792	0.763	0.586	0.740	0.776	0.704	0.583	0.510	0.574	0.568	0.712	0.664	0.583	0.556	0.862	0.865	0.837
0.2	0.756	0.760	0.896	0.895	0.811	0.912	0.915	0.860	0.870	0.761	0.807	0.776	0.837	0.840	0.773	0.753	0.961	0.945	0.919
0.3	0.831	0.850	0.968	0.956	0.901	0.956	0.956	0.936	0.949	0.901	0.888	0.878	0.915	0.922	0.909	0.900	0.977	0.967	0.956
0.4	0.893	0.907	0.984	0.974	0.964	0.978	0.981	0.968	0.988	0.958	0.934	0.918	0.951	0.948	0.955	0.962	0.993	0.989	0.975
0.5	0.933	0.944	0.996	1.000	0.982	0.991	0.994	0.994	0.992	0.977	0.958	0.960	0.967	0.968	0.967	0.991	0.997	0.996	0.986
0.6	0.959	0.974	1	1	1	0.991	1	0.997	1	0.996	0.973	0.980	0.980	0.980	0.984	0.997	1	1	0.992
0.7	0.980	0.985	1	1	1	1	1	1	1	1	0.991	0.994	0.991	0.99	1	1	1	1	0.995
0.8	0.997	0.995	1	1	1	1	1	1	1	1	1	1	0.993	0.996	1	1	1	1	0.997
0.9	1	1	1	1	1	1	1	1	1	1	1	1	1	0.998	1	1	1	1	0.999
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
expectation	56.4	53.2	25.2	28.5	45.2	30.0	27.9	33.9	40.2	52.1	50.5	50.5	39.2	40.7	48.5	50.2	19.2	18.9	22.5

Table 3.3: The diversity of the edges in different cortical areas, measured by the distribution function $F(x)$. The abbreviation "ctx-lh" stands for "cortex left-hemisphere", "ctx-rh" for "cortex right-hemisphere". The areas, whose columns reach the value 1 faster (i.e., have more 1's at the bottom) have higher diversity. As in Table 3.2, the frontal regions are relatively more conservative, while the parietal regions are more diverse. Both precentral gyri are also conservative, and the postcentral and the superiortemporal gyri are more diverse. The last row contains the expected number of the graphs which contain a randomly chosen edge from the brain area indicated. Large expected number implies a conservative area, a small value implies a more diverse area. The data for the left hemisphere are also visualized in Figure and on an interactive figure `http://uratim.com/diversity/Figure_3.html`

based on the Desikan-Killiany anatomical atlas [151](see Figure 4 in [151]). Tractography was performed by the Connectome Mapper Toolkit [151], using the MRtrix processing tool [150] and choosing the deterministic streamline method with randomized seeding.

The graphs were constructed as follows: the 1015 nodes correspond to the 1015 ROIs, and two nodes were connected by an edge if there exists at least one fiber connecting the ROIs corresponding to the nodes.

The distribution function

The variability of the edges in regions or lobes are described by cumulative distribution functions (CDF) (also called just the "distribution function") of the edges [148]. The general definition of the CDF is as follows:

Definition 1 *Let Y be a real-valued random variable. Then*

$$F(x) = P(Y \leq x)$$

defines the cumulative distribution function of Y for real x values.

For example, if a is the maximum value of Y then $F(a) = 1$, and if b is less than the minimum value of Y, then $F(b) = 0$.

CDFs are used the following way: Suppose that our cohort consists of n persons' braingraphs (in the present work $n = 395$). For a given, fixed brain area, our random variable Y takes on values $Y = u/n, u = 0, 1, \ldots, n$. The equation $Y = u/n$ corresponds to the event that a uniformly, randomly chosen edge is in exactly u graphs from the n possible one, and the probability $P(Y = u/n)$ gives the probability of this event. Or, in other words, the equation $Y = u/n$ corresponds to the set of edges — with both nodes in the fixed brain area — which are present in exactly u braingraphs, and the probability $P(Y = u/n)$ gives the fraction of the edges that are present in exactly u braingraphs. Therefore, $F(x) = P(Y \leq x)$ gives the fraction (i.e., the probability) of the edges that are present in at most of a fraction x of all the graphs.

The number of nodes and edges in each brain regions are given in supporting Tables S1 and S2 in the Appendix. We remark that we counted the edges without multiplicities: that is, if an edge e was either present in, say, 42 copies or just 1 copy of the braingraph, in both cases we counted it only once.

The distributions were computed by counting the number of appearances of each edge in all the 395 braingraphs. Then the distribution of these numbers was evaluated in lobes and smaller cortical areas.

3.2.4 Conclusions:

By our knowledge for the first time, we have mapped the inter-individual variability of the braingraph edges in different cortical areas. We have found more and less conservative areas of the brain: for example, frontal lobes are conservative, superiortemporal and the post-central gyri are very diverse. The fusiform gyrus and the paracentral lobule have shown both conservative and diverse distributions, depending on the range of the parameters.

Data availability:

The unprocessed and pre-processed MRI data are available at the Human Connectome Project's website:

`http://www.humanconnectome.org/documentation/S500` [105].

The assembled graphs that were analyzed in the present work can be accessed and downloaded at the site `http://braingraph.org/download-pit-group-connectomes/`.

Source codes and the workflow to reproduce our results are available at `https://github.com/kerepesi/Brain-Graph-Tools` (see the "BRAIN DIVERSITY WORKFLOW" section of the README file).

3.2.5 Appendix

Abbreviations: ctx-rh: cortex right-hemisphere ctx-lh: cortex left-hemisphere

Area name	No. Of nodes
ctx-lh-superiorfrontal	45
ctx-rh-superiorfrontal	42
ctx-rh-precentral	36
ctx-lh-precentral	35
ctx-lh-postcentral	31
ctx-rh-postcentral	30
ctx-lh-superiorparietal	29
ctx-rh-superiorparietal	29
ctx-rh-rostralmiddlefrontal	27
ctx-lh-superiortemporal	26
ctx-lh-rostralmiddlefrontal	26
ctx-rh-inferiorparietal	26
ctx-rh-superiortemporal	25
ctx-rh-lateraloccipital	23
ctx-rh-precuneus	23
ctx-lh-lateraloccipital	23
ctx-lh-precuneus	22
ctx-lh-inferiorparietal	22
ctx-lh-supramarginal	21
ctx-rh-supramarginal	20
ctx-rh-middletemporal	19
ctx-lh-fusiform	18
ctx-rh-lateralorbitofrontal	17
ctx-rh-fusiform	17
ctx-rh-lingual	17
ctx-lh-insula	17

ctx-lh-lingual	17
ctx-lh-inferiortemporal	16
ctx-rh-insula	16
ctx-rh-inferiortemporal	16
ctx-lh-middletemporal	16
ctx-lh-lateralorbitofrontal	16
ctx-lh-caudalmiddlefrontal	13
ctx-rh-paracentral	12
ctx-lh-paracentral	11
ctx-rh-caudalmiddlefrontal	11
ctx-rh-medialorbitofrontal	11
ctx-lh-medialorbitofrontal	10
ctx-lh-parsopercularis	10
ctx-lh-posteriorcingulate	9
ctx-rh-posteriorcingulate	9
ctx-rh-parsopercularis	9
ctx-rh-parstriangularis	8
ctx-rh-cuneus	8
ctx-rh-pericalcarine	8
ctx-lh-cuneus	7
ctx-lh-pericalcarine	7
ctx-lh-isthmuscingulate	7
ctx-lh-parstriangularis	7
ctx-rh-parahippocampal	6
ctx-lh-bankssts	6
ctx-rh-caudalanteriorcingulate	6
ctx-rh-isthmuscingulate	6

ctx-lh-parahippocampal	6
ctx-rh-bankssts	6
ctx-lh-rostralanteriorcingulate	5
ctx-lh-caudalanteriorcingulate	5
ctx-rh-parsorbitalis	4
ctx-lh-transversetemporal	4
ctx-lh-parsorbitalis	4
ctx-rh-rostralanteriorcingulate	4
ctx-lh-entorhinal	3
ctx-lh-temporalpole	3
ctx-rh-temporalpole	3
ctx-rh-transversetemporal	3
ctx-lh-frontalpole	2
ctx-rh-entorhinal	2
ctx-rh-frontalpole	2
Left-Thalamus-Proper	1
Left-Amygdala	1
Right-Hippocampus	1
Right-Amygdala	1
Right-Putamen	1
Right-Accumbens-area	1
Left-Hippocampus	1
Left-Pallidum	1
Right-Pallidum	1
Right-Thalamus-Proper	1
Left-Putamen	1
Right-Caudate	1

Left-Caudate	1
Left-Accumbens-area	1
Brain-Stem	1
Sum of nodes	1015

Table 3.4: The number of nodes in each ROI.

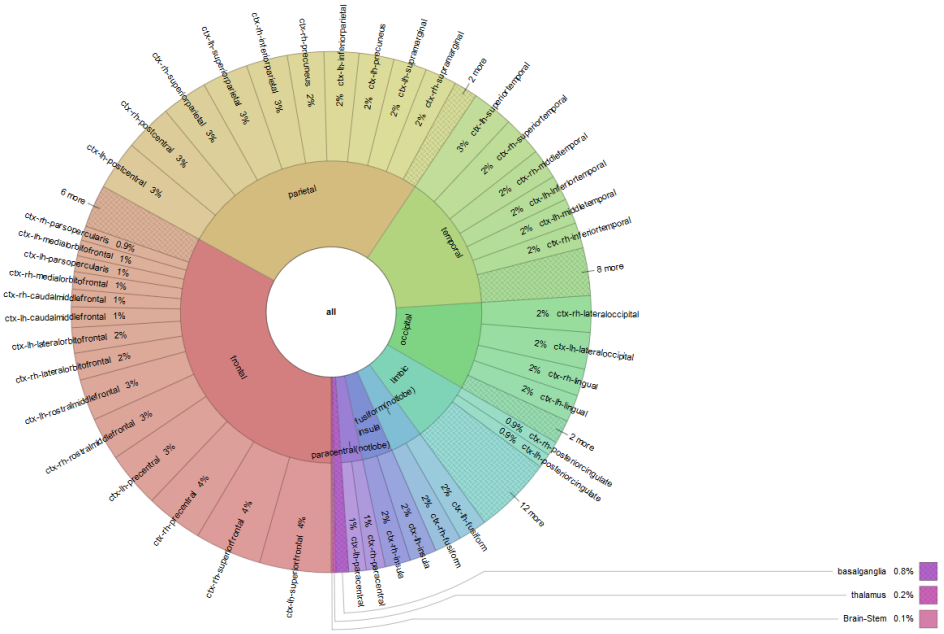

Figure 3.6: The number of nodes in ROIs and lobes. The interactive figure can be viewed at `http://uratim.com/diversity/Figure_S1-Krona.html`

'all'	70652
'ctx-lh-superiorfrontal'	910
'ctx-rh-superiorfrontal'	774
'ctx-rh-precentral'	500
'ctx-lh-precentral'	448

'ctx-rh-rostralmiddlefrontal'	352
'ctx-rh-inferiorparietal'	340
'ctx-lh-rostralmiddlefrontal'	331
'ctx-lh-superiorparietal'	317
'ctx-rh-superiorparietal'	314
'ctx-lh-postcentral'	305
'ctx-rh-postcentral'	273
'ctx-rh-lateraloccipital'	263
'ctx-lh-lateraloccipital'	254
'ctx-lh-superiortemporal'	250
'ctx-lh-inferiorparietal'	242
'ctx-rh-superiortemporal'	228
'ctx-rh-precuneus'	227
'ctx-lh-precuneus'	222
'ctx-lh-supramarginal'	209
'ctx-rh-supramarginal'	206
'ctx-rh-middletemporal'	176
'ctx-lh-fusiform'	157
'ctx-rh-lateralorbitofrontal'	144
'ctx-lh-inferiortemporal'	135
'ctx-rh-insula'	131
'ctx-lh-lingual'	131
'ctx-rh-fusiform'	130
'ctx-rh-inferiortemporal'	130
'ctx-lh-lateralorbitofrontal'	127
'ctx-lh-insula'	125
'ctx-lh-middletemporal'	119

'ctx-rh-lingual'	114
'ctx-lh-caudalmiddlefrontal'	91
'ctx-rh-paracentral'	76
'ctx-rh-caudalmiddlefrontal'	65
'ctx-lh-paracentral'	64
'ctx-rh-medialorbitofrontal'	59
'ctx-lh-parsopercularis'	55
'ctx-lh-medialorbitofrontal'	54
'ctx-lh-posteriorcingulate'	45
'ctx-rh-parsopercularis'	45
'ctx-rh-posteriorcingulate'	43
'ctx-rh-parstriangularis'	36
'ctx-rh-cuneus'	35
'ctx-rh-pericalcarine'	35
'ctx-lh-cuneus'	28
'ctx-lh-pericalcarine'	28
'ctx-lh-isthmuscingulate'	28
'ctx-lh-parstriangularis'	28
'ctx-lh-bankssts'	21
'ctx-rh-caudalanteriorcingulate'	21
'ctx-lh-parahippocampal'	21
'ctx-rh-parahippocampal'	20
'ctx-rh-isthmuscingulate'	20
'ctx-rh-bankssts'	20
'ctx-lh-rostralanteriorcingulate'	15
'ctx-lh-caudalanteriorcingulate'	15
'ctx-rh-parsorbitalis'	10

'ctx-lh-parsorbitalis'	10
'ctx-rh-rostralanteriorcingulate'	10
'ctx-lh-transversetemporal'	8
'ctx-lh-entorhinal'	6
'ctx-rh-transversetemporal'	5
'ctx-lh-temporalpole'	4
'ctx-rh-entorhinal'	3
'ctx-rh-temporalpole'	3
'Left-Thalamus-Proper'	1
'Left-Amygdala'	1
'ctx-lh-frontalpole'	1
'Right-Hippocampus'	1
'Right-Amygdala'	1
'ctx-rh-frontalpole'	1
'Right-Putamen'	1
'Right-Accumbens-area'	1
'Left-Hippocampus'	1
'Left-Pallidum'	1
'Right-Pallidum'	1
'Right-Thalamus-Proper'	1
'Left-Putamen'	1
'Right-Caudate'	1
'Left-Caudate'	1
'Left-Accumbens-area'	1
'Brainstem'	1

Table 3.5: The number of edges in each ROI.

3.3 How to Direct the Edges of the Connectomes: Dynamics of the Consensus Connectomes and the Development of the Connections in the Human Brain

3.3.1 Introduction

The Human Connectome Project [105] has produced high-quality MRI-imaging data of hundreds of healthy subjects. The enormous quantity of data is almost impossible to use in brain research without introducing some rich structure that helps us to get rid of the unimportant details and allow us to focus on the essential data in the set. We believe that the braingraph or the connectome is such a structure to apply.

The braingraphs or connectomes are discretizations of the diffusion MRI imaging data. Being a graph, it has a set of vertices and some pairs of these vertices are the edges of the graph. Each vertex corresponds to a small (1-1.5 cm^2) areas (called Regions of Interest, ROIs) of the gray matter, and two vertices are connected by an edge, if a diffusion-MRI based workflow finds fibers of axons, running between those ROIs in the white matter of the brain. In other words, the braingraph concentrates on the connections between areas of gray matter (this is an essential part of the data) and forgets about the exact spatial orbits of the axon-fibers, running between these gray matter areas in the white matter of the brain (these are the unimportant part of the data). The braingraphs may record the length or the width of these fibers as edge-weights but definitely does not contain any spatial description of their orbit in the white matter.

An important question is the determination of the direction of the graph – or connectome – edges in these braingraphs. By our knowledge, the present diffusion-MRI based workflows have no data showing the direction of the neuronal fiber tracts between the ROIs.

Hundreds of publications deal with the properties of the human connectome every year (e.g., [106–109], but very few analyze the common edges and the edge-distributions between distinct subjects and distinct brain areas [7,8]. In [8] we have mapped the inter-individual variability of the braingraphs in different brain regions, and we have found that the measure of the variability significantly differs between the regions: there are more and less conservative areas of the brain.

3.3.2 Results

In the construction of the Budapest Reference Connectome Server `http://connectome.pitgroup.org` [7], [10], not those edges were mapped that differ [8], but, on the contrary, those that are the same in at least k subject's braingraphs, for $k = 1, 2, \ldots, 418$. These parametrized consensus-graphs describe the common connectomes of healthy humans, parametrized with k.

For $k = 418$ we get only those edges that are present in all the 418 braingraphs. For $k = 1$ we get those edges that are present in at least one braingraph from these 418. Therefore, if we change the value of k, one-by-one, from $k = 418$ through $k = 1$, we will have more and more edges in the graph (Figure 3.7).

We have observed that the order of the appearance of the new edges when we were decreasing the value of k from 418 through 1, is not random at all. More precisely, it resembles a growing tree: the newly appearing edges are usually connected to the already existing edges. This phenomenon is

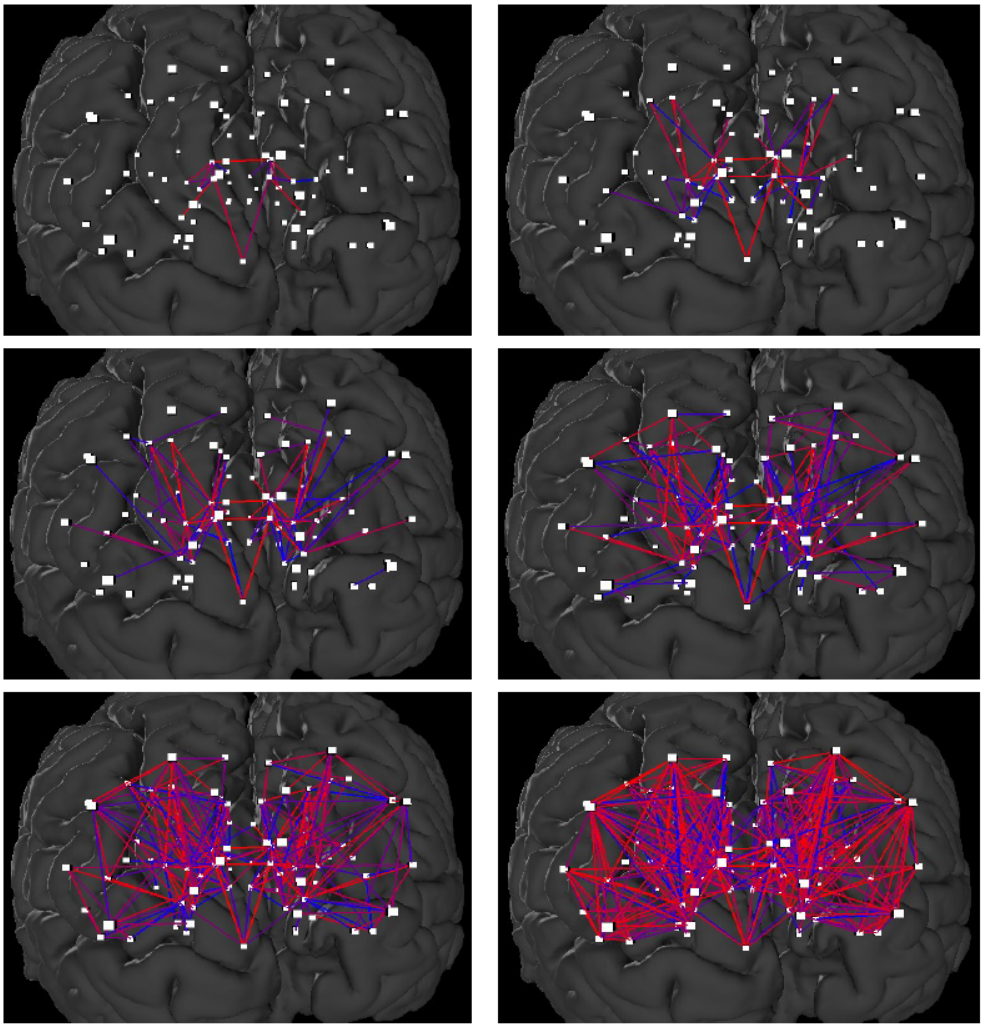

Figure 3.7: Snapshots on the tree-like structure of the Budapest Reference Connectome Server v2.0. The edges of the smallest graph can be identified easily with using the webserver. For example, the edges that are present in all braingraphs include edges between Right-Caudate and Right-Pallidium, Left-Thalamus-Proper and Brain-Stem, Right-Thalamus Proper and Right-Putamen.

observable in the animation at `https://youtu.be/EnWwIf_HNjw` (we remark that graph-theoretically, the growing structure is not a tree as a graph). The

same observation was done in Version 2 (with 96 braingraphs) and Version 3 (with 418, 476 and 477 braingraphs, depending on the fiber-numbers selected) of the server.

In what follows, we clarify the implications of this observation to the

(i) description of the individual development of the connections in the human brain, and

(ii) the determination of the direction of the edges in the human connectome.

The observation is verified by Figure 3.8, made for the Version 3.0 of the server, with 418 braingraphs. For steps $\ell = 0$ through $\ell = 417$, for $k = 418 - \ell$, we have visualized the number of those new edges (that were present in k connectomes, but were not present in $k+1$ connectomes), which connect two vertices, which were not adjacent to any edges before (i.e., they were isolated vertices). We have compared

- a random model, where exactly that many new edges were added randomly in uniform distribution, as in the graph generated by the Budapest Reference Connectome Server,
- and the graph of edges drawn by the Budapest Reference Connectome Server.

3.3.3 Discussion

In the random model, in each step, the same number of edges were added to the graph randomly (independently, in a uniform distribution), as in the Budapest reference Connectome Server.

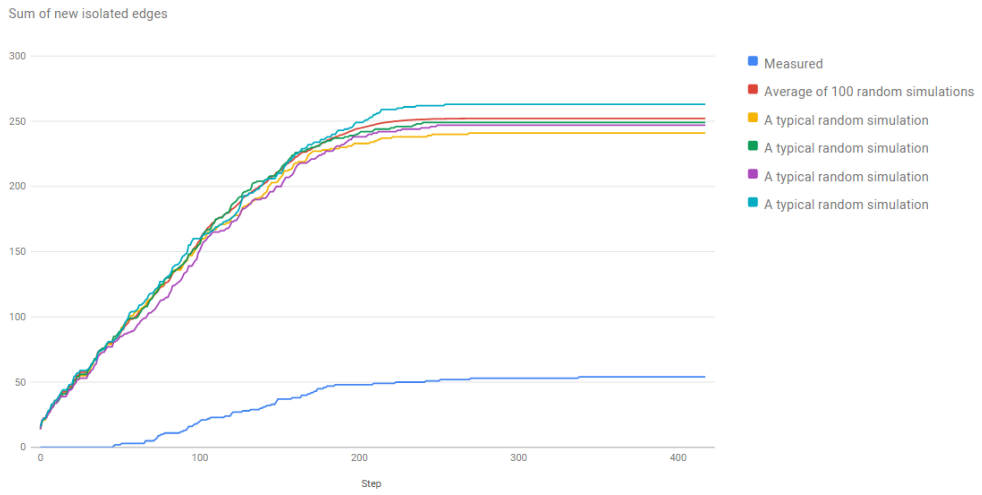

Figure 3.8: The comparison of the random simulation and the real buildup of the edges in the Budapest Reference Connectome server v3.0.

The difference is very clear in Figure 3.8: in the random model, dramatically more new edges appear that are not connected to the old ones.

Another visualization of this surprising phenomenon is the component tree of the evolving graph, made for Version 2 with 96 braingraphs. As k decreases from 96 to 1, zero or more new edges are added to the existing graph in each step. In the step corresponding to k, those edges appear that are present in exactly k graphs. This may result in the forming of new connected components, and/or the merging of some older components of the graph. The phenomenon can be visualized on a graph-theoretical tree, where each level of the tree corresponds to some value of k. On each level, some leaf nodes may appear (for each new component), and the existing nodes may merge into a parent node. We can also assign colors to the nodes according to the following scheme: the leaves get a new color, and a parent node gets the color of its child node corresponding to the largest merged component. The component-tree of the graph is visualized on a very large, labeled interactive figure at

the site http://pitgroup.org/static/graphmlviewer/index.html?src=connectome_dynamics_component_tree.graphml.

We hypothesize that those edges that are contained in many of the graphs were developed in an earlier stage of the brain development than those that are present in fewer subjects. As a possible explanation, we think that those neurons that connect to the developing braingraph at https://youtu.be/EnWwIf_HNjw will not receive apoptosis signals [110–112] and will survive, while other neurons, which are not connected to the older graph, will be eliminated by receiving apoptosis signals in the individual brain development.

In other words, we assume that the connections that are present in almost all braingraphs (c.f., the upper left panel of Fig.3.7) were developed first. Next, new connections were developed, but those neurons whose connections were disconnected from these oldest neurons were eliminated. Next, new neuronal connections were developed, but only those neurons survived that were connected to the building network. Since the deviation between the new edges among the subjects was increased step-by-step, the newer the connections, the fewer the subjects have those edges.

This assumption explains our findings, and it is in line with the "competition hypothesis" of the brain development [112].

How to direct the edges of the human connectome?

For any neuron, there exists a well-defined direction of the signal propagation from the soma through its axon. Diffusion MRI-based methods can be used to identify the spatial location of the fiber tracts, consisted of axons, but their directions, by our present knowledge, cannot be discovered from the MRI data.

If the order of development of the edges in the connectome is known then

we can easily assign a direction to those edges that connects a vertex to another one, such that the first vertex was not connected to any other vertex before, but the second vertex was already connected to the network, when we consider the transition of the edges that were present in at least $k+1$ graphs through the edges that were present in at least k graphs.

More exactly, the observation described above implies a straightforward method for directing some (but not all) the edges of the connectome. Consider the undirected edge u, v, and our goal is to assign a direction to this edge. Let G_{k+1} denote the consensus connectome where each edge is present in at least $k+1$ graphs, and let G_k denote the consensus connectome where each edge is present in at least k graphs. Both G_{k+1} and G_k have the same set of vertices, all the edges of G_{k+1} are also the edges of G_k, but G_k typically has more edges than G_{k+1}. Assume that vertex v was not connected to any other vertices in G_{k+1}, and becomes connected to a vertex u in G_k, where u was connected to other vertices in G_{k+1}. Then we direct this (v, u) edge from v to u, and denote it as an ordered pair (v, u) (Figure 3.9). Obviously, if our hypothesis is correct, then the undirected edge u, v remained in the consensus connectome since vertex v did not get an apoptosis signal, since u was already been connected to the growing network.

We remark that those new edges that connect two, previously isolated points ("isolated edges"), or those that connect two vertices, where both of them were connected to the network before, cannot be directed this way.

3.3.4 Methods

The description of the program and the methods applied in the construction of the Budapest Reference Connectome Server `http://connectome.pitgroup.org` is given in [7].

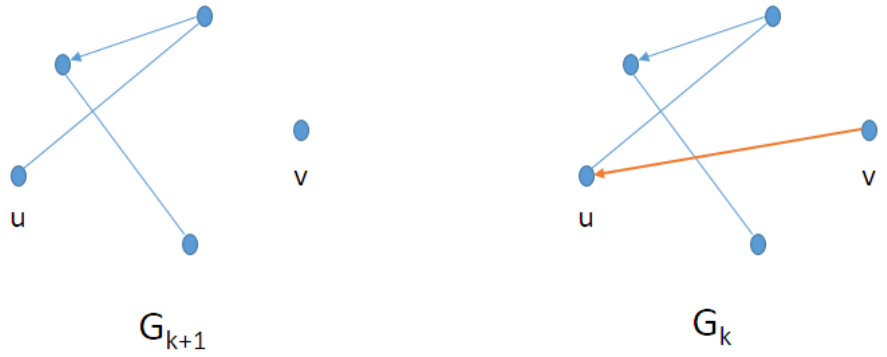

Figure 3.9: Let G_{k+1} denote the consensus connectome where each edge is present in at least $k+1$ graphs, and let G_k denote the consensus connectome where each edge is present in at least k graphs. Both G_{k+1} and G_k have the same set of vertices, all the edges of G_{k+1} are also the edges of G_k, but G_k typically has more edges than G_{k+1}. The (v, u) edge is directed from v to u, if v is not connected to any other vertices in G_{k+1}, and becomes connected to a vertex u in G_k, where u was connected to other vertices in G_{k+1}. Then we direct this (v, u) edge from v to u.

The animation at `https://youtu.be/EnWwIf_HNjw` were prepared by our own Python program from the tables generated by the Budapest Reference Connectome Server [7] with the following settings: Version 2 (i.e., 96 subjects), Population: All (i.e., both male and female subjects), Minimum edge confidence running from 100 % through 26%, Minimum edge weight is 0, Weight calculation model: Median. It contains the common edges found in k subject's braingraphs, from $k = 96$ through $k = 25$. The number of vertices is 1015.

3.3.5 Conclusions:

We have observed that the buildup of the consensus graphs in the Budapest Reference Connectome Server is far from random when the k parameter is changed from $k = 418$ through 1. This observation suggests an underlying structure in the consensus braingraphs: the edges, which are present in more subjects are most probably older in the individual brain development than the edges, which are present fewer individuals. This assumption is in line with the "competition hypothesis" of the brain development [112]. We believe that this observation is applicable to discover the finer structure of the development of the connections in the human brain.

Based on this hypothesis we were able to assign directions to some of the otherwise undirected edges of the connectome, built through a diffusion MRI based workflow.

Data availability:

The unprocessed and pre-processed MRI data that served as a source of our work are available at the Human Connectome Project's website:

`http://www.humanconnectome.org/documentation/S500` [105].

The assembled graphs that were used to build the Budapest Reference Connectome Server can be downloaded at the site `http://braingraph.org/download-pit-group-connectomes/`.

Source codes and the workflow to reproduce our results are available at `https://github.com/kerepesi/Brain-Graph-Tools` (see the "BRAIN EVOLUTION WORKFLOW" section of the README file).

Chapter 4

Outlook and Future Perspectives

AmphoraNet was used in 6 published studies (interestingly all of them are genomics rather than metagenomics) since December 2013 and run more than 2300 jobs. We hope it will be used with success in more studies. For the reason that its popularity and the fact that the number of Bacteria and Archaea complete genomes growth rapidly it would be worth updating the marker gene set in the near future.

It would be very interested to search giant viruses in various arid or not arid metagenomes for example in Hungarian soda pans or in the human body and then characterize where infect giant viruses and what are they doing. Giant Virus Finder is a suitable tools for this purpose.

Our dUTPase findings raise questions about how can live organisms without this important enzymes and can lead to discover new pathways which would be helpful to better understand DNA repair mechanisms in humans.

Brain graphs created by us are downloadable (`http://braingraph.org`) freely so it is open for other researcher for exploring. For example can

be valuable to analyze the correlation between behavioral data (containing results of psichological tests) and brain graphs. Other promising research direction is comparing our brain graphs to brain graphs created from MRI data of people suffered from various diseases (for example using the MRI data of the Alzheimer's Disease Neuroimaging Initiative database - `http://adni.loni.usc.edu/`). Our open source brain graph analysis tools (`https://github.com/kerepesi/Brain-Graph-Tools`) can be helpful for the future analysis and developing.

Chapter 5

One Page Summaries

5.1 Summary in English

We have developed AmphoraNet, an easy-to-use webserver that is capable of assigning a probability-weighted taxonomic group for each phylogenetic marker gene found in the input metagenomic sample; the webserver is based on the AMPHORA2 workflow. Then we have developed a visual analysis tool that is capable of demonstrating the quantitative relations gained from the output of the AMPHORA2 program or the AmphoraNet webserver and then we have evaluated three metagenomic analysis software for their capabilities of assigning *quantitative* phylogenetic information for the data.

On the area of Giant Viruses we have developed a software, called "Giant Virus Finder" that is capable to discover the very likely presence of the genomes of giant viruses in metagenomic datasets. The software is applied to numerous hot and cold desert soil samples as well as some tundra- and forest soils and the soil samples of the Kutch desert.

During investigating the genotype of deposited fully sequenced bacterial and archaeal genomes we have surprisingly found that a wide number of bacterial and archaeal species lack the dUTPase gene.

We have developed the Budapest Reference Connectome Server which generates the common edges of the connectomes of distinct cortexes, each with 1015 vertices, computed from MRI data sets of the Human Connectome Project. After the server had been published, we recognized a surprising property of the server. Decreasing the minimum edge confidence from the maximal value, more and more edges appear in the consensus graph. The observation is that the appearance of the new edges similar to a growing tree. We have also discovered the inter-individual variability of the graphs within different brain regions and we have found that the edges in the temporal and occipital lobes are the most diverse.

5.2 Summary in Hungarian

Létrehoztuk az AmphoraNet-et, egy könnyen használható webszervert, amely minden egyes a metagenomban talált filogenetikai marker gén szekvenciához kijelöl egy rendszertani csoportot. A webszerver az AMPHORA2 munkafolyamaton alapul. Az AmphoraNet után kifejlesztettük az AmphoraVizu webszervert, amely az AmphoraNet nehezen feldolgozható szöveges outputjához nyújt interaktív képi megjelenítést. Ezek után kiértékeltük az általunk fejlesztett AmphoraNet+AmphoraVizu-t és két másik metagenomikai elemző szoftvert abból a szempontból, hogy mennyire írják le jól adott baktériumok előfordulási gyakoriságát ugyanazon mintában.

Ezután kifejlesztettük a Giant Virus Finder szoftvert, amely képes kimutatni óriás vírus specifikus szekvenciák jelenlétét metagenomokban. Az új szoftver segítségével óriás vírusok jelenlétét mutattuk ki számos forró és hideg sivatagi talajmintában.

Megvizsgáltuk az összes baktérium (2261 db) és archaea (151 db) teljes genomi szekvenciára, hogy tartalmaznak-e dUTPáz gént. Meglepő módon azt találtuk, hogy nagy számú baktérium és archaea fajban hiányzik a dUTPáz gén.

Kifejlesztettük a Budapest Reference Connectome szervert, amely MRI felvételekből számolt agygráfokhoz számolja ki a referencia agygráfot. A szervert vizsgálva felfedeztünk egy meglepő tulajdonságot. Amikor a szerveren a maximum értéktől indulva csökkentjük a "Minimum edge confidence" értéket egyre több él jelenik meg az referencia agygráfban. A megdöbbentő észrevétel az, hogy az élek nem véletlenszerűen tűnnek fel, hanem egy kis összefüggő konzervatív gráfból kiindulva egymás után épülve belülről kifelé. Ezen kívül számításokat is végeztünk 395 egyén agygráfjára, felmérve az agyi régiók egyének közötti különbözőségét.

Chapter 6

My Publications Presented in the Thesis

1. Csaba Kerepesi, Dániel Bánky, and Vince Grolmusz. AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene*, 533(2):538–540, 2014.

2. Csaba Kerepesi, Balázs Szalkai, and Vince Grolmusz. Visual Analysis of the Quantitative Composition of Metagenomic Communities: the AmphoraVizu Webserver. *Microbial Ecology*, 69(3):695–697, 2015.

3. Csaba Kerepesi and Vince Grolmusz. Evaluating the Quantitative Capabilities of Metagenomic Analysis Software. *Current Microbiology*, 72(5):612–616, 2016.

4. Csaba Kerepesi and Vince Grolmusz. The "Giant Virus Finder" Discovers an Abundance of Giant Viruses in the Antarctic Dry Valleys. *Archives of Virology*, 162(6):1671–1676, 2017.

5. Csaba Kerepesi and Vince Grolmusz. Giant viruses of the Kutch

Desert. *Archives of Virology*, Volume 161, Issue 3, pp 721–724, 2016.

6. Csaba Kerepesi, Judit E Szabó, Veronika Papp-Kádár, Orsolya Dobay, Dóra Szabó, Vince Grolmusz, and Beata G Vertessy. Life without dUTPase. *Frontiers in Microbiology*, 7:1768, 2016.

7. Balázs Szalkai, Csaba Kerepesi, Bálint Varga, and Vince Grolmusz. The Budapest Reference Connectome Server v2.0. *Neuroscience Letters*, 595:60–62, 2015.

8. Csaba Kerepesi, Balázs Szalkai, Bálint Varga, and Vince Grolmusz. Comparative Connectomics: Mapping the Inter-Individual Variability of Connections within the Regions of the Human Brain. *Neuroscience Letters, Vol 662, pp. 17–21*, 2018.

9. Csaba Kerepesi, Balázs Szalkai, Bálint Varga, and Vince Grolmusz. How to Direct the Edges of the Connectomes: Dynamics of the Consensus Connectomes and the Development of the Connections in the Human Brain. *PLOS ONE, 11(6): e0158680*, 2016.

References

[1] Csaba Kerepesi, Dániel Bánky, and Vince Grolmusz. AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene*, 533(2):538–540, 2014.

[2] Csaba Kerepesi, Balázs Szalkai, and Vince Grolmusz. Visual Analysis of the Quantitative Composition of Metagenomic Communities: the AmphoraVizu Webserver. *Microbial Ecology*, 69(3):695–697, 2015.

[3] Csaba Kerepesi and Vince Grolmusz. Evaluating the Quantitative Capabilities of Metagenomic Analysis Software. *Current Microbiology*, 72(5):612–616, 2016.

[4] Csaba Kerepesi and Vince Grolmusz. The "Giant Virus Finder" Discovers an Abundance of Giant Viruses in the Antarctic Dry Valleys. *Archives of Virology*, 162(6):1671–1676, 2017.

[5] Csaba Kerepesi and Vince Grolmusz. Giant viruses of the Kutch Desert. *Archives of Virology*, Volume 161, Issue 3, pp 721–724, 2016.

[6] Csaba Kerepesi, Judit E Szabó, Veronika Papp-Kádár, Orsolya Dobay, Dóra Szabó, Vince Grolmusz, and Beata G Vertessy. Life without dUTPase. *Frontiers in Microbiology*, 7:1768, 2016.

[7] Balázs Szalkai, Csaba Kerepesi, Bálint Varga, and Vince Grolmusz. The Budapest Reference Connectome Server v2.0. *Neuroscience Letters*, 595:60–62, 2015.

[8] Csaba Kerepesi, Balázs Szalkai, Bálint Varga, and Vince Grolmusz. Comparative Connectomics: Mapping the Inter-Individual Variability of Connections within the Regions of the Human Brain. *Neuroscience Letters, Vol 662, pp. 17–21*, 2018.

[9] Csaba Kerepesi, Balázs Szalkai, Bálint Varga, and Vince Grolmusz. How to Direct the Edges of the Connectomes: Dynamics of the Consensus Connectomes and the Development of the Connections in the Human Brain. *PLOS ONE, 11(6): e0158680*, 2016.

[10] Balázs Szalkai, Csaba Kerepesi, Bálint Varga, Vince Grolmusz. Parameterizable Consensus Connectomes from the Human Connectome Project: The Budapest Reference Connectome Server v3.0. *Neuroscience Letters, Vol 662, pp. 17–21*, 2016.

[11] Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Research* 2016;44(Database issue):D20-D26.

[12] Nakazato, Takeru, Tazro Ohta, and Hidemasa Bono. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One* 8.10: e77910, 2013.

[13] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. *Elsevier*, 2011.

[14] James R. Anderson, Bryan W. Jones, Carl B. Watt, Margaret V. Shaw, Jia-Hui Yang, David Demill, James S. Lauritzen, Yanhua Lin, Kevin D. Rapp, David Mastronarde, Pavel Koshevoy, Bradley Grimm, Tolga Tasdizen, Ross Whitaker, and Robert E. Marc. Exploring the retinal connectome. *Mol Vis*, 17:355–379, 2011.

[15] Dmitri B. Chklovskii, Shiv Vitaladevuni, and Louis K. Scheffer. Semi-automated reconstruction of neural circuits using electron microscopy. *Curr Opin Neurobiol*, 20(5):667–675, Oct 2010.

[16] Siemon C. de Lange, Marcel A. de Reus, and Martijn P. van den Heuvel. TheLaplacian spectrum of neural networks. *Front Comput Neurosci*, 7:189, Jan 2014.

[17] Cole Gilbert. Brain connectivity: revealing the fly visual motion circuit. *Curr Biol*, 23(18):R851–R853, Sep 2013.

[18] Daniel Ginsburg, Stephan Gerhard, John Edgar Congote, and Rudolph Pienaar. Realtime visualization of the connectome in the browser using webgl. *Frontiers in Neuroinformatics*, 2011.

[19] Patric Hagmann, Leila Cammoun, Xavier Gigandet, Reto Meuli, Christopher J. Honey, Van J. Wedeen, and Olaf Sporns. Mapping the structural core of human cerebral cortex. *PLoS Biol*, 6(7):e159, Jul 2008.

[20] Heidi Johansen-Berg. Human connectomics - what will the future demand? *Neuroimage*, 80:541–544, Oct 2013.

[21] Emma K. Towlson, Petra E. Vértes, Sebastian E. Ahnert, William R. Schafer, and Edward T. Bullmore. The rich club of the c. elegans neuronal connectome. *J Neurosci*, 33(15):6380–6387, Apr 2013.

[22] Chaehyun Yook, Shaul Druckmann, and Jinhyun Kim. Mapping mammalian synaptic connectivity. *Cell Mol Life Sci*, 70(24):4747–4757, Dec 2013.

[23] Chantal Abergel, Matthieu Legendre, and Jean-Michel Claverie. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiology Reviews*, 39(6):779–796, 2015.

[24] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[25] J. Amar, M. Serino, C. Lange, C. Chabo, J. Iacovoni, S. Mondot, P. Lepage, C. Klopp, J. Mariette, O. Bouchez, L. Perez, M. Courtney, M. Marre, P. Klopp, O. Lantieri, J. Dore, M. Charles, B. Balkau, R. Burcelin, and D.E.S.I.R. Study Group. Involvement of tissue bacteria in the onset of diabetes in humans: evidence for a concept. *Diabetologia*, 54(12):3055–3061, Dec 2011.

[26] David Arndt, Jianguo Xia, Yifeng Liu, You Zhou, An Chi Guo, Joseph A. Cruz, Igor Sinelnikov, Karen Budwill, Camilla L. Nesbø, and David S. Wishart. METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Res*, 40(Web Server issue):W88–W95, Jul 2012.

[27] Defne Arslan, Matthieu Legendre, Virginie Seltzer, Chantal Abergel, and Jean-Michel Claverie. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci U S A*, 108(42):17486–17491, Oct 2011.

[28] Janelle C Arthur, Ernesto Perez-Chanona, Marcus Muhlbauer, Sarah Tomkovich, Joshua M Uronis, Ting-Jia Fan, Barry J Campbell, Turki Abujamel, Belgin Dogan, Arlin B Rogers, Jonathan M Rhodes, Alain Stintzi, Kenneth W Simpson, Jonathan J Hansen, Temitope O Keku, Anthony A Fodor, and Christian Jobin. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science*, 338(6103):120–123, Oct 2012.

[29] Koji Atarashi and Kenya Honda. Microbiota in autoimmunity and tolerance. *Curr Opin Immunol*, 23(6):761–768, Dec 2011.

[30] L. B. Bindels, P. Porporato, E. M. Dewulf, J. Verrax, A. M. Neyrinck, J. C. Martin, K. P. Scott, P. Buc Calderon, O. Feron, G. G. Muccioli, P. Sonveaux, P. D. Cani, and N. M. Delzenne. Gut microbiota-derived propionate reduces cancer cell proliferation in the liver. *Br J Cancer*, 107(8):1337–1344, Oct 2012.

[31] Brian P Boerner and Nora E Sarvetnick. Type 1 diabetes: role of intestinal microbiome in humans and mice. *Ann N Y Acad Sci*, 1243:103–118, Dec 2011.

[32] Mickael Boyer, Natalya Yutin, Isabelle Pagnier, Lina Barrassi, Ghislain Fournous, Leon Espinosa, Catherine Robert, Saïd Azza, Siyang Sun, Michael G Rossmann, Marie Suzan-Monti, Bernard La Scola, Eugene V Koonin, and Didier Raoult. Giant marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci U S A*, 106(51):21848–21853, Dec 2009.

[33] Mickaël Boyer, Natalya Yutin, Isabelle Pagnier, Lina Barrassi, Ghislain Fournous, Leon Espinosa, Catherine Robert, Saïd Azza, Siyang Sun,

Michael G. Rossmann, Marie Suzan-Monti, Bernard La Scola, Eugene V. Koonin, and Didier Raoult. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci U S A*, 106(51):21848–21853, Dec 2009.

[34] P. D. Burbelo, A. Bayat, E. E. Lebovitz, and M. J. Iadarola. New technologies for studying the complexity of oral diseases. *Oral Dis*, 18(2):121–126, Mar 2012.

[35] Rafael K. Campos, Paulo V. Boratto, Felipe L. Assis, Eric R G R. Aguiar, Lorena C F. Silva, Jonas D. Albarnaz, Fabio P. Dornas, Giliane S. Trindade, Paulo P. Ferreira, João T. Marques, Catherine Robert, Didier Raoult, Erna G. Kroon, Bernard La Scola, and Jônatas S. Abrahão. Samba virus: a novel mimivirus from a giant rain forest, the Brazilian Amazon. *Virol J*, 11:95, 2014.

[36] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, Jul 2012.

[37] Patrice D. Cani. Gut microbiota and obesity: lessons from the microbiome. *Brief Funct Genomics*, 12(4):381–387, Jul 2013.

[38] Philippe Colson, Natalya Yutin, Svetlana A Shabalina, Catherine Robert, Ghislain Fournous, Bernard La Scola, Didier Raoult, and Eugene V Koonin. Viruses with more than 1,000 genes: Mamavirus, a new Acanthamoeba polyphaga mimivirus strain, and reannotation of Mimivirus genes. *Genome Biol Evol*, 3:737–742, 2011.

[39] Committee on Metagenomics: Challenges and Functional Applications, National Research Council. *The New Science of Metagenomics: Reveal-*

ing the Secrets of Our Microbial Planet. The National Academies Press, 2007.

[40] Michael A Conlon, Caroline A Kerr, Christopher S McSweeney, Robert A Dunne, Janet M Shaw, Seungha Kang, Anthony R Bird, Matthew K Morell, Trevor J Lockett, Peter L Molloy, Ahmed Regina, Shusuke Toden, Julie M Clarke, and David L Topping. Resistant starches protect against colonic DNA damage and alter microbiota and gene expression in rats fed a Western diet. *J Nutr*, 142(5):832–840, May 2012.

[41] Daniel de Oliveira, Kary ACS Ocaña, Eduardo Ogasawara, Jonas Dias, João Gonçalves, Fernanda Baião, and Marta Mattoso. Performance evaluation of parallel strategies in public clouds: A study with phylogenomic workflows. *Future Generation Computer Systems*, 2013.

[42] Sridevi Devaraj, Peera Hemarajata, and James Versalovic. The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clin Chem*, 59(4):617–628, Apr 2013.

[43] Scott Diguistini, Nancy Y Liao, Darren Platt, Gordon Robertson, Michael Seidel, Simon K Chan, T. Roderick Docking, Inanc Birol, Robert A Holt, Martin Hirst, Elaine Mardis, Marco A Marra, Richard C Hamelin, Jorg Bohlmann, Colette Breuil, and Steven Jm Jones. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol*, 10(9):R94, 2009.

[44] Johannes Dröge and Alice C McHardy. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in bioinformatics*, 13(6):646–655, 2012.

[45] Sean R Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Inform*, 23(1):205–211, Oct 2009.

[46] Matthias G Fischer, Michael J Allen, William H Wilson, and Curtis A Suttle. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci U S A*, 107(45):19508–19513, Nov 2010.

[47] D Randy Garza and Curtis A Suttle. Large double-stranded DNA viruses which cause the lysis of a marine heterotrophic nanoflagellate (bodo sp.) occur in natural marine viral communities. *Aquatic Microbial Ecology*, 9(3):203–210, 1995.

[48] Elodie Ghedin and Jean-Michel Claverie. Mimivirus relatives in the Sargasso sea. *Virol J*, 2:62, 2005.

[49] Romina S Goldszmid and Giorgio Trinchieri. The price of immunity. *Nat Immunol*, 13(10):932–938, Oct 2012.

[50] Sharon Greenblum, Peter J. Turnbaugh, and Elhanan Borenstein. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A*, 109(2):594–599, Jan 2012.

[51] Iman Hajirasouliha, Fereydoun Hormozdiari, S. Cenk Sahinalp, and Inanc Birol. Optimal pooling for genome re-sequencing with ultra-high-throughput short-read technologies. *Bioinformatics*, 24(13):i32–i40, Jul 2008.

[52] Gabriele Hormannsperger, Thomas Clavel, and Dirk Haller. Gut matters: microbe-host interactions in allergic diseases. *J Allergy Clin Immunol*, 129(6):1452–1459, Jun 2012.

[53] Daniel H. Huson, Alexander F. Auch, Ji Qi, and Stephan C. Schuster. MEGAN analysis of metagenomic data. *Genome Res*, 17(3):377–386, Mar 2007.

[54] Daniel H. Huson and Suparna Mitra. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol Biol*, 856:415–429, 2012.

[55] Daniel H. Huson, Daniel C. Richter, Suparna Mitra, Alexander F. Auch, and Stephan C. Schuster. Methods for comparative metagenomics. *BMC Bioinformatics*, 10 Suppl 1:S12, 2009.

[56] Shaun D Jackman and Inanc Birol. Assembling genomes using short-read sequencing technology. *Genome Biol*, 11(1):202, 2010.

[57] R. Jain, M. C. Rivera, and J. A. Lake. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*, 96(7):3801–3806, Mar 1999.

[58] L. Steven Johnson, Sean R Eddy, and Elon Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11:431, 2010.

[59] R. S. Kootte, A. Vrieze, F. Holleman, G. M. Dallinga-Thie, E. G. Zoetendal, W. M. de Vos, A. K. Groen, J. B L Hoekstra, E. S. Stroes, and M. Nieuwdorp. The therapeutic potential of manipulating gut microbiota in obesity and type 2 diabetes mellitus. *Diabetes Obes Metab*, 14(2):112–120, Feb 2012.

[60] Bernard La Scola, Stéphane Audic, Catherine Robert, Liang Jungang, Xavier de Lamballerie, Michel Drancourt, Richard Birtles, Jean-Michel

Claverie, and Didier Raoult. A giant virus in amoebae. *Science*, 299(5615):2033, Mar 2003.

[61] Matthieu Legendre, Julia Bartoli, Lyubov Shmakova, Sandra Jeudy, Karine Labadie, Annie Adrait, Magali Lescot, Olivier Poirot, Lionel Bertaux, Christophe Bruley, Yohann Couté, Elizaveta Rivkina, Chantal Abergel, and Jean-Michel Claverie. Thirty-thousand-year-old distant relative of giant icosahedral dna viruses with a pandoravirus morphology. *Proc Natl Acad Sci U S A*, 111(11):4274–4279, Mar 2014.

[62] Thomas Lingner, Kathrin Petra Asshauer, Fabian Schreiber, and Peter Meinicke. CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic Acids Res*, 39(Web Server issue):W518–W523, Jul 2011.

[63] Norman J. MacDonald, Donovan H. Parks, and Robert G. Beiko. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res*, 40(14):e111, Aug 2012.

[64] Diane Mathis and Christophe Benoist. Microbiota and autoimmune disease: the hosted self. *Cell Host Microbe*, 10(4):297–301, Oct 2011.

[65] Konstantinos Mavromatis, Natalia Ivanova, Kerrie Barry, Harris Shapiro, Eugene Goltsman, Alice C. McHardy, Isidore Rigoutsos, Asaf Salamov, Frank Korzeniewski, Miriam Land, Alla Lapidus, Igor Grigoriev, Paul Richardson, Philip Hugenholtz, and Nikos C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4(6):495–500, Jun 2007.

[66] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A.

Edwards. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[67] Monzoorul Haque Mohammed, Tarini Shankar Ghosh, Nitin Kumar Singh, and Sharmila S. Mande. SPHINX–an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, 27(1):22–30, Jan 2011.

[68] Josef Neu, Graciela Lorca, Sandra D K Kingma, and Eric W Triplett. The intestinal microbiome: relationship to type 1 diabetes. *Endocrinol Metab Clin North Am*, 39(3):563–571, Sep 2010.

[69] A. S. Pandit, M. N. Joshi, P. Bhargava, G. N. Ayachit, I. M. Shaikh, Z. M. Saiyed, A. K. Saxena, and S. B. Bagatharia. Metagenomes from the saline desert of Kutch. *Genome Announc*, 2(3), 2014.

[70] Nadège Philippe, Matthieu Legendre, Gabriel Doutre, Yohann Couté, Olivier Poirot, Magali Lescot, Defne Arslan, Virginie Seltzer, Lionel Bertaux, Christophe Bruley, Jérome Garin, Jean-Michel Claverie, and Chantal Abergel. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, 341(6143):281–286, Jul 2013.

[71] Claudia S Plottel and Martin J Blaser. Microbiome and malignancy. *Cell Host Microbe*, 10(4):324–335, Oct 2011.

[72] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu,

Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, Oct 2012.

[73] Didier Raoult, Stephane Audic, Catherine Robert, Chantal Abergel, Patricia Renesto, Hiroyuki Ogata, Bernard La Scola, Marie Suzan, and Jean-Michel Claverie. The 1.2-megabase genome sequence of Mimivirus. *Science*, 306(5700):1344–1350, Nov 2004.

[74] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–277, Jun 2000.

[75] Daniel C. Richter, Felix Ott, Alexander F. Auch, Ramona Schmid, and Daniel H. Huson. Metasim: a sequencing simulator for genomics and metagenomics. *PLoS One*, 3(10):e3373, 2008.

[76] Jose U Scher and Steven B Abramson. The microbiome and rheumatoid arthritis. *Nat Rev Rheumatol*, 7(10):569–578, Oct 2011.

[77] Robert F Schwabe and Timothy C Wang. Cancer: Bacteria deliver a genotoxic hit. *Science*, 338(6103):52–53, Oct 2012.

[78] Nicola Segata, Daniela Boernigen, Timothy L Tickle, Xochitl C Morgan, Wendy S Garrett, and Curtis Huttenhower. Computational meta'omics for microbial community studies. *Molecular systems biology*, 9(1), 2013.

[79] Daniel B Sloan and Nancy A Moran. Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Molecular biology and evolution*, 29(12):3781–3792, 2012.

[80] Jason E Stajich. An introduction to BioPerl. *Methods Mol Biol*, 406:535–548, 2007.

[81] Jason E Stajich, David Block, Kris Boulez, Steven E Brenner, Stephen A Chervitz, Chris Dagdigian, Georg Fuellen, James G R Gilbert, Ian Korf, Hilmar Lapp, Heikki Lehvaslaiho, Chad Matsalla, Chris J Mungall, Brian I Osborne, Matthew R Pocock, Peter Schattner, Martin Senger, Lincoln D Stein, Elia Stupka, Mark D Wilkinson, and Ewan Birney. The BioPerl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618, Oct 2002.

[82] A. Stamatakis, A. J. Aberer, C. Goll, S. A. Smith, S. A. Berger, and F. Izquierdo-Carrasco. RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics*, 28(15):2064–2066, Aug 2012.

[83] A. Stamatakis, T. Ludwig, and H. Meier. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463, Feb 2005.

[84] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core

gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, Jan 2009.

[85] Outi Vaarala. Is the origin of type 1 diabetes in the gut? *Immunol Cell Biol*, 90(3):271–276, Mar 2012.

[86] J Craig Venter, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O. Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, Apr 2004.

[87] Zhang Wang and Martin Wu. A phylum-level bacterial phylogenetic marker database. *Mol Biol Evol*, 30:1258–1262, 2013.

[88] Zi-Kai Wang and Yun-Sheng Yang. Upper gastrointestinal microbiota and digestive diseases. *World J Gastroenterol*, 19(10):1541–1550, Mar 2013.

[89] Hsin-Jung Wu and Eric Wu. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes*, 3(1):4–14, 2012.

[90] Martin Wu and Jonathan A Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, 9(10):R151, 2008.

[91] Martin Wu and Alexandra J Scott. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, 28(7):1033–1034, Apr 2012.

[92] Sitao Wu, Zhengwei Zhu, Liming Fu, Beifang Niu, and Weizhong Li. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics*, 12:444, 2011.

[93] Sheree Yau, Federico M Lauro, Matthew Z DeMaere, Mark V Brown, Torsten Thomas, Mark J Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffrey M Hoffman, John A Gibson, and Ricardo Cavicchioli. Virophage control of antarctic algal host-virus dynamics. *Proc Natl Acad Sci U S A*, 108(15):6163–6168, Apr 2011.

[94] Jean-Michel Claverie, Hiroyuki Ogata, Stéphane Audic, Chantal Abergel, Karsten Suhre, and Pierre-Edouard Fournier. Mimivirus and the emerging concept of "giant" virus. *Virus research*, 117(1):133–144, 2006.

[95] Philippe Colson, Xavier de Lamballerie, Ghislain Fournous, and Didier Raoult. Reclassification of giant viruses composing a fourth domain of life in the new order megavirales. *Intervirology*, 55(5):321–332, 2011.

[96] Philippe Colson, Gregory Gimenez, Mickaël Boyer, Ghislain Fournous, and Didier Raoult. The giant cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of life. *PLoS One*, 6(4):e18935, 2011.

[97] Matthieu Legendre, Defne Arslan, Chantal Abergel, and Jean-Michel Claverie. Genomics of megavirus and the elusive fourth domain of life. *Communicative & integrative biology*, 5(1):102–106, 2012.

[98] Tom A Williams, T Martin Embley, and Eva Heinz. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS One*, 6(6):e21080, 2011.

[99] Matthias G Fischer and Curtis A Suttle. A virophage at the origin of large DNA transposons. *Science*, 332(6026):231–234, 2011.

[100] Christelle Desnues, Bernard La Scola, Natalya Yutin, Ghislain Fournous, Catherine Robert, Saïd Azza, Priscilla Jardot, Sonia Monteil, Angélique Campocasso, Eugene V Koonin, et al. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proceedings of the National Academy of Sciences*, 109(44):18078–18083, 2012.

[101] Pierre-Alain Jachiet, Philippe Colson, Philippe Lopez, and Eric Bapteste. Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. *Genome biology and evolution*, 6(9):2195–2205, 2014.

[102] Noah Fierer, Jonathan W. Leff, Byron J. Adams, Uffe N. Nielsen, Scott Thomas Bates, Christian L. Lauber, Sarah Owens, Jack A. Gilbert, Diana H. Wall, and J Gregory Caporaso. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci U S A*, 109(52):21390–21395, Dec 2012.

[103] Stephen F. Altschul, John C. Wootton, E Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schaffer, and Yi-Kuo Yu. Protein database searches using compositionally adjusted substitution matrices. *FEBS J*, 272(20):5101–5109, Oct 2005.

[104] Daniel H. Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Res*, 21(9):1552–1560, Sep 2011.

[105] Jennifer A. McNab, Brian L. Edlow, Thomas Witzel, Susie Y. Huang, Himanshu Bhat, Keith Heberlein, Thorsten Feiweier, Kecheng Liu, Boris

Keil, Julien Cohen-Adad, M Dylan Tisdall, Rebecca D. Folkerth, Hannah C. Kinney, and Lawrence L. Wald. The Human Connectome Project and beyond: initial applications of 300 mT/m gradients. *Neuroimage*, 80:234–245, Oct 2013. doi: 10.1016/j.neuroimage.2013.05.074. URL http://dx.doi.org/10.1016/j.neuroimage.2013.05.074.

[106] Madhura Ingalhalikar, Alex Smith, Drew Parker, Theodore D. Satterthwaite, Mark A. Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E. Gur, Ruben C. Gur, and Ragini Verma. Sex differences in the structural connectome of the human brain. *Proc Natl Acad Sci U S A*, 111(2):823–828, Jan 2014. doi: 10.1073/pnas.1316909110. URL http://dx.doi.org/10.1073/pnas.1316909110.

[107] Balázs Szalkai, Bálint Varga, and Vince Grolmusz. Graph theoretical analysis reveals: Women's brains are better connected than men's. *PLOS One*, 10(7):e0130045, July 2015a. doi: doi:10.1371/journal.pone.0130045. URL http://dx.plos.org/10.1371/journal.pone.0130045.

[108] Patric Hagmann, Patricia E. Grant, and Damien A. Fair. Mr connectomics: a conceptual framework for studying the developing brain. *Front Syst Neurosci*, 6:43, 2012. doi: 10.3389/fnsys.2012.00043. URL http://dx.doi.org/10.3389/fnsys.2012.00043.

[109] R Cameron Craddock, Michael P. Milham, and Stephen M. LaConte. Predicting intrinsic brain activity. *Neuroimage*, 82:127–136, Nov 2013. doi: 10.1016/j.neuroimage.2013.05.072. URL http://dx.doi.org/10.1016/j.neuroimage.2013.05.072.

[110] K. A. Roth and C. D'Sa. Apoptosis and brain development. *Ment Retard Dev Disabil Res Rev*, 7(4):261–266, 2001. doi: 10.1002/mrdd.1036. URL `http://dx.doi.org/10.1002/mrdd.1036`.

[111] Keiko Nonomura, Yoshifumi Yamaguchi, Misato Hamachi, Masato Koike, Yasuo Uchiyama, Kenichi Nakazato, Atsushi Mochizuki, Asako Sakaue-Sawano, Atsushi Miyawaki, Hiroki Yoshida, Keisuke Kuida, and Masayuki Miura. Local apoptosis modulates early mammalian brain development through the elimination of morphogen-producing cells. *Dev Cell*, 27(6):621–634, Dec 2013. doi: 10.1016/j.devcel.2013.11.015. URL `http://dx.doi.org/10.1016/j.devcel.2013.11.015`.

[112] N Gordon. Apoptosis (programmed cell death) and other reasons for elimination of neurons and axons. *Brain & development*, 17(1):73–77, 1995.

[113] Bohr, V.A., Stevnsner, T. and de Souza-Pinto, N.C. (2002) Mitochondrial DNA repair of oxidative damage in mammalian cells. Gene, 286, 127-134.

[114] Lu, A.L., Li, X., Gu, Y., Wright, P.M. and Chang, D.Y. (2001) Repair of oxidative DNA damage: mechanisms and functions. Cell Biochem Biophys, 35, 141-170.

[115] Vertessy, B.G. and Toth, J. (2009) Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases. Acc Chem Res, 42, 97-106.

[116] Lindahl, T. (1993) Instability and decay of the primary structure of DNA [see comments]. Nature, 362, 709-715.

[117] Meier, B. and Gartner, A. (2014) Having a direct look: analysis of DNA damage and repair mechanisms by next generation sequencing. Exp Cell Res, 329, 35-41.

[118] Kana, B.D. and Mizrahi, V. (2004) Molecular genetics of Mycobacterium tuberculosis in relation to the discovery of novel drugs and vaccines. Tuberculosis, 84, 63-75.

[119] Boshoff, H.I., Reed, M.B., Barry, C.E., 3rd and Mizrahi, V. (2003) DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. Cell, 113, 183-193.

[120] Jiricny, J. (1998) Replication errors: cha(lle)nging the genome. EMBO J, 17, 6427-6436.

[121] Visnes, T., Doseth, B., Pettersen, H.S., Hagen, L., Sousa, M.M., Akbari, M., Otterlei, M., Kavli, B., Slupphaug, G. and Krokan, H.E. (2009) Uracil in DNA and its processing by different DNA glycosylases. Philos Trans R Soc Lond B Biol Sci, 364, 563-568.

[122] Galperin, M.Y., Moroz, O.V., Wilson, K.S. and Murzin, A.G. (2006) House cleaning, a part of good housekeeping. Mol Microbiol, 59, 5-19.

[123] Nagy, G.N., Leveles, I. and Vertessy, B.G. (2014) Preventive DNA repair by sanitizing the cellular (deoxy)nucleoside triphosphate pool. FEBS J, 281, 4207-4223.

[124] Pecsi, I., Hirmondo, R., Brown, A.C., Lopata, A., Parish, T., Vertessy, B.G. and Toth, J. (2012) The dUTPase enzyme is essential in Mycobacterium smegmatis. PloS one, 7, e37461.

[125] Muha, V., Horvath, A., Bekesi, A., Pukancsik, M., Hodoscsek, B., Merenyi, G., Rona, G., Batki, J., Kiss, I., Jankovics, F. et al. (2012) Uracil-containing DNA in Drosophila: stability, stage-specific accumulation, and developmental involvement. PLoS Genet, 8, e1002738.

[126] Dengg, M., Garcia-Muse, T., Gill, S.G., Ashcroft, N., Boulton, S.J. and Nilsen, H. (2006) Abrogation of the CLK-2 checkpoint leads to tolerance to base-excision repair intermediates. EMBO Rep, 7, 1046-1051.

[127] Castillo-Acosta, V.M., Aguilar-Pereyra, F., Vidal, A.E., Navarro, M., Ruiz-Perez, L.M. and Gonzalez-Pacanowska, D. (2012) Trypanosomes lacking uracil-DNA glycosylase are hypersensitive to antifolates and present a mutator phenotype. Int J Biochem Cell Biol, 44, 1555-1568.

[128] Szabo, J.E., Nemeth, V., Papp-Kadar, V., Nyiri, K., Leveles, I., Bendes, A.A., Zagyva, I., Rona, G., Palinkas, H.L., Besztercei, B. et al. (2014) Highly potent dUTPase inhibition by a bacterial repressor protein reveals a novel mechanism for gene expression control. Nucleic Acids Res, 42, 11912-11920.

[129] Guillet, M., Van Der Kemp, P.A. and Boiteux, S. (2006) dUTPase activity is critical to maintain genetic stability in Saccharomyces cerevisiae. Nucleic Acids Res, 34, 2056-2066.

[130] el-Hajj, H.H., Zhang, H. and Weiss, B. (1988) Lethality of a dut (deoxyuridine triphosphatase) mutation in Escherichia coli. J Bacteriol, 170, 1069-1075.

[131] Nilsen, H., Rosewell, I., Robins, P., Skjelbred, C.F., Andersen, S., Slupphaug, G., Daly, G., Krokan, H.E., Lindahl, T. and Barnes, D.E. (2000)

Uracil-DNA glycosylase (UNG)-deficient mice reveal a primary role of the enzyme during DNA replication. Mol Cell, 5, 1059-1065.

[132] Aravind, L. and Koonin, E.V. (2000) The alpha/beta fold uracil DNA glycosylases: a common origin with diverse fates. Genome Biol, 1, RESEARCH0007.

[133] el-Hajj, H.H., Wang, L. and Weiss, B. (1992) Multiple mutant of Escherichia coli synthesizing virtually thymineless DNA during limited growth. J Bacteriol, 174, 4450-4456.

[134] Castillo-Acosta, V.M., Aguilar-Pereyra, F., Bart, J.M., Navarro, M., Ruiz-Perez, L.M., Vidal, A.E. and Gonzalez-Pacanowska, D. (2012) Increased uracil insertion in DNA is cytotoxic and increases the frequency of mutation, double strand break formation and VSG switching in Trypanosoma brucei. DNA Repair (Amst), 11, 986-995.

[135] Golding, G.R., Bryden, L., Levett, P.N., McDonald, R.R., Wong, A., Graham, M.R., Tyler, S., Van Domselaar, G., Mabon, P., Kent, H. et al. (2012) whole-genome sequence of livestock-associated st398 methicillin-resistant staphylococcus aureus Isolated from Humans in Canada. J Bacteriol, 194, 6627-6628.

[136] Chen, C.J., Unger, C., Hoffmann, W., Lindsay, J.A., Huang, Y.C. and Gotz, F. (2013) Characterization and comparison of 2 distinct epidemic community-associated methicillin-resistant Staphylococcus aureus clones of ST59 lineage. PloS one, 8, e63210.

[137] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol, 215, 403-410.

[138] Huson, D.H. and Mitra, S. (2012) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods in molecular biology, 856, 415-429.

[139] Serrano-Heras, G., Salas, M. and Bravo, A. (2006) A uracil-DNA glycosylase inhibitor encoded by a non-uracil containing viral DNA. The Journal of biological chemistry, 281, 7068-7074.

[140] Cole, A.R., Ofer, S., Ryzhenkova, K., Baltulionis, G., Hornyak, P. and Savva, R. (2013) Architecturally diverse proteins converge on an analogous mechanism to inactivate Uracil-DNA glycosylase. Nucleic Acids Res, 41, 8760-8775.

[141] Serrano-Heras, G., Bravo, A. and Salas, M. (2008) Phage phi29 protein p56 prevents viral DNA replication impairment caused by uracil excision activity of uracil-DNA glycosylase. Proc Natl Acad Sci U S A, 105, 19044-19049.

[142] Wang, H.C., Hsu, K.C., Yang, J.M., Wu, M.L., Ko, T.P., Lin, S.R. and Wang, A.H. (2014) Staphylococcus aureus protein SAUGI acts as a uracil-DNA glycosylase inhibitor. Nucleic Acids Res, 42, 1354-1364.

[143] Tatsch, C.O., Wood, T.L., Chamakura, K.R. and Kuty Everett, G.F. (2013) Complete Genome of Salmonella enterica Serovar Typhimurium Myophage Maynard. Genome announcements, 1.

[144] Minton, K.W. and Daly, M.J. (1995) A model for repair of radiation-induced DNA double-strand breaks in the extreme radiophile Deinococcus radiodurans. BioEssays : news and reviews in molecular, cellular and developmental biology, 17, 457-464.

[145] Goncalves, A.M., de Sanctis, D. and McSweeney, S.M. (2011) Structural and functional insights into DR2231 protein, the MazG-like nucleoside triphosphate pyrophosphohydrolase from Deinococcus radiodurans. The Journal of biological chemistry, 286, 30691-30705.

[146] Merenyi, G., Kovari, J., Toth, J., Takacs, E., Zagyva, I., Erdei, A. and Vertessy, B.G. (2011) Cellular response to efficient dUTPase RNAi silencing in stable HeLa cell lines perturbs expression levels of genes involved in thymidylate metabolism. Nucleosides, nucleotides & nucleic acids, 30, 369-390.

[147] Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A. et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature, 399, 323-329.

[148] Feller, Willliam. An introduction to probability theory and its applications. Vol. 2. John Wiley & Sons, 2008.

[149] Morris Hirsch. *Differential Topology.* Springer-Verlag, 1997. ISBN 978-0-387-90148-0.

[150] J Tournier, Fernando Calamante, Alan Connelly, et al. Mrtrix: diffusion tractography in crossing fiber regions. *International Journal of Imaging Systems and Technology*, 22(1):53–66, 2012.

[151] Alessandro Daducci, Stephan Gerhard, Alessandra Griffa, Alia Lemkaddem, Leila Cammoun, Xavier Gigandet, Reto Meuli, Patric Hagmann, and Jean-Philippe Thiran. The connectome mapper: an open-source processing pipeline to map connectomes with MRI. *PLoS*

One, 7(12):e48121, 2012. doi: 10.1371/journal.pone.0048121. URL http://dx.doi.org/10.1371/journal.pone.0048121.

[1]**ADATLAP**
a doktori értekezés nyilvánosságra hozatalához

I. A doktori értekezés adatai

A szerző neve: Kerepesi Csaba
MTMT-azonosító: 10040326
A doktori értekezés címe és alcíme: Data mining in genomics, metagenomics, and connectomics
DOI-azonosító[2]: 10.15476/ELTE.2017.160
A doktori iskola neve: Informatika Doktori Iskola
A doktori iskolán belüli doktori program neve: Információs rendszerek
A témavezető neve és tudományos fokozata: Dr. Grolmusz Vince, egyetemi tanár
A témavezető munkahelye: ELTE TTK, Számítógéptudományi tanszék

II. Nyilatkozatok

1. A doktori értekezés szerzőjeként[3]

a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom az ELTE Informatika Doktori Iskola hivatalának ügyintézőjét Boda Annamáriát, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[4]

c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés (*dátum*)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[5]

d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követőn egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.[6]

2. A doktori értekezés szerzőjeként kijelentem, hogy

a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;

b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: 2017. október 29.

a doktori értekezés szerzőjének aláírása

[1] Beiktatta az Egyetemi Doktori Szabályzat módosításáról szóló CXXXIX/2014. (VI. 30.) Szen. sz. határozat. Hatályos: 2014. VII.1. napjától.
[2] A kari hivatal ügyintézője tölti ki.
[3] A megfelelő szöveg aláhúzandó.
[4] A doktori értekezés benyújtásával egyidejűleg be kell adni a tudományági doktori tanácshoz a szabadalmi, illetőleg oltalmi bejelentést tanúsító okiratot és a nyilvánosságra hozatal elhalasztása iránti kérelmet.
[5] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a minősített adatra vonatkozó közokiratot.
[6] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a mű kiadásáról szóló kiadói szerződést.