curriculum vitae

PERSONAL INFORMATION

Surname		**Alfonsi**
Name		**Tommaso**
Address		**via E. Pagani 86, 21048 Solbiate Arno (VA), Italy**
Telephone		**3497561639**
E-mail		**tommaso.alfonsi@outlook.com**

Nationality		Italian

Date of birth		29 / 12 / 1993

Education and training

		Bachelor's degree in computer science
• Date (from – to)		2012 - 2016
• Name and type of organisation providing education and training		Politecnico di Milano
Duration of the program of study		3 years
• Principal subjects/occupational skills covered		Computer Science
• Title of qualification awarded		Dottore in ingegneria informatica
Final mark obtained		87 / 110

		Master's degree in computer science ((LM-32) Ingegneria informatica)
• Date (from – to)		2017 - 2020
• Name and type of organisation providing education and training		Politecnico di Milano
Duration of the program of study		2 years
• Principal subjects/occupational skills covered		Computer Science
• Title of qualification awarded		Dottore magistrale in ingegneria informatica
Final mark obtained		102 / 110
Graduation thesis		Title: *"Integration of DNA variation data into a GDM repository and API development for identification of genomic populations"*

		The thesis address both a data-integration and a data-analysis challenge in the context of human genomics. Full thesis document at https://1drv.ms/b/s!Agket-TecjHogZ4lVedtlQ13fZJGQA?e=VD6U1n

		Ph.D. in Information Technology
• Date (from – to)		2020 - current (the Ph.D. is ending in January 2024)
• Name and type of organisation providing education and training		Politecnico di Milano
Duration of the program of study		3 years + 3 months extension for COVID-19 pandemic
• Principal subjects/occupational skills covered		Information Technology
• Title of qualification awarded		Dottore di ricerca in Ingegneria dell'Informazione
Thesis title		*Methods and tools for data integration and knowledge discovery in viral genomics*

Journal articles		6 published journal articles + 1 under review

2021

1	Author(s) and title		*Canakoglu, A.; Pinoli, P.; Bernasconi, A.;* ***Alfonsi, T.****; Melidis, D. P.; Ceri, S.* *"ViruSurf: An Integrated Database to Investigate Viral Sequences"*
	Publication place		*Nucleic Acids Research*, 2021, https://doi.org/10.1093/nar/gkaa846, *(IF 2022 19.509)*
	Personal Contribution		I contributed to the design and implementation of the data integration pipeline of viral sequences and to first manuscript draft

2	Author(s) and title		*Bernasconi, A.; Cilibrasi, L.; Al Khalaf, R.;* ***Alfonsi, T.****; Ceri, S.; Pinoli, P.; Canakoglu, A.* *"EpiSurf: Metadata-Driven Search Server for Analyzing Amino Acid Changes within Epitopes of SARS- CoV-2 and Other Viral Species"*
	Publication place		*Database, 2021,* https://doi.org/10.1093/database/baab059, *(IF 2022 5.8)*
	Personal Contribution		I contributed to design the data model for epitope annotations, to design and implement the corresponding data integration pipeline and to write the first manuscript draft

3	Author(s) and title		*Bernasconi, A.; Gulino, A.;* ***Alfonsi, T.****; Canakoglu, A.; Pinoli, P.; Sandionigi, A.; Ceri, S.* *"VirusViz: comparative analysis and effective visualization of viral nucleotide and amino acid variants"*
	Publication place		*Nucleic Acids Research, 2021,* https://doi.org/10.1093/nar/gkab478, *(IF 2022 19.509)*
	Personal Contribution		I contributed to design and implement the software pipeline responsible for preparing user-submitted data, and for generating the output reports

2022

4	Author(s) and title		***Alfonsi T.****; Bernasconi A.; Canakoglu A.; Masseroli M.* *"Genomic data integration and user-defined sample-set extraction for population variant analysis"*
	Publication place		*BMC Bioinformatics, 2022,* https://doi.org/10.1186/s12859-022-04927-0, *(IF 2022 3.071)*
	Personal Contribution		I contributed to design the data model, design and implement the data integration pipeline and API, prepare the use cases, write the documentation and the first manuscript draft.

5	Author(s) and title		***Alfonsi, T.****; Pinoli, P.; Canakoglu, A.* *"High Performance Integration Pipeline for Viral and Epitope Sequences"*
	Publication place		*BioTech 2022,* https://doi.org/10.3390/biotech11010007, *(IF 2022 3.477)*
	Personal Contribution		I contributed to writing the first manuscript draft, design and implement the data integration software

6	Author(s) and title		***Alfonsi, T.**; Al Khalaf, R.; Ceri, S.; Bernasconi, A.* *"CoV2K Model, a Comprehensive Representation of SARS-CoV-2 Knowledge and Data Interplay"*
	Publication place		*Scientific Data, 2022, https://doi.org/10.1038/s41597-022-01348-9, (IF 2022 9.554)*
	Personal Contribution		I contributed to design and implement the knowledge base, the underlying data model, the software responsible for data ingestion and transformation and the CoV2K API.

2023

7	Author(s) and title		***Alfonsi, T.**; Bernasconi A.; Chiara M.; Ceri S.* *"Data-driven recombination detection in viral genomes".*
	Publication place		*2023, manuscript under review at Nature Communications, Pre-print at https://doi.org/10.1101/2023.06.05.543733*
	Personal Contribution		I contributed to design and implement the core method and the software necessary to prepare the input data, run the experiments and analyse the results. The method is already available to the scientific community in form of a Zenodo archive containing the software, demo and the supplementary materials of the manuscript. https://zenodo.org/doi/10.5281/zenodo.8123832

Conference papers		2 conference papers

2021

8	Author(s) and title		*Al Khalaf, R.; **Alfonsi, T.**; Ceri, S.; Bernasconi, A.* *"CoV2K: A Knowledge Base of SARS-CoV-2 Variant Impacts"*
	Publication place		*Research Challenges in Information Science: 15th International Conference (RCIS), 2021, https://hdl.handle.net/11311/1172289*
	Personal Contribution		I contributed to design the data model representing the novel data and knowledge entities.

2022

9	Author(s) and title		***Alfonsi, T.**; Bellomarini, L.; Bernasconi, A.; Ceri, S.* *"Expressing Biological Problems with Logical Reasoning Languages"*
	Publication place		*Rules and Reasoning: 6th International Conference (RuleML+RR), 2022, https://hdl.handle.net/11311/1221640*
	Personal Contribution		I contributed to design and implement the software, prepare the use cases and write the first manuscript draft.

Research activity

2020, Research Assistant, Politecnico di Milano		The main goal of this collaboration was to support the development of an extension for the GMQL framework responsible for the automatic alignment and backup of datasets between the clusters of the GMQL repository. I developed the software that replicated data and metadata as necessary between local file systems and Hadoop FS in order to reflect updates and distribute operations efficiently between clusters of machines.

2020 - 2023, Ph.D. in Information Technolohy., Politecnico di Milano		The goal of my Ph.D. research is to analyse and integrate heterogeneous genomic datasets and knowledge resources to answer complex biological questions. I collaborated with other researchers of the Genomic Computing group of Politecnico di Milano on building a large integrated database of viral sequences and genomic annotations. I designed and implemented a distributed ETL pipeline (Extract, Transform, Load) capable of building a timely, uniform and complete representation of all the viral sequences available in three major databanks (GISAID, GenBank, COG-UK), enriched with annotations from the Immune Epitope Database (IEDB) [5]. The resulting database is a repository of viral data and metadata about many viruses, including Dengue, Ebola, MERS, SARS and SARS-CoV-2. This work was

		propaedeutic to the realisation of the web applications ViruSurf [1], EpiSurf [2] and VirusViz [3] and is documented in my publications between 2021-2022. During the year 2022, I focused on the integration of data and knowledge about SARS-CoV-2 to foster knowledge-discovery applications. I designed a knowledge graph (CoV2K) [6,8] about SARS-CoV-2 variants, genomic mutations, scientific studies, clinical effects and genomic annotations. Notably, the knowledge entities are interlinked also with the aforementioned database of viral sequences, making It possible to quickly find evidence or examples of any information present in the knowledge base. I implemented the data model and the ETL pipeline providing information to the knowledge base. CoV2K is usable from any interested user or program thanks to an API which I developed. The relevance of CoV2K is emphasized by the graph-based representation of viral knowledge, that enhances the effectiveness of automated reasoning in knowledge discovery applications. This was demonstrated by formulating four use cases in the Knowledge Reasoning and Representation (KRR) language "Vadalog" that were solved by the artificial reasoner [9]. In 2022, I also published a research work on the integration and analysis of human genome variants which follows from my master graduation thesis on human genomics. I contributed to the GMQL project (http://gmql.eu) by extending its genomic data repository and developed an API for simplifying the exploration of huge genomic datasets obtaining descriptive measures on user-defined populations [4]. This work is reflected in my publications from the year 2022. The knowledge developed about genomic data and viruses was fundamental to address the data-analysis problem of finding viral recombinant sequences. During the COVID-19 pandemic, experts were finding novel recombinants mostly by manual observation. As, the process is long and prone to errors, I contributed to design a novel method (RecombinHunt) [7] for spotting recombination in a reliable and accurate manner. I prepared the data, implemented and tested the method on both real and simulated datasets for the SARS-CoV-2 and Monkeypox viruses showing excellent results compared to its competitors. This work is described in a manuscript that is currently under review at Nature Communications. Currently, I am testing the applicability of RecombinHunt on segmented-genome types of viruses, especially the Influenza virus. The Influenza virus is capable of both intra-segment recombination and inter-segment recombination (reassortment). An adaptation of RecombinHunt could help in the identification of recombinant and reassortant viral sequences in the context of the Influenza and similar viruses. A potential method using aggregated genome metrics and unsupervised learning algorithms is under investigation.

Studies abroad

• Date (from – to)		**07/08/2022 – 28/08/2022**
• Name and address of firm/university		University of Oxford
• Main activities		I attended the course "Artificial Intelligence and Machine Learning summer programme" (7.5 CFU, 37.5 hours of lectures, 3 weeks long in-person educational programme with weekly written and oral evaluations)

Conferences/Seminars/ Workshops attended

05/2021		Attended the conference Research Challenges in Information Science (RCIS)
01/2022		Poster with presentation at the 13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS)"A unique approach to SARS-CoV-2 data and knowledge ingestion, integration and querying"
09/2022		Full paper oral presentation at the International Joint Conference on Rules and Reasoning (RuleML+RR)
09/2022		Attended the Genome Informatics Hybrid Conference 2022 at Wellcome Sanger Institute

11/2022		Attended the Viral Genomics Conference 2022 at Wellcome Sanger Institute
11/2022		Conference talk at the Doctoral Achievements for Colloquia Doctoralia 2022 “From Human Genomics to COVID-19 Virus Sequences: Re-Thinking a Data-Driven PhD Research on the Fly”
07/2023		Program Committee Member for the 9th IEEE International Conference on Big Data Computing Service and Machine Learning Applications (IEEE BigDataService 2023).

Teaching activity

a.y. 22/23 and 23/24, Teaching assistant, Politecnico di Milano		20h teaching assistant (exercises classes) for the course “Programming” (Prof. Anna Bernasconi). The course aims to teach the basics of Python programming for bioinformatics analysis.

a.y. 22/23 and 23/24, Laboratory assistant, Politecnico di Milano		15h laboratory assistant for the course “Informatica B” (Prof. Marco Masseroli). The course aims to introduce the students to the principles of software programming using the languages C and Matlab.

Work experience, stages

2016 - 2017, IT consulting for Android app development, MD Sergio Dall’Acqua		I was asked to design and develop an Android application which allows to acquire printed documents, collect, organize them and exchange them with predefined contacts at regular intervals without further intervention. Its purpose was to ease the exchange of commercial documents for people running a small company.

Personal skills and competences

Languages

Italian		Mother tongue
English		Excellent level **(**TOEIC Certification of English - score 870/990 – 27 Jan 2017)
French		Sufficient level

Social skills and competences		During my academic career I faced many situations in which a clear communication and efficient coordination skills are mandatory to obtain good results. Such occasions include the teaching activities, participation to conferences, talks and research projects carried on in collaboration with other people, including foreigners with largely different backgrounds and cultures which are now friends and colleagues. Throughout my Ph.D, I had the chance of further improve my communication skills while working with contributors of acknowledged competence and heterogeneous backgrounds from various institutes; examples of these include the Department of Agricultural and Animal Sciences from the University of Florida, the L3S group of the Leibniz University of Hannover, the Molecular Biology group from Università degli Studi di Milano, the Adam Group at Wellcome Sanger Institute, and the Applied Research Team from Banca d’Italia.

Organisational skills and competences		Several group projects during my academic studies gave me the opportunity to develop an attitude toward working in a collaborative environment, distributing resources fairly and coordinating with other members to reach complex and long-term goals. By effectively applying these skills, I was able to foster a proficient work environment with my research peers during my Ph.D., ultimately leading to significant outcomes and publications.

Technical skills and competences		▪ Excellent skills in data analysis using a variety of statistical and data-exploration methods to extract insights from datasets. Experience in the detection and analysis of trends, patterns, recurrent events, time-series and anomalies. ▪ Knowledge of the principal machine learning algorithms and their theoretical foundations to develop linear and non-linear (neural networks, CNNs) models for regression, classification, clustering and feature selection/extraction.

		▪ Excellent skills in data management through relational and noSQL systems, in the creation of efficient queries, query optimizations, database-software integration and administrative operations (create, clone, backup, transfer, restore a DB). ▪ Proficient software developer of APIs (Application Programming Interfaces) and ETL (Extract Transform Load) pipelines targeted to the automatic download from a variety of sources (like FTP server, API, Amazon S3, remote DB, HTTP using web-scraping), synchronization, backup and transformation of remote datasets. ▪ Excellent software development skills using several programming languages (Scala, Python, PHP, Java, JavaScript, Node, C, R), data query languages (PostgreSQL, MySQL, XQuery, XPath, OWL, SPARQL) and descriptive languages (HTML, CSS, XML, YAML for OpenAPI3 standard). Knowledge of the Python software libraries for data analysis (Pandas, Numpy, Matplotlib, Plotly, Scikit-learn, PyTorch, TensorFlow). ▪ Knowledge of the principal bioinformatic databanks and the tools related to genome sequences, taxonomy and annotations in the context of the human and viral genomes. ▪ Knowledge of the biological background about the reproduction, evolution and the history of viruses. ▪ Knowledge and experience in the use of the common statistical models, multivariate analysis, probability estimation methods. ▪ Knowledge of the data modelling theory used for achieving integration of genomic data repositories (Genomic Data Model, Genomic Conceptual Model, Viral Conceptual Model) and of several bioinformatic data encodings. ▪ Knowledge of the cloud computing paradigms and resources (Apache Spark, Flink, SciDB, Hadoop FS, Amazon EC2) and the architectural features supporting the execution of the GMQL engine. ▪ Experience in development of Android native applications, back-end servers and web sites. ▪ Knowledge of development frameworks Bootstrap, AJAX, SQLAlchemy, Flask and Connexion. ▪ Knowledge of the Git version control system. ▪ Knowledge of standard office applications (Word, Excel, PowerPoint) and LaTeX. ▪ Knowledge of the operative systems Microsoft Windows, Linux and MacOS.

Artistic skills and competences *Music, writing, drawing etc.*		I practiced playing classic guitar for three years before going to high school. Later I taught guitar for some months to the young people of my city as a recreational activity.

Other skills and competences *Competences not mentioned above.*		I consider myself a curious and dynamic person who likes to get involved in many activities, especially outdoor. I like doing sport when I want to take a rest from study or work. Indeed, I practised many disciplines, mainly archery (three years) and swimming (nine years), but occasionally also scuba diving, mountain biking, downhill, skiing. I love enjoying natural environments, challenge myself and learn new things.

Additional information		I am manager of the cloud infrastructure of the Genomic Computing group (GeCo) at CINECA. Also, I'm maintaining the distributed computing engine that supports the GenoMetric Query Language (GMQL) interface.