iterata MEDICAL SERVICES

real world medical data analysis
- we just observe the data
- we do not copy

Iterata AG Medical Services
Industriestrasse 8
5722 Gränichen | Switzerland
www.iterata.ch

We do not copy data; we bring the algorithm to them.
Applying algorithms on your secured data

CHAOS GAME REPRESENTATION (CGR) → covid19

Visualizing patterns – Published Algorithms applyied on Patient Covid19 Gen-Sequenz

Introduction

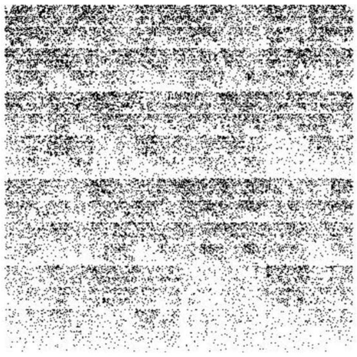
In organizations where sensitive data are collected – like medical data, banking system data, insurance data – a safe handling is very important. For that reason, one of the main principles of Iterata is that we do not copy data, we just observe them and bring the algorithm to the data. The collected data stays in the policy space of the organization. All data are referenced and anonymized.

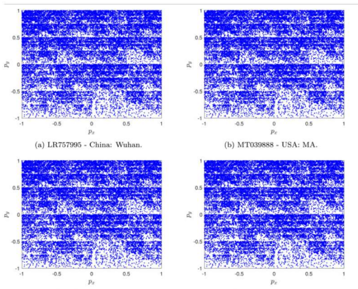
The collaboration of domain experts and researcher with application engineers is a big part of this firm. Through this cooperation, it is possible to transfer domain knowledge into an application.

We want to show this based on a short example. The SARS-CoV-2 keeps the researcher in medicine busy at the moment. Novel approaches are necessary to find pattern and understand the virus better. Based on the paper of Raquel de M. Barbosa and colleague ("Chaos game representation dataset of SARS-CoV-2 genome", 2020, Data in Brief 30, https://www.sciencedirect.com/science/article/pii/S2352340920305126) it was possible for our team to extract an algorithm and apply it on a curated data set of Corona sequences from Wolfram Mathematica to simulate the chaos game representation (CGR) with real-life data. The result is the transfer from a genome sequence to a figure where patterns are recognizable.

This whole process from knowledge acquisition, to understanding and extracting the algorithm and apply them on a given data set within its safe policy space took place within two weeks.

The merge of domain knowledge, understanding of relevant information and IT skills and knowledge opens up new possibility to pass on information.





(a) LR757995 - China: Wuhan. (b) MT039888 - USA: MA.
(c) MT126808 - Brazil. (d) MT066156 - Italy.

Already implemented possibilities

There are various ways what to do with one data set: three-dimensional representation of molecules and proteins; tracing maps where the virus to see the origin and the distribution of it; phylogenetic tree of the virus; etc.

Computational Essays

Modeling: Compartmental
- "Epidemiological Models for Influenza and COVID-19," by Robert Nachbar
- "The SIR Model for Spread of Disease," by Arnoud Buzing
- "COVID-19 - Policy Simulator," by Jan Brugård, Ankit Naik, Malte Lenz and Sergio Vargas
- "Basic Experiments Workflow for Simple Epidemiological Models," by Anton Antonov
- "Scaling of Epidemiology Models with Multi-site Compartments," by Anton Antonov

Data Processing
- "Scraping OpenTable's 'State of the Industry' page," by Aaron Enright
- "Tracking Coronavirus Testing in the United States," by Robert Rimmer
- "Apple Mobility Trends Data Visualization," by Anton Antonov
- "Web Scraper: New York Times Coronavirus Data," by Robert Rimmer
- "Population Density Map," by M. Bahrami

Genomics
- "Genome analysis and the SARS-nCoV-2," by Daniel Lichtblau
- "A walk-through of the SARS-CoV-2 nucleotide Wolfram resource," by John Cassel
- "Creating Dendrograms from Sequenced SARS-CoV-2 Genomes," by Daniel Lichtblau
- "Geometrical Analysis of Genome for COVID-19 vs. SARS-Like Viruses," by Mads Bahrami
- "Chaos Game for Clustering of Novel Coronavirus COVID-19," by Mads Bahrami

Data Analysis
- "COVID-19—The Swedish Experiment"
- "Exploring Pandemic Data: March 24 Livestream," by Stephen Wolfram
- "COVID-19 in Iran: An Issue of Underdiagnosis," by Mads Bahrami
- "COVID-19 Progress in Peru Macro Regions: Coast vs. Mountain vs. Jungle"
- "Comparing the spread of COVID-19 between countries," by Jan Brugard
- "What Countries Are Hit Hard by COVID-19 Outbreak?" by Mads Bahrami

Data Visualizations
- "3D Modeling of the SARS-CoV-2 Virus in the Wolfram Language"
- "Propagation risk of COVID-19 by local contact in Spain (March 10–14)," Bernat Espigulé
- "Mapping Novel Coronavirus COVID-19 Outbreak," by Jofre Espigule-Pons
- "US County Death Timeline Plots," by Bob Sandheinrich
- "Distance to nearest confirmed US COVID-19 case," by Chip Hurst
- "COVID-19 visualization of turning point," by Isao Maruyama
- "COVID-19 Epidemic Data: US County Level," by M. Bahrami

Modeling: Simulation
- "Epidemic simulation with a polygon container," by Francisco Rodríguez
- "Agent-based epidemic simulation," by Jon McLoone
- "Modeling the spatial spread of infection diseases in the US," by Diego Zviovich
- "Geo-spatial-temporal COVID-19 simulations and visualizations over USA," by Diego Zviovich

Modeling: Logistic
- "COVID-19 pandemic data in Italy & logistic curve," by Riccardo Fantoni
- "Logistic Model for Quarantine-Controlled Epidemics," by Robert Rimmer
- "Logistic Growth Model for COVID-19," by Mads Bahrami and Brian Wood

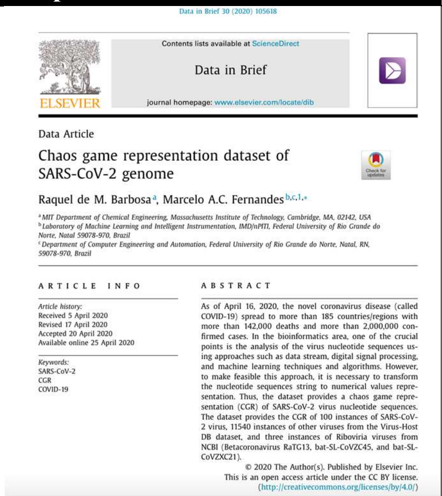
Modeling: Networking
- "Agent-Based Networks Models for COVID-19," by Christopher Wolfram

Please do not hesitate to contact us

Sincerely yours, Iterata Team

Phone +41 62 842 88 27 | info@iterata.ch

Insights & Impressions



Data in Brief 30 (2020) 105618

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Chaos game representation dataset of SARS-CoV-2 genome

Raquel de M. Barbosa[a], Marcelo A.C. Fernandes[b,c,1,*]

[a] MIT Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02142, USA
[b] Laboratory of Machine Learning and Intelligent Instrumentation, IMD/nPITI, Federal University of Rio Grande do Norte, Natal 59078-970, Brazil
[c] Department of Computer Engineering and Automation, Federal University of Rio Grande do Norte, Natal, RN, 59078-970, Brazil

ARTICLE INFO

Article history:
Received 5 April 2020
Revised 17 April 2020
Accepted 20 April 2020
Available online 25 April 2020

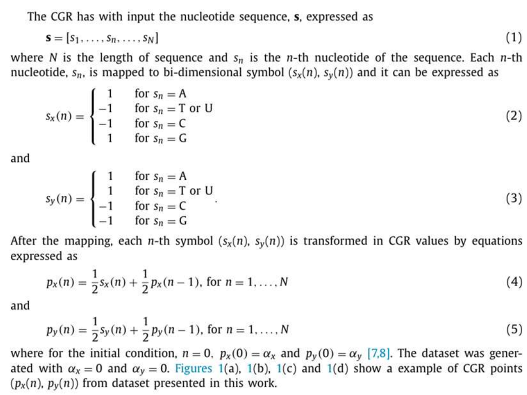
Keywords:
SARS-CoV-2
CGR
COVID-19

ABSTRACT

As of April 16, 2020, the novel coronavirus disease (called COVID-19) spread to more than 185 countries/regions with more than 142,000 deaths and more than 2,000,000 confirmed cases. In the bioinformatics area, one of the crucial points is the analysis of the virus nucleotide sequences using approaches such as data stream, digital signal processing, and machine learning techniques and algorithms. However, to make feasible this approach, it is necessary to transform the nucleotide sequences string to numerical values representation. Thus, the dataset provides a chaos game representation (CGR) of SARS-CoV-2 virus nucleotide sequences. The dataset provides the CGR of 100 instances of SARS-CoV-2 virus, 11540 instances of other viruses from the Virus-Host DB dataset, and three instances of Riboviria viruses from NCBI (Betacoronavirus RaTG13, bat-SL-CoVZC45, and bat-SL-CoVZXC21).

© 2020 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license.
(http://creativecommons.org/licenses/by/4.0/)

From a scientific paper…



The CGR has with input the nucleotide sequence, **s**, expressed as

$$\mathbf{s} = [s_1, \dots, s_n, \dots, s_N] \tag{1}$$

where N is the length of sequence and s_n is the n-th nucleotide of the sequence. Each n-th nucleotide, s_n, is mapped to bi-dimensional symbol $(s_x(n), s_y(n))$ and it can be expressed as

$$s_x(n) = \begin{cases} 1 & \text{for } s_n = \text{A} \\ -1 & \text{for } s_n = \text{T or U} \\ -1 & \text{for } s_n = \text{C} \\ 1 & \text{for } s_n = \text{G} \end{cases} \tag{2}$$

and

$$s_y(n) = \begin{cases} 1 & \text{for } s_n = \text{A} \\ 1 & \text{for } s_n = \text{T or U} \\ -1 & \text{for } s_n = \text{C} \\ -1 & \text{for } s_n = \text{G} \end{cases} \tag{3}$$

After the mapping, each n-th symbol $(s_x(n), s_y(n))$ is transformed in CGR values by equations expressed as

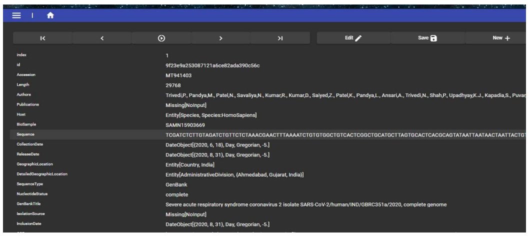
$$p_x(n) = \frac{1}{2}s_x(n) + \frac{1}{2}p_x(n-1), \text{ for } n = 1, \dots, N \tag{4}$$

and

$$p_y(n) = \frac{1}{2}s_y(n) + \frac{1}{2}p_y(n-1), \text{ for } n = 1, \dots, N \tag{5}$$

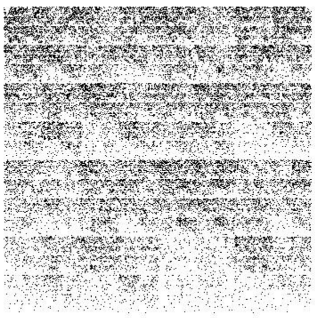
where for the initial condition, $n = 0$, $p_x(0) = \alpha_x$ and $p_y(0) = \alpha_y$ [7,8]. The dataset was generated with $\alpha_x = 0$ and $\alpha_y = 0$. Figures 1(a), 1(b), 1(c) and 1(d) show a example of CGR points $(p_x(n), p_y(n))$ from dataset presented in this work.

… to the extraction of the algorithm…

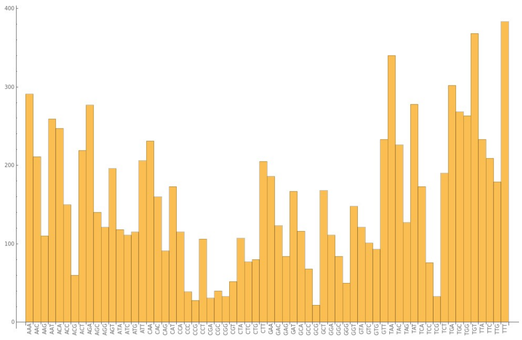


…and the application
Covid19 Patient Data [cov gen-sequence] at hospital location

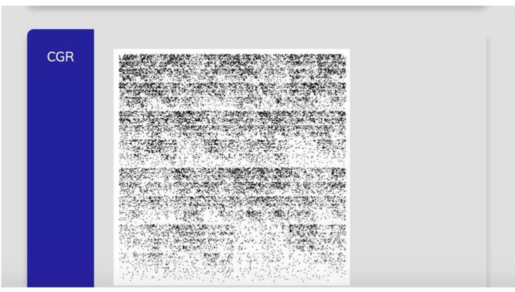
The whole sequence is deposited



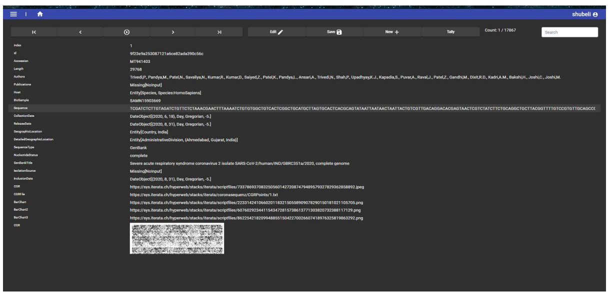
Run CGR algorithm and produce CGR image at data location of hospital



Run algorithmen for base pairs extraction of codones and acids



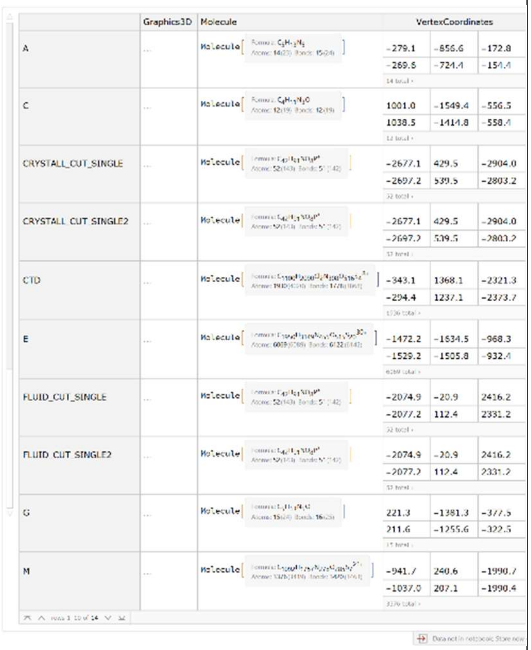
Corresponding CGR figure to the sequence



Publish produced images throughout existing application in data policy at hospital/laboratory location

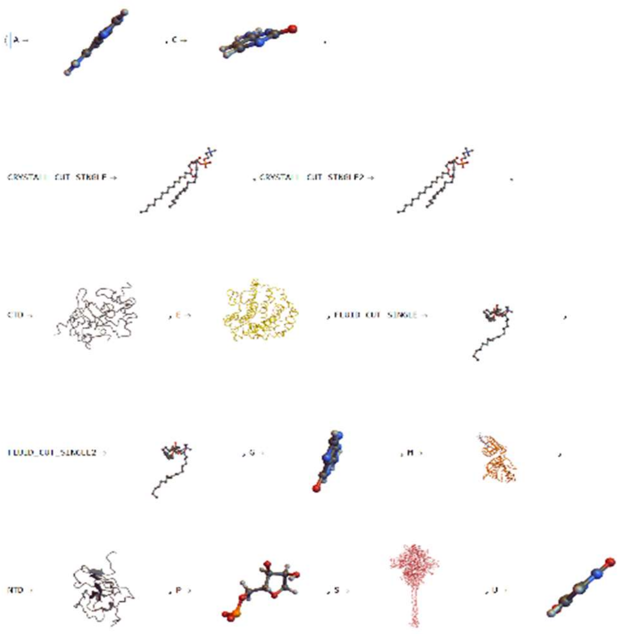
Iterata AG Medical Services
Industriestrasse 8
5722 Gränichen | Switzerland
www.iterata.ch

3d Modeling of the SARS-cov-2



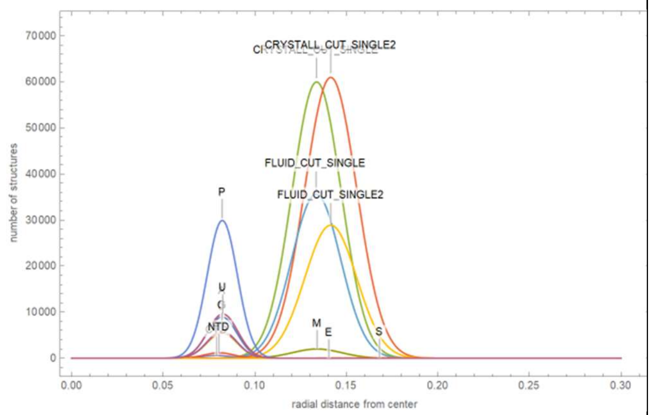
	Graphics3D	Molecule	VertexCoordinates		
A	...	Molecule[Atoms: 14(23) Bonds: 15(24)]	-279.1	-856.6	-172.8
			-269.6	-724.4	-154.4
			14 total ›		
C	...	Molecule[Atoms: 12(19) Bonds: 12(19)]	1001.0	-1549.4	-556.5
			1038.5	-1414.8	-558.4
			12 total ›		
CRYSTALL_CUT_SINGLE	...	Molecule[Atoms: 52(143) Bonds: 51(142)]	-2677.1	429.5	-2904.0
			-2697.2	539.5	-2803.2
			52 total ›		
CRYSTALL_CUT_SINGLE2	...	Molecule[[illegible]]	-2677.1	429.5	-2904.0
			-2697.2	539.5	-2803.2
			52 total ›		
CTD	...	Molecule[[illegible]]	-343.1	1368.1	-2321.3
			-294.4	1237.1	-2373.7
			1736 total ›		
E	...	Molecule[[illegible]]	-1472.2	-1634.5	-968.3
			-1529.2	-1505.8	-932.4
			[illegible] total ›		
FLUID_CUT_SINGLE	...	Molecule[Atoms: 52(143) Bonds: 51(142)]	-2074.9	-20.9	2416.2
			-2077.2	112.4	2331.2
			52 total ›		
FLUID_CUT_SINGLE2	...	Molecule[[illegible]]	-2074.9	-20.9	2416.2
			-2077.2	112.4	2331.2
			52 total ›		
G	...	Molecule[Atoms: 15(24) Bonds: 16(25)]	221.3	-1381.3	-377.5
			211.6	-1255.6	-322.5
			15 total ›		
M	...	Molecule[[illegible]]	-941.7	240.6	-1990.7
			-1037.0	207.1	-1990.4
			[illegible] total ›		

Loading Data

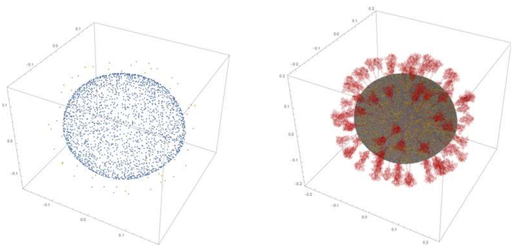


... 3D representations

Analyze the Distance between Structures



The whole sequence is deposited



Quaternion Rotation Approach & Volumetric Approach