PERSONAL INFORMATION

Fabio Cumbo, Ph.D.

Address 9500 Euclide Avenue – Cleveland – Ohio 44195 – USA
Mobile +1 (440) 360-9313
Email fabio.cumbo@gmail.com
Website https://cumbof.github.io/
ORCID https://orcid.org/0000-0003-2920-5838

Sex Male | Date of birth December 30, 1989 | Nationality Italian

WORKING EXPERIENCES

March 2022 – **ongoing**

Postdoctoral Researcher

Blankenberg Lab – Genomic Medicine Institute – Lerner Research Institute – Cleveland Clinic – Cleveland, Ohio, USA
https://lerner.ccf.org/gmi/blankenberg/

November 2018 – December 2021

Postdoctoral Researcher

Segata Lab – Computational Metagenomics Laboratory – Department of Cellular, Computational, and Integrative Biology (CIBIO) – University of Trento – Trento, Italy
http://segatalab.cibio.unitn.it/

- Project: MetaRefSGB: a scalable framework to organize genomes from metagenomes, as well as reference genomes and genomes from isolates sequencing, in addition to their annotations, into species-level genome bins (SGBs). It is based on a strategy of clustering genomes into known and previously unknown species. It permits several types of analyses, such as the collection of the functional potential of species microbial clades (known and unknown) and the study of the population genomics of microbial species.
  Goal: The main goal is the characterization and study of the microbial dark matter.
  Outcomes: The framework has been already applied on a set of over a million microbial genomes from over 50,000 metagenomes from multiple hosts and environments. This expanded the overall number of unknown SGBs and confirms that the framework can scale and support the integration of many new MAGs from the ever-increasing body of metagenomic samples available.
  Technologies: The framework has been developed in Python 3; it internally makes use of MASH for computing the genetic distances between genomes; average linkage clustering has been performed with the fastcluster Python package; the framework also provides a set of APIs developed over the Flask package to programmatically interact with the database; it also makes use of a subroutine of PhyloPhlAn called phylophlan metagenomic.
  Role: I'm the person in charge for the framework development and the identification of new set of MAGs and Reference Genomes from samples of multiple hosts and environments.
  Funds: EU-ERC Grant MetaPG-716575

- Project: Analysis of the microbiome composition of a cohort of stool samples collected from children affected by the Neuroblastoma disease with no treatment (PZ), under treatment (PZ_T), age-matched healthy controls (CTR), and the mothers of both patients (M_PZ) and healthy controls (M_CTR). Project in collaboration with the Laboratory of Experimental Oncology of the Italian National Public Children's Hospital and Research Institute "Giannina Gaslini" of Genoa, Italy.
  Goal: Investigate microbial composition alterations as possible cause of the neuroblastoma development in children.
  Outcomes: The differential abundance analysis highlighted significant differences in the microbial composition of CTR vs PZ and PZ_T, as well as in the composition of M_CTR vs M_PZ, suggesting that mother-to-infant transmitted strains play a key role in this context.
  Technologies: Analyses have been performed with R and statistical packages; the microbial communities' abundances have been estimated with MetaPhlAn.
  Role: I'm constantly interacting with the medical researchers of the Gaslini Hospital of Genoa in order to better define the statistical analysis. I'm also responsible for the production of the statistical reports.

Collaborations: Laboratory of Experimental Therapies in Oncology – Italian National Public Children's Hospital and Research Institute "Giannina Gaslini" – Genoa – Italy
Keywords: Bioinformatics – Metagenomics – Bacterial Species – Neuroblastoma Disease – ERC

April 2018 – December 2018

Professional Collaborator

ACTOR (Analytics, Control Technologies and Operations Research) S.R.L. – Rome – Italy
http://actorventure.com/

Project: Development of a technological platform to establish an early and non-invasive diagnosis of neurodegenerative diseases; The platform allows to automatically extract and standardize data from the IDA (Image and Data Archive) database powered by LONI (Laboratory of Neuro Imaging) funded by the NIH and NIBIB.
Goal: Develop a new software framework for the early-stage detection of the Alzheimer's disease.
Outcomes: Creation of an ontology in order to better understand how these data are organized and to create an easy access service to the data themselves; the software platform is currently in use in multiple specialized centres for the study of Alzheimer's disease all over the Lazio region in Italy.
Technologies: The ontology has been designed with Protégé, while the automatic extraction and standardization module of the framework has been developed in Java and R.
Role: I was involved in the framework development and in the design of the ontology.
Funds: MoDiag project (POR FESR Lazio 2014-2020, Life 2020)

Collaborations: (i) EBRI (European Brain Research Institute) – Rita Levi Montalcini Foundation – Rome – Italy and (ii) Institute for Systems Analysis and Computer Science "Antonio Ruberti" – National Research Council of Italy (CNR) – Rome – Italy
Keywords: Bioinformatics – Ontologies – Machine Learning – Alzheimer and Parkinson's disease – Diagnostics

April 2018 – September 2018

Ph.D. Fellow

Institut für Informatik of the Albert-Ludwigs-Universität Freiburg – Freiburg im Breisgau – Germany
http://bioinf.uni-freiburg.de/

Project: Development of bioinformatics tools for the Galaxy platform.
Goal: Improve Galaxy functionalities and integrate specific tools into the platform
Technologies: Software packages have been developed in Python 3.
Role: Software developer.

Collaborations: Wartik Laboratory – Department of Biochemistry and Molecular Biology – The Pennsylvania State University – University Park Campus – 16802 PA – Pennsylvania – USA
Keywords: Bioinformatics – Galaxy Project – Conda – Bioconda

March 2017 – March 2018

Research Lab Assistant

Wartik Laboratory – Department of Biochemistry and Molecular Biology – The Pennsylvania State University – University Park Campus – 16802 PA – Pennsylvania – USA
https://nekrut.github.io/lab_site/

Project: Development of new statistical analysis tools and algorithms, and contribute to the development of the Galaxy platform, Conda, and Bioconda; Development of a web tool to fast query massive sequence datasets with Sequence Bloom Trees.
Goal: Improve Galaxy functionalities and integrate specific tools into the platform.
Outcomes: Produced material has been presented at the Cold Spring Harbor Meeting on Genome Informatics in Cold Spring Harbor, NY, USA.
Technologies: Statistical analysis tools have been developed in R, while the web platform has been developed in PHP, jQuery, and JavaScript over the Flask web server on top of Python 3, with the HowDeSBT software running in background.
Role: Software developer.

Collaborations: (i) The Galaxy Team – Nekrutenko Lab and (ii) the Paul Medvedev's Lab
Keywords: Bioinformatics – Galaxy – Functional Data Analysis – Information Retrieval – Sequence Bloom Tree

February 2017 – November 2018

Professional Collaborator and Teaching Assistant

Department of Engineering – International Telematic University UNINETTUNO – Rome – Italy
https://www.uninettunouniversity.net/

Project: Development of a software to automatically extract, extend, and standardize clinical and genomic data from the Genomic Data Commons Portal by also integrating data from external sources. Part of the Data-Driven Genomic Computing (GeCo), focusing on tertiary analysis for genomic data integration.
Goal: Build open access resources that simplify the task of integrating heterogeneous genomic data.
Outcomes: Developed software and produced resources have been successfully published and are

currently available for the whole scientific bioinformatics community free of charge.
Technologies: Software has been completely developed in Java.
Role: Principal software developer.
Funds: EU-ERC Advanced Grant 693174

Working on the definition of a training plan for two new Master's degree courses in Software Engineering (Big Data branch) (i) "Introduction to Big Data" and (ii) "Big Data Analytics and Visualization".

Advisor for the realization of the following theses:
- "*Analysis and implementation of a web platform for the management and querying of genomic Big Data*" (Bachelor's Degree): Candidate "Lorenzo Di Nardo", Advisors "Prof. Emanuel Weitschek and Fabio Cumbo";
- "*The structure of the Bloom Filters for the management and querying of Big Data*" (Master's Degree): Candidate "Antonio Tranchida", Advisors "Prof. Emanuel Weitschek and Dr. Fabio Cumbo";
- "*Probabilistic data structures for the reference-free alignment of sequences*" (Master's Degree): Candidate "Federico Ferranti", Advisors "Prof. Emanuel Weitschek and Dr. Fabio Cumbo";
- "*Hyperdimensional Computing for Supervised Machine Learning*" (Master's Degree): Candidate "Simone Truglia", Advisors "Prof. Emanuel Weitschek and Dr. Fabio Cumbo".

Collaborations: (i) Department of Electronics, Information and Bioengineering of the Polytechnic University of Milan and (ii) Institute for Systems Analysis and Computer Science "Antonio Ruberti" – National Research Council of Italy
Keywords: Bioinformatics – ERC – GeCo – TCGA2BED – Apache Hadoop – Apache Spark – MapReduce – Machine Learning – D3.js – Data Visualization

September 2016 – March 2017

Professional Collaborator

Institute of Marine Engineering (ex "Marine Technology Research Institute – INSEAN") of the National Research Council of Italy – INM-CNR – Rome – Italy
http://inm.cnr.it/

Project: Development of a database containing data about military and merchant ships in which were used amiantus as a thermal insulator and data about officers and machinists affected by asbestos-correlated mesothelioma disease.
Goal: Map the presence of the amiantus on board of military and merchant ships.
Outcomes: The AMINAVI database is a powerful instrument used also by the Italian Navy.
Technologies: The database has been designed in SQL while the web platform required for consulting data has been developed in PHP, jQuery, and JavaScript with the support of the Boostrap framework for refining the UI.
Role: I was responsible for entire web platform and database design and development.

Keywords: Amiantus – Database – Mesothelioma – Military and Merchant Ships

October 2015 – November 2018

Ph.D. Student

Department of Engineering – University of Roma Tre – Rome – Italy
https://ingegneria.uniroma3.it/

Project: Analysis and development of new software technologies for the acquisition, storage, management, integration, and analysis of heterogeneous biomedical data.
Goal: The main goal consisted in being recognised as early stage researcher by the Doctoral School of Engineering (XXXI cycle) of the University of Roma Tre, Italy.
Outcomes: Eight papers published on international peer-review journals, in addition to seven international conferences and Ph.D. schools attended as auditor and speaker; degree has been conferred with Excellent score by the Department of Engineering, University of Roma Tre, Italy.
Technologies: All the proposed software solutions were developed in Python and Java.
Role: Ph.D. Student
Funds: Ph.D. has been funded by SYSBIO.IT – Center for Systems Biology, University of Milano-Bicocca, Milan, Italy

Advisor for the realization of the following theses:
- "*Analysis and development of a web service for the computation, visualization, and comparison of gene co-expression networks*" (Bachelor's Degree): Candidate "Dalila Rosati", Advisors "Prof. Maurizio Patrignani and Dr. Fabio Cumbo";
- "*TCGAinBED Web: Managing and querying genomic Big Data*" (Bachelor's Degree): Candidate "Luca Wissel", Advisors "Prof. Maurizio Patrignani and Dr. Fabio Cumbo".

Collaborations: (i) Institute for Systems Analysis and Computer Science "Antonio Ruberti" – National Research Council of Italy – Rome – Italy and (ii) SYSBIO.IT – Center for Systems Biology – University of Milano-Bicocca – Milan – Italy
Keywords: Bioinformatics – Computer Science – Data Modelling – Data Integration

September 2011 – February 2020

Research Assistant

Institute for Systems Analysis and Computer Science "Antonio Ruberti" – National Research Council of Italy – IASI-CNR – Rome – Italy
http://www.iasi.cnr.it/

Project: Design and development of algorithms for the computation of characteristic parameters in biological networks; Analysis of significant changes in the structure of protein complexes with the integration of temporal gene expression microarray data for the transgenic Mouse organism affected by Alzheimer's disease; Design and implementation of a software for the automatic extraction, storage, management, analysis, and querying of genomic and clinical data. Application of the software to The Cancer Genome Atlas. Part of the Data-Centric Genomic Computing (GenData 2020) project funded by the Ministry of Education, University, and Research of Italy under the PRIN program; Analysis and development of COSYS, web platform for the interoperability of software tools for the Systems Biology. COSYS allowed SYSBIO.IT to be part of ISBE (Infrastructure for Systems Biology Europe), a large-scale European research infrastructure of the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap.
Goal: Study the dynamics of the protein structures over time; Produce open access resources; Build scalable platforms for the analysis of complex metabolic models.
Outcomes: A huge number of outcomes have been reached over almost nine years of working experience at IASI-CNR: publishing my researches on peer-review international journals, presenting my research at international conferences, receiving my PhD, and expanding my knowledge by working for high-quality international research institutions and universities.
Technologies: Developed software tools are available on GitHub and they are all open access, as well as the produced resources.
Role: Research assistant

Collaborations: (i) EBRI (European Brain Research Institute) – Rita Levi Montalcini Foundation – Rome – Italy, (ii) Department of Electronics, Information and Bioengineering of the Polytechnic University of Milan, (iii) SYSBIO.IT – Center for Systems Biology – University of Milano-Bicocca – Milan – Italy, and (iv) ACTOR (Analytics, Control Technologies and Operations Research) S.R.L. – Rome – Italy
Keywords: Bioinformatics – PPI – Protein Complexes – CORUM – Cytoscape – Network Theory – Alzheimer's Disease – Microarray – Time Dynamics – Data Extraction – The Cancer Genome Atlas – Systems Biology – COSYS – ISBE

EDUCATION AND TRAINING

October 2015 – April 2019

Ph.D. in Computer Science and Automation Engineering

Department of Engineering – University of Roma Tre – Rome – Italy
Thesis: Data and models integration in biomedical information systems
Advisors: Prof. Maurizio Patrignani, Dr. Paola Bertolazzi
Score: Excellent

October 2012 – August 2014

Master of Science Degree in Software Engineering

Department of Engineering – University of Roma Tre – Rome – Italy
Thesis: Time dynamics of protein complexes in the AD11 transgenic mouse model for Alzheimer's disease like pathology
Advisors: Prof. Giuseppe Di Battista, Dr. Paola Bertolazzi
Score: 102/110

October 2008 – October 2012

Bachelor's Degree in Software Engineering

Department of Engineering – University of Roma Tre – Rome – Italy
Thesis: Selecting relevant nodes and structures in biological networks. BiNAT: a new plugin for Cytoscape
Advisors: Prof. Giuseppe Di Battista, Dr. Paola Bertolazzi
Score: 89/110

PERSONAL SKILLS

Mother tongue Italian

Other languages

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
English	C2	C2	C2	C2	C2

Levels: A1/2: Basic user - B1/2: Independent user - C1/2 Proficient user
Common European Framework of Reference for Languages

Communication skills Excellent communication skills gained through my experience in a wide variety of international working environments in addition to a consolidated experience as speaker in international conferences.

Organisational / managerial skills

- Scientific manager of the CINECA project entitled "*A novel brain-inspired hyperdimensional computing algorithm enabling efficient backward variable selection on massive datasets*" (HDCOMP) [grant number HP10CA0SGK]. The project allowed to get 50 thousand computational hours on the parallel computing resources of the CINECA, the largest Italian computing centre.
  Granted on February 2021
  Grant number HP10CA0SGK
- Advisor of 6 master and bachelor theses on Big Data, Machine Learning, and Bioinformatics.
- Participation to several national and international research projects on Machine Learning, Software Engineering, and Bioinformatics.

Job-related skills Good knowledge of molecular biology and statistics in addition to a good knowledge of bioinformatics and computational biology tools acquired during my professional career.

Computer skills

- Linux, Microsoft Windows and MacOS
- Bash, R, Python, and Java
- Office package

Proficiency on the technologies listed above

ADDITIONAL INFORMATION

Seminars

- Invited speaker at the YES@IASI (Young Experts Seminar at IASI-CNR) event with a talk about "*Fast querying of massive sequence datasets*" – March 9, 2018 at the Institute for Systems Analysis and Computer Science "Antonio Ruberti" – National Research Council of Italy

Conferences

- Speaker at the "Galaxy Community Conference 2022" with a talk and poster about "*Microbial strain characterization and subtyping of metagenome-assembled genomes with Sequence Bloom Trees in Galaxy*" – July 2022
- Speaker at the "11th International Workshop on Biological Knowledge Discovery from Big Data (BIOKDD'20)" (virtual conference due to the COVID-19 pandemic) in the context of the "31st International Conference on Database and Expert Systems Applications" (DEXA'20) originally planned for Comenius University, Bratislava, Slovakia, with a talk about "*An in-memory cognitive-based hyperdimensional approach to accurately classify DNA-Methylation data of cancer*" – September 2020
- Speaker at the poster session of the "ISMB/ECCB 2019 - International Society for Computational Biology" conference held in Basel, Switzerland with a poster about "*An Ontology to Organize Data on Alzheimer's Disease from International Databases to Support Integrated Analysis*" – July 2019
- Speaker at the "ISMB/ECCB 2019 – International Society for Computational Biology" conference held in Basel, Switzerland with a selected Long Talk about "*MetaRefSGB: a scalable framework to organize genomes from metagenomes and their annotations into species-level genome bins*" – July 2019
- Auditor at the "9th International Workshop on Biological Knowledge Discovery from Big Data" (BIOKDD'18) in the context of the "29th International Conference on Database and Expert Systems Applications" (DEXA'18) in Regensburg, Germany – September 2018
- Speaker at the poster session of the Cold Spring Harbor Meeting on Genome Informatics – Cold

Spring Harbor, NY, USA with a poster about "*GDCWebApp: filtering, extracting, and converting genomic and clinical data from the Genomic Data Commons portal*" – September 2017

- Speaker at the "13th International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics" at the University of Stirling, UK, with a talk about "*IRIS-TCGA: An Information Retrieval and Integration System for Cancer Genomic Data*" – September 2016
- Participant to the "Bioinformatics Italian Society 2015" (BITS 2015) with a poster about "*The Cancer Genome Atlas Data Querying Tool*" at the University of Milano-Bicocca – June 2015

Ph.D. Schools

- Auditor at the "3rd SYSBIO.IT School in Computational Biology" at the Institute for Systems Analysis and Computer Science "Antonio Ruberti" – National Research Council of Italy, Rome, Italy – May 2018
- Auditor at the "2nd School on Scientific Data Analytics and Visualization" at the CINECA center of Bologna, Italy – May 2016
- Scientific Committee member and speaker at the "1st SYSBIO.IT School on Computational Systems Biology" at the University of Milano-Bicocca with a talk titled "*Introduction to COSYS platform*" about a web platform able to manage, simulate, analyse, and visualize biochemical models – June 2016
- Auditor at the "2nd SyBSyM Lake Como School – Systems Biology and Systems Medicine: Toward a Precision Medicine" at Villa del Grumello, Como, Italy – September 2016

Workshops

- Participant to the course "*Parallel I/O and Management of Large Scientific Data*" organized by CINECA at the CINECA center of Rome, Italy – May 2015
- Participant to the workshop "*Hands On Big Data: Getting Started With NoSQL and Hadoop*" organized by the Codemotion at the Polo Didattico in Rome, Italy – April 2015

Publications

International peer reviewed journals
Conference proceedings
Book chapters

Publication records are available online on Google Scholar, Scopus, and on my ORCID profile:
https://scholar.google.com/citations?user=DJWJY7EAAAAJ&hl=en
https://www.scopus.com/authid/detail.uri?authorId=56373576900
https://orcid.org/0000-0003-2920-5838

Awards

- GCC2022 fellowship – May 2022
- CINECA Type C Grant HP10CA0SGK for project HDCOMP – February 2021

In compliance with the Legislative Decree n. 13 GDPR 679/16 and the Italian Legislative Decree n. 196 dated June 6, 2003, I hereby authorize the recipient of this document to use and process my personal details for the recruiting and selecting staff purposes and I confirm to be informed about my rights according to the art. 7 of the above mentioned Decree.

Date
August 1, 2022

_____
Fabio Cumbo