PERSONAL INFORMATION

Fabio Cumbo, Ph.D.

Position	Postdoctoral Research Fellow
Affiliation	Center for Computational Life Sciences, Lerner Research Institute, Cleveland Clinic Foundation
Address	9500 Euclid Avenue, NE5, Cleveland, OH 44195, USA
Mobile	+1 (440) 360-9313
Email	cumbof@ccf.org | 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
Work Authorization EU Citizenship, US Research Scholar Visa

WORKING EXPERIENCES

March 2022 – **ongoing**

Postdoctoral Research Fellow

Blankenberg Lab, Center for Computational Life Science, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, USA
https://lerner.ccf.org/gmi/blankenberg

- Project: MetaSBT: a scalable framework for automatically organizing and indexing microbial genomes and accurately characterizing metagenome-assembled genomes with Sequence Bloom Trees.
  Goal: Characterization of the microbial dark matter and study of its relationship with host health and environmental factors.
  Outcomes: Framework is currently under development.
  Technologies: Python 3.9, C++
  Role: Scientific Software Developer
- Project: Vector-symbolic architectures – Implementation of a feature selection technique based on the stepwise regression strategy built on top of *chopin2*, a domain-agnostic brain-inspired supervised-learning classification model built according to the hyperdimensional computing paradigm; Implementation of *hdlib*, a Python library for building vector-symbolic architectures.
  Goal: Implementation of the first of its kind stepwise regression strategy for selecting features with vector-symbolic architectures.
  Outcomes: The tool has been tested on public datasets with microbial profiles of metagenomic stool samples from shotgun sequencing collected from individuals affected by colorectal cancer in a case/control scenario. Results are comparable with the state-of-the-art feature selection methods. It is able to scale on datasets with massive amounts of features with commodity hardware.
  Technologies: Python 3.9, CUDA, Apache Spark
  Role: Scientific Software Developer
- Project: Quantum Computing – The Cleveland Clinic Foundation has signed a 10-year partnership with IBM aimed at accelerating discovery in healthcare and life sciences, by developing the first private sector onsite, IBM-managed quantum computer in the United States, located on Cleveland Clinic's main campus.
  Technologies: Python 3.9, Qiskit
  Role: Scientific Software Developer, Member of the Quantum Computing Research Group at the Lerner Research Institute of the Cleveland Clinic

Collaborators: IBM Quantum
Keywords: Metagenomics, Vector-Symbolic Architectures, Machine Learning, Quantum Computing

November 2018 – December 2021

Postdoctoral Research Fellow

Segata Lab, Computational Metagenomics Laboratory, Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Povo, TN, Italy
https://segatalab.github.io

- Project: MetaRefSGB: a scalable framework to organize genomes from metagenomes (MAGs), as well as reference genomes and genomes from isolate 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 microbial

clades (known and unknown) and the study of the population genomics of microbial species.
Goals: The main goal is the characterization and study of the microbial dark matter.
Outcomes: The framework has been applied on a set of over a million bacterial genomes from over 50,000 metagenomes from multiple hosts and environments. This expanded the overall number of unknown SGBs and confirmed 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 3.9; 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 its databases; it also makes use of a subroutine of PhyloPhlAn, called *phylophlan_metagenomic*.
Role: Scientific Software Developer – Framework developer and databases curator
Funds: EU-ERC Grant MetaPG-716575 to Prof. Nicola Segata

- Project: Analysis of the microbiome composition of a cohort of stool samples collected from children affected by the Neuroblastoma disease with no treatment, under treatment, age-matched healthy controls, and the mothers of both patients and healthy controls children. 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.
  Goals: Investigate microbial composition alterations as a possible cause of the Neuroblastoma development in children.
  Technologies: Analyses have been performed with R and statistical packages; the quantitative profiling of metagenomic samples has been performed with MetaPhlAn.
  Role: Responsible for (i) interacting with medical researchers of the Gaslini Hospital of Genoa and (ii) performing the statistical analyses.

Collaborations: Laboratory of Experimental Therapies in Oncologies, Italian National Public Children's Hospital and Research institute "Giannina Gaslini", Genoa, Italy
Keywords: Bioinformatics, Metagenomics, Neuroblastoma

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 framework for the early-stage detection of Alzheimer's disease.
Outcomes: Creation of an ontology for organizing psychometric assessment data; the software platform is currently in use in multiple specialized centers 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: Scientific Software Developer and Ontology Curator
Funds: MoDiag project (POR FESR Lazio 2014-2020, Life 2020)

Collaborations: EBRI (European Brain Research Institute), Rita Levi Montalcini Foundation, and the Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy
Keywords: Bioinformatics, Ontology, Machine Learning, Alzheimer'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-frieburg.de

Project: Development of bioinformatics tools for the Galaxy platform.
Goal: Improve Galaxy functionalities and integrate specific tools into the platform.
Technologies: Python 3
Role: Scientific Software Developer

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

March 2017 – March 2018

Ph.D. Fellow and Research Lab Assistant

Wartik Laboratory, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park Campus, PA, 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.
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, making use of the HowDeSBT framework.
Role: Scientific Software Developer

Collaborations: The Galaxy Team, Nekrutenko Lab, and Medvedev Lab at the Department of Biochemistry and Molecular Biology, The Pennsylvania State University, PA, USA
Keywords: Bioinformatics, Galaxy Project, Functional Data Analysis, Information Retrieval, Sequence Bloom Tree

February 2017 – November 2018

Professional Collaborator and Teaching Assistant

Department of Engineering, University of 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.
Technologies: Java
Role: Scientific Software Developer
Funds: EU-ERC Advanced Grant 693174

As a Teaching Assistant, I also worked on the definition of two courses for graduate students of the Master of Science Degree in Computer Engineering, i.e., "*Introduction to Big Data*" and "*Big Data Analytics and Visualization*". I have also served as co-advisor for 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 Dr. Fabio Cumbo;
- *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: Department of Electronics, Information and Bioengineering of the Polytechnic University of Milan and the Institute of Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy
Keywords: Bioinformatics, GeCo, TCGA2BED, Apache Hadoop, Apache Spark, MapReduce, Machine Learning, D3.js, Data Visualization

September 2016 – March 2017

Professional Collaborator

Institute of Marine Engineering, National Research Council of Italy, Rome, Italy
http://inm.cnr.it

Project: Development of a database for organizing data about Italian military and merchant ships in which amiantus was used as a thermal insulator and data about officers and machinists affected by asbestos-related mesothelioma disease.
Goal: Map the presence of amiantus on board of Italian military and merchant ships.
Outcomes: The database (AMINAVI) is used 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 Bootstrap framework for rendering the UI.
Role: Scientific Software Developer

Keywords: Amiantus, Mesothelioma, Italian 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: Ph.D. in Computer Science and Automation Engineering awarded by the Doctoral School of Engineering (XXXI cycle) of the University of Roma Tre, Italy.
Role: PhD. Student, Scientific Software Developer
Funds: Ph.D. has been funded by SYSBIO.IT – Center for Systems Biology, University of Milano-Bicocca, Milan, Italy

As a Ph.D. Student, I have been involved as a co-advisor for 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.

Collaborators: Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy and 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 and Associate

Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, 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 data from microarray for a 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, a web platform for the interoperability of software tools for Systems Biology.
Goal: Study the dynamics of the protein structures over time; produce open access resources; build scalable platforms for the analysis of complex metabolic networks.
Role: Research Assistant, Research Associate, Scientific Software Developer

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

EDUCATION AND TRAINING

October 2015 – April 2019

Ph.D. in Computer Science and Automation Engineering

Doctoral School of 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

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

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

PERSONAL SKILLS

Languages English, Italian

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.

Organizational / managerial skills

- Scientific manager of the CINECA project titled "*A novel brain-inspired hyperdimensional computing algorithm enabling efficient backward variable selection on massive datasets*" (HDCOMP – grant number HP10CA0SGK). I have been awarded with 50 thousand computational hours on the parallel computing resources of CINECA, the largest Italian computing center.
  Granted on February 2021
  Grant number HP10CA0SGK
- Advisor of 6 master and bachelor theses on Big Data, Machine Learning, and Bioinformatics.
- Participation in several national and international research projects about 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
- Software packaging with Docker and Conda
- Software versioning with git
- Office package

Proficiency with 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, Rome, Italy

Conferences

- Speaker at the "Galaxy Community Conference 2023" with a talk and poster about "*Investigating the known and unknown microbial composition of metagenomic samples made easy with MetaSBT in Galaxy*" held in Brisbane, Queensland, Australia – July 2023
- Speaker at the poster session of the "Midwest Microbiome Symposium 2023" held at The Ohio State University in Columbus, OH, USA, with a poster titled "*MetaSBT: a scalable reference-based phylogeny-aware framework for characterizing known and yet-to-be-named microbial species with Sequence Bloom Trees*" – May 2023
- Speaker at the poster session of the "9th Annual Genetics Education Symposium" organized by the Genomic Medicine Institute of the Cleveland Clinic and the Center for Personalized Genetic Healthcare in Cleveland, OH, USA, with a poster titled "*Feature selection with vector-symbolic architecture: a case study on microbial profiles of shotgun metagenomic samples of colorectal cancer*" – September 2022
- 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*" held in Minneapolis, MN, USA – July 2022
- Speaker at the "11th International Workshop on Biological Knowledge Discovery from Big Data (BIOKDD'20)" (virtual conference due to COVID-19 pandemic) in the context of the "31st International Conference on Database and Expert Systems Applications" (DEXA'20), with a talk about "*An in-memory cognitive-based hyperdimensional computing 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 in Basel, Switzerland with a poster about "*An Ontology to Organize Data on Alzheimer's Disease from International Databases to Support Integrative Analysis*" – July 2019
- Speaker at the "ISMB/ECCB 2019 - International Society for Computational Biology" conference 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 at 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
- Attendee at 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 Analysis 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, Milan, Italy, with a talk about "*Introduction to COSYS platform*" – 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

- Attendee at the "HPC Workshop: Machine Learning and Big Data" organized by ACCESS at the Ohio Supercomputing Center in Columbus, OH, USA – March 2023
- Attendee at the "Microbiome Workshop" organized by the Case Comprehensive Cancer Center of the Case Western Reserve University in Cleveland, OH, USA – February 2023
- Attendee at the "Workshop on Drug Discovery: Small Molecules and Biologics" organized by the Cleveland Clinic and IBM Research Discovery Accelerator at the Lerner Research Institute of the Cleveland Clinic – January 2023
- Attendee at the "Parallel I/O and Management of Large Scientific Data" workshop organized by CINECA at the CINECA center of Rome, Italy – May 2015
- Attendee at the "Hands On Big Data: Getting Started with NoSQL and Hadoop" workshop organized by the Codemotion at the Polo Didattico in Rome, Italy – April 2015

Publications

Publication records, published in international peer-reviewed journals, conference proceedings, and book chapters, are available online on Google Scholar, Scopus, and ORCID:
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

Associations and Organizations

- Member of the National Postdoctoral Association since October 2022 (membership #70567958)
- Member of the Italian Scientists and Scholars in North America Foundation since July 2023

Editorial Boards

- Associate Editor of the Soft Computing journal (Springer Nature) since May 2021
  https://www.springer.com/journal/500/editors
- Invited Guest Editor of the Research Topic "The Role of the Microbiome in Head and Neck Cancers" for the Frontiers in Oncology journal since September 2023

Awards

- Certificate of Quantum Excellence issued by IBM – August 2022
- GCC2022 Fellowship – May 2022
- CINECA Type C Grant, project HDCOMP #HP10CA0SGK – 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 aforementioned Decree.

Date
September 13th, 2023

Fabio Cumbo