

ALLEGATO B**UNIVERSITÀ DEGLI STUDI DI MILANO**

selezione pubblica per n. 1 posto/i di Ricercatore a tempo determinato ai sensi dell'art.24, comma 3, lettera b) della Legge 240/2010 per il settore concorsuale 02/A2 - Fisica Teorica delle Interazioni Fondamentali,

settore scientifico-disciplinare FIS/02 - Fisica Teorica Modelli e Metodi Matematici

presso il Dipartimento di FISICA "ALDO PONTREMOLI",

(avviso bando pubblicato sulla G.U. n. 53 del 02 marzo 2021) Codice concorso 4541

**Edoardo Sarti
CURRICULUM VITAE****15/03/2021****INFORMAZIONI PERSONALI (NON INSERIRE INDIRIZZO PRIVATO E TELEFONO FISSO O CELLULARE)**

COGNOME	SARTI
NOME	EDOARDO
DATA DI NASCITA	30/04/1987

PROFESSIONAL EXPERIENCE

From	To	Institution	Position
01/10/2020	Present	Sorbonne Université ANR fellowship, 10 months Department of Computer Science Researcher at LCQB With Alessandra Carbone	Postdoctoral fellow Computational biology <i>Statistical methods for the structural and coevolutional study of the SARS-CoV-2 proteome</i>
01/09/2019	31/08/2020	Sorbonne Université ATER contract, 1 year Department of Biology Researcher at LCQB With Martin Weigt	Attaché temporaire d'enseignement et de recherche Mathematics, Statistics, Computer science <i>Statistical models for protein-protein interaction</i>
15/06/2018	31/08/2019	Sorbonne Université CalSimLab fellowship 12+6 months Department of Computer Science Researcher at LCQB With Martin Weigt	Postdoctoral fellow Statistical mechanics and Computational biologist <i>Statistical models for protein-protein interaction</i>
30/11/2015	31/05/2018	National Institutes of Health National Institute of Neurological Disorders and Stroke Bethesda (MD) USA Postdoctoral fellowship, 2.5 years With Lucy Forrest	Postdoctoral fellow Computational molecular biology <i>Sequence alignment and structural analysis and classification of membrane proteins</i>
01/02/2014	31/05/2014	Rice University Houston (TX) USA Visiting student, 3 months With Cecilia Clementi	Visiting student (PhD student) Computational biology <i>Coarse-grained force fields for molecular dynamics</i>
01/02/2013	30/06/2013	Università Cattolica del Sacro Cuore Milano, Italy Department of Medicine Lecturer contract, 4 months Coordinator : Alberto Granato	Lecturer <i>Foundations of Physics</i>

01/10/2011	12/10/2015	SISSA Trieste, Italy Physics and Chemistry of Biological Systems Doctoral fellowship, 4 years Supervisor : Alessandro Laio	PhD student <i>Physical and statistical methods for assessing the structure of proteins and protein complexes</i>
------------	------------	---	--

DEGREES

12/10/2015 **PhD in Physics and Chemistry of Biological Systems**

Title : *Assessing the structure of proteins and protein complexes through physical and statistical approaches*

28/11/2011 **Laurea Magistrale in Physics**

Mark : 110 / 110 con lode

SUPERVISION

2021 Supervision for the Master-1 stage of Mouna Ouattara
Domain annotation of the SARS-CoV-2 structural proteins
Sorbonne Université
Master in Bioinformatics and Modelization (BIM), Department of Computer Science

SARS-CoV-2's most trusted evolutionary theory sees its four structural proteins evolve independently of the rest of the genome, but their origins remain a mystery. By modifying the MetaCLADE sequence annotation algorithm, Mouna Ouattara is studying the homology relationships between the structural proteins of SARS-CoV-2 and those of other enveloped viruses.

2020 Co-supervision for the Master-1 stage of Rouquaya Mouss and Abderrhaman ElMenosum
With Maureen Muscat
Predicting the structure of membrane proteins with DCA
Sorbonne Université
Master in Bioinformatics and Modelization (BIM), Department of Computer Science

Direct coupling analysis (DCA) uses a mutual information approach to recover the co-evolving signals between residues of a protein, in order to predict its structure. However, this algorithm is less efficient on membrane proteins because of the statistical properties of their multiple sequence alignments. The students studied the problem and looked for a more efficient algorithm.

EDITORIAL WORK

2020-2021 Editor of a special issue of Frontiers in Molecular Biosciences
Molecular evolution: you learn from your mistakes
www.frontiersin.org/research-topics/12506/molecular-evolution-you-learn-from-your-mistakes
With Annalisa Pastore, Gian Gaetano Tartaglia and Marco Fantini

OTHER PERSONAL RESPONSIBILITIES

2020-2021 Referee for Bioinformatics, PLOS Computational Biology and PLOS ONE.

2015-2021 Maintenance of the AlignMe webserver with René Staritzbichler
www.bioinfo.mpg.de/AlignMe/

2014 Coordination with Ivan Gladich of a project for Alessandro Laio's research group. Gladich and I successfully applied for an ISCRA-B international call for 5 million CPU hours, and within the next 12 months we were responsible for the management and allocation of compute resources within the group.

2012-2014 Tutoring for three editions of a SISSA computational physics summer school for undergraduate students

TEACHING ACTIVITIES

FRANCE

2020-2021 **Data Science and Artificial Intelligence**, 18,5h TD, Sorbonne Université, Paris, France.

Coordinator: Christophe Marsala

Algorithms on Trees and Graphs in Bioinformatics, 20h TD, Sorbonne Université, Paris, France.

Coordinator: Alessandra Carbone

Statistics in Bioinformatics and Algorithms on Sequences, 24h TD, Sorbonne Université, Paris, France.

Coordinator: Martin Weigt, Juliana Bernardes

2019-2020 **Elements of Programming I**, 38,5h TD, Sorbonne Université, Paris, France.

Coordinator: Romain Demangeon

Mathematics and Statistics for Biology I, 40h TD, Sorbonne Université, Paris, France.

Coordinator: Stéphane Genet, Céline Ellien

Essential bioinformatics, 2h CM, 17h TP, Sorbonne Université, Paris, France.

Coordinator: Stéphane Le Crom

Mathematics and Statistics for Biology II, 36h TD, Sorbonne Université, Paris, France.

Coordinator: Dominique Lamy, Lorette Noiret

Modelization, Algorithmics and Programming for Biology, 8h CM, 16h TP, Sorbonne Université, Paris, France.

Coordinator: Mathilde Carpentier

2018-2019 **Multidisciplinary modelization**, 15h TD, Sorbonne Université, Paris, France.

Coordinator: Frédérique Charles

Statistics in Bioinformatics and Algorithms on Sequences, 24h TD, Sorbonne Université, Paris, France.

Coordinator: Martin Weigt, Juliana Bernardes

ITALY

2011-2014 **Summer School on Atomistic Simulation Techniques**, 3x15h TP, SISSA, Trieste, Italy.

Coordinators: Giovanni Bussi, Cristian Micheletti, Alessandro Laio

2013 **Introduction to Physics**, 28h CM, Università Cattolica del Sacro Cuore, Milano, Italy.

Coordinator: Alberto Granato

PUBLICATIONS

INTERNATIONAL PEER-REVIEWED JOURNALS

- J₁ M. Muscat, G. Croce, **E. Sarti**, M. Weigt,
FilterDCA: Interpretable supervised contact prediction using inter-domain coevolution
(2020) PLoS Comput Biol 16(10)
10.1371/journal.pcbi.1007621
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007621>
- J₂ A. A. Aleksandrova, **E. Sarti**, L. R. Forrest,
MemSTATS: a benchmark set of membrane protein symmetries and pseudo-symmetries
(2020) J Mol Biol 432(2) 597 – 604
10.1016/j.jmb.2019.09.020
<https://www.sciencedirect.com/science/article/pii/S0022283619305753>
- J₃ **E. Sarti***, A. A. Aleksandrova*, S. K. Ganta, A. S. Yavatkar, L. R. Forrest,
EncoMPASS: an online database for analyzing structure and symmetry in membrane proteins
(2019) Nucleic Acids Res. 47(D1), D315 - D321
10.1093/nar/gky952
<https://academic.oup.com/nar/article/47/D1/D315/5144152>
- J₄ A. Battisti, S. Zamuner, **E. Sarti**, A. Laio,
Toward a unified scoring function for native state discrimination and drug-binding pocket recognition
(2018) Phys Chem Chem Phys, 20(25) 17148-17155
10.1039/C7CP08170G
<https://pubs.rsc.org/en/content/articlelanding/2018/CP/C7CP08170G>
- J₅ **E. Sarti**, I. Gladich, S. Zamuner, B. E. Correia, A. Laio,
Protein-protein structure prediction by scoring molecular dynamics trajectories of putative poses
(2016) Proteins Struct Func Bioinfo, 84(9), 1312-1320
10.1002/prot.25079
<https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25079>
- J₆ **E. Sarti**, D. Granata, F. Seno, A. Trovato, A. Laio,
Native fold and docking pose discrimination by the same residue-based scoring function
(2015) Proteins Struct Func Bioinfo, 83(4), 621-630
10.1002/prot.24764
<https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24764>
- J₇ **E. Sarti**, S. Zamuner, P. Cossio, A. Laio, F. Seno, A. Trovato,
BACHSCORE. A tool for evaluating efficiently and reliably the quality of large sets of protein structures
(2013) Comput Phys Comm, 184(12), 2860-2865
10.1016/j.cpc.2013.07.019
<https://www.sciencedirect.com/science/article/abs/pii/S0010465513002488>
- J₈ S. Bietti, C. Somaschini, **E. Sarti**, N. Koguchi, S. Sanguinetti, G. Isella, D. Chrastina, A. Fedorov,
Photoluminescence study of low thermal budget III-V nanostructures on silicon by droplet epitaxy
(2011) Nanoscale research letters, 5(10), 1650
10.1007/s11671-010-9689-8
<https://nanoscalereslett.springeropen.com/articles/10.1007/s11671-010-9689-8>

SEMINARS AT INTERNATIONAL PEER-REVIEWED CONFERENCES

- C₁ **E. Sarti**, A. A. Aleksandrova, S. K. Ganta, A. S. Yavatkar, L. R. Forrest,
Analyzing the Structure and Symmetry of Membrane Proteins through the Systematic Online Database EncoMPASS,
Biophysical Society 2018, San Francisco (CA) USA
- C₂ **E. Sarti**, A. A. Aleksandrova, S. K. Ganta, A. S. Yavatkar, L. R. Forrest,
EncoMPASS: an Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry,
ISMB/ECCB 2017, Prague (CZ)
- C₃ **E. Sarti**, A. A. Aleksandrova, S. K. Ganta, A. S. Yavatkar, L. R. Forrest,
EncoMPASS: an Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry,
GRC Membrane protein folding 2017, Easton (MA) USA

POSTERS AT INTERNATIONAL PEER-REVIEWED CONFERENCES

- P₁ M. Muscat, G. Croce, **E. Sarti**, M. Weigt,
A supervised but interpretable coevolutionary predictor of protein-protein interactions
ISMB/ECCB 2019, Basel (CH)
- P₂ M. Muscat, G. Croce, **E. Sarti**, M. Weigt,
A supervised but interpretable coevolutionary predictor of protein-protein interactions
Statistical Physics Approaches to Systems Biology 2019, La Habana (CU)
- P₃ **E. Sarti**, L. R. Forrest,
Transmembrane topology alignment for identifying related protein structures
ISMB/ECCB 2018, Chicago (IL) USA
- P₄ M. Muscat, G. Croce, **E. Sarti**, M. Weigt,
A supervised but interpretable coevolutionary predictor of protein-protein interactions
Approches Interdisciplinaires de l'Evolution Mollaire 2018, Villeneuve d'Ascq (FR)
- P₅ **E. Sarti**, A. A. Aleksandrova, S. K. Ganta, A. S. Yavatkar, L. R. Forrest,
Analyzing the Structure and Symmetry of Membrane Proteins through the Systematic Online Database EncoMPASS
Biophysical Society 2018, San Francisco (CA) USA
- P₆ **E. Sarti**, A. A. Aleksandrova, S. K. Ganta, A. S. Yavatkar, L. R. Forrest,
EncoMPASS: an Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry
ISMB/ECCB 2017, Prague (CZ)

SEMINARS AT NATIONAL PEER-REVIEWED CONFERENCES

- C₄ **E. Sarti**, A. Carbone,
Statistical analysis of coevolutionary signals on the SARS-CoV-2 Spike protein
Journée ASIM SARS-CoV-2 et Bioinformatique Structurale 2020 (en ligne)

SUBMITTED PAPERS

- J₉ R. Staritzbichler*, **E. Sarti***, E. Yaklich, M. Stamm, A. A. Aleksandrova, K. Khafizov, L. R. Forrest,
Refining pairwise sequence alignments of membrane proteins by the incorporation of anchors
(2020) bioRxiv
10.1101/2020.09.16.299453
- J₁₀ **A. A. Aleksandrova***, **E. Sarti***, L. R. Forrest,
EncoMPASS: an Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry
(2018) bioRxiv
10.1101/391961

SOFTWARE

- L₁ **Locusts** : Family=utility; Audience=community; Evolution=lts; Duration=<1; Contribution=leader; Url=<https://github.com/EdoardoSarti/locusts>

Locusts is a Python3 / bash package for distributing very large amounts of short jobs on HPC systems. Each job runs in a protected environment and data is transferred securely from a local machine to the HPC and vice versa. Locusts can also be applied to any local machine or local cluster without the need to rely on a scheduling system. The package was born as part of the EncoMPASS project and has been extensively tested on different HPCs. I am the creator and the only developer and tester of this algorithm.

- L₂ **Pfam Interactions** : Family=vehicle; Audience=partners; Evolution=lts; Duration=1; Contribution=leader; Url=https://github.com/infernet-h2020/pfam_interactions

Pfam Interactions is a Python3 platform for training and testing algorithms that use multiple sequence alignments (MSA) for the prediction of protein structure and its complexes. It uses the MSAs contained in the Pfam database, which classifies more than 18,000 protein domains by hidden Markov models, and the PDB (Protein Data Bank), which provides the 3D coordinate files of proteins. Pfam Interactions performs a very precise mapping between sequence and structure resources, and is able to evaluate the accuracy of any protein structure prediction algorithm, by calculating different metrics (positive predictive value or PPV, area under the curve or AUC) and by combining visual and interactive analysis and tools.

I developed this set of tools as part of FilterDCA, and it has already been used in 3 other works, now being part of the analysis tool library of Martin Weigt's group.

- L₃ **EncoMPASS** : Family=research; Audience=community; Evolution=lts; Duration=4; Contribution=leader; Url= <https://github.com/Lucy-Forrest-Lab>

EncoMPASS is an online database (~ 200 users per month) for analyzing and relating the structures and symmetries of membrane proteins. To do this, this Python3 software combines information from different databases, standardizes their description, inserts them into a lipid bilayer model, classifies them in terms of structure and topology, and finds an exhaustive set of internal symmetries. EncoMPASS uses several external resources, which are provided in an associated Singularity container. Completely automated, it returns a file system where data, numbers and interactive content are stored and ready to be uploaded to the web server. EncoMPASS also includes an update procedure for the integration of new data, which is performed every 2 months.

I developed, with Antoniya Aleksandrova, the general framework, the libraries, the original algorithms and the web content of EncoMPASS. As of this writing, the current version based on the HPC Locusts task manager is in its late testing phase, and according to US federal security protocol, it cannot yet be released to the public. It will appear on the GitHub address provided soon. Do not hesitate to contact me in case of need.

PAST RESEARCH ACTIVITIES

My research career is characterized by a diverse and extensive set of methodologies and approaches, ranging from statistical mechanics to bioinformatics to biophysics and computational biology. Yet, two common traits emerge: the interest for the structural and evolutionary aspects of proteins, and the development and optimization of novel algorithmic approaches to broad computational problems.

Proteins are of central interest for the molecular biology community, as they perform most of the functions of any living organism. Their understanding in terms of chemistry, structure, and function are important also to a pharmaceutical and medical level, since they represent most of the drug targets and virtually any known disease can be associated to their malfunctioning. Proteins usually work by interacting with each other and creating complex molecular machineries.

In the same way as the shape of a screwdriver or a wrench is suggestive to their function, knowledge about the three-dimensional structure of a protein is of paramount importance. Yet, experimentally determining it is time- and resource-intensive. One traditional strategy to predict protein structure is to propose physically credible conformations, and then assessing their quality via a scoring function. Within this scope, my PhD studies focused on the optimization of BACH, an existing scoring method for protein structure assessment, and on the creation and development of **BACH SixthSense**, a novel algorithm allowing the generalization to protein complexes. The optimization of BACH brought to an average 10-fold reduction of its execution time also owing to a complete reimplementing of the calculation of the protein's accessible surface, described in [J7]. BACH SixthSense achieved top performances in assessing both single protein and protein complex structures and was described and applied in [J4, J5, J6]. The theoretical framework used in the design of BACH SixthSense described in my thesis relies on the calculation of a generalized relative entropy estimator capturing the entropic difference between train and test feature distributions. BACH SixthSense and its framework will be detailed in Form 1.

Membrane proteins represent ~30% of the proteome of each living organisms and perform their functions within the environment of the lipid bilayer (such as the cell membrane or the membrane of most cellular organelles). Their structure is constrained by the geometry of the membrane, thus collapsing the enormous variety of proteins in only two main topological classes. For this reason, algorithms trained to recognize similarities between soluble proteins are not accurate on membrane proteins: a complete, new classification and set of analyses must be designed. This is what urged the creation of **EncoMPASS**, an online database for relating membrane proteins by their structure and set of internal symmetries. Together with Antoniya Aleksandrova I developed the whole code for the generation of the database, which includes several original algorithmic contributions such as an HMM/SVM topological alignment algorithm [P3], and the upcoming package locusts [L1], the publication of both of which is planned by the end of 2021. EncoMPASS is described in [J3, J10] and presented in [C1, C2, C3, P5, P6]. It receives 100-200 unique users per month and has been already applied for systematic analyses on membrane structures [J2]. The novel approach of EncoMPASS will be the subject of Form 2.

The chemical formula of a protein, represented as a sequence of amino acids, uniquely defines its structural and functional properties. Since experimentally determining the structure of proteins is often prohibitive, sequence data are used in order to infer their sequence. In order to do this, a broad class of methodologies use coevolution: since the protein's internal contacts often induce correlated mutations of the amino acids involved, we can retrieve structural information from how mutations correlate in a multiple sequence alignment (MSA). Direct Coupling Analysis (DCA) and Blocks in Sequences (BIS) are two methodologies based on this intuition, but exploiting the very different frameworks of, respectively, statistical mechanics and clustering combinatorics. Whereas the two techniques both have excellent performances on single proteins, they are not as successful on protein-protein complexes.

I have implemented **FilterDCA**, a new version of DCA that also considers structural information in the form of convolutional filters aiming at reusing the signal from the otherwise very noisy contact interaction contact maps. FilterDCA is described in [J1] and has been presented in [P1, P2, P4]. In order to thoroughly validate the method, I created Pfam Interactions [L2], a testing platform providing a very accurate sequence-structure matching. Since October 2020 I am instead applying the BIS framework to the case of SARS-CoV-2. BIS is only one of the tools that are being used for studying viral protein evolution and interactions. The project is conducted together with CIRI, and preliminary results of my work have already been presented in [C4]. One of the main research directions that are emerging from the study of SARS-CoV-2 evolution will be presented in Form 3.

Form 1: Efficient assessment of protein-protein interaction conformations

1. Description of the contribution

Predicting which pairs of proteins interact and how they do it (protein-protein interaction, PPI) is one of the main challenges of today's molecular biology and biophysics. Different approaches can be used to screen possible interacting conformations of pairs of proteins, but in the end each method will rely on a scoring function for assessing the quality of each conformation and predict whether any of them might reflect the natural conformation of the protein complex. Even if their job is only to assess – and not to propose – protein-protein conformations, these scoring functions are notoriously difficult to design.

Our approach was to start from an existing knowledge-based statistical scoring function for protein folding conformations, called BACH (Bayesian Analysis Conformation Hunt). The method classifies each pair of amino acids as either in one out of 5 types of chemical contacts or in none. A pre-trained score is then associated to each term. The algorithm uses naive Bayesian learning for training the internal parameters, estimating a generalized relative entropy. Despite its effectiveness on protein folding problems, BACH's performances were poor over PPI problems.

Two main improvements were made: 1) the redefinition of the algorithm's contact classes using notions of information theory and biochemistry greatly increased its accuracy, and 2) the introduction of a penalty score accounting for steric clashes filtered out more efficiently the unphysical conformations.

We tested the new BACH SixthSense over extensive databases of conformation decoys, and against an exhaustive list of state-of-the-art scoring functions, where it showed top performances in discriminating the correct conformation over hundreds-to-thousands of decoys.

2. Personal contribution of the applicant

The design of this and other statistical potentials constituted my main doctoral project. Whereas the BACH algorithm had been already implemented, I redesigned it from scratch, obtaining a considerable decrease in execution times (nearly 10 times benchmarked over thousands of test cases). I also wrote the entirety of BACH SixthSense's code and contributed to its translation from Fortran90 to C++ (together with Stefano Zamuner).

I also developed the whole information-theoretical framework for BACH (originally described in a purely Bayesian framework), which induced the modifications bringing to BACH SixthSense.

BACH SixthSense was tested and employed in different studies, one of which included extensive molecular dynamics simulations, also performed by me with a 5-million computing hour grant that I submitted together with Ivan Gladich.

During the course of this project I collaborated with Alessandro Laio, Antonio Trovato, Flavio Seno (my supervisor and co-supervisors), Stefano Zamuner (implementation of BACHSCORE), Pilar Cossio, Daniele Granata, Ivan Gladich (tests) and Bruno Correia (technical support).

3. Originality and difficulty

The originality of the work lies on using a simple statistical model designed to grasp the appropriate physical and chemical details of protein-protein interaction, instead of trying to use plausible but approximated physical models accounting for the different energy contributions or using very large sets of parameters.

4. Validation and impact

Despite being in the era of AI-driven structural biology algorithms, the BACH project has continued to grow and be used as a lightweight and fast tool by the Laio and Cossio research groups, and was recently employed in PARCE (Protocol for Amino acid Refinement through Computational Evolution).

5. Dissemination

BACH SixthSense was described and applied in [J₄, J₅, J₆].

Form 2: Sequence and structure similarities of membrane proteins

1. Description of the contribution

Membrane proteins perform their function in a lipid bilayer, such as the cell membrane or the membrane of many cellular organelles. They constitute ~30% of the proteome of each living organism, and present unique characteristics: their structural conformations are strongly constrained by the geometry of the lipid bilayer, which imposes similar structures to membrane proteins performing very different functions. Assessing structural similarities is thus much more challenging than for soluble proteins.

Several protein structure classifications exist, yet none of these manage to grasp the structural differences between families or even superfamilies of membrane proteins. For this reason, we created EncoMPASS (Encyclopedia of Membrane Proteins Analyzed by Structure and Symmetry), an online database that relates membrane proteins by aligning and comparing their sequences, structures and symmetries. EncoMPASS is automatically updated every 2 months and provides many useful metrics and analysis tools with the aim of defining similarity networks and contribute to the correct classification of membrane proteins.

One of the problems encountered in the design of EncoMPASS is how to compare pairs of protein structures: if the two structures are too different, the similarity score will not be reliable, but we cannot know whether we are in this situation until we compare them. Also, similarity is not related to any simple feature we can interrogate. We thus needed a faster way to probe the structural similarity of each pair of proteins, knowing that a full structural superposition is beyond $O(n^6)$ (depending on the implementations), with n the number of superposable amino acids. Proteins can be described as a concatenation of basic rigid structures (secondary structure elements) which can be described with a constant set of features. Then, a pairwise Hidden Markov Model trained on manually curated structural alignments can be used for finding the best coarse-grained structural alignment and concurrently an absolute score of the alignment (i.e., not relative to the pair of proteins used). By imposing a threshold on this score, I therefore obtained a filter over the more similar pairs of structures, on which the full structure alignment algorithm is then executed. The Viterbi algorithm for calculating this score has complexity $O(s^2)$, s being the number of superposable secondary structure elements (much lower than the number of amino acids) in the proteins.

2. Personal contribution of the applicant

I am the author of the EncoMPASS pipeline (in Python3) of data retrieval, processing, and sequence and structural analysis, whereas Antoniya Aleksandrova implemented the symmetry analysis part. I also released Locusts, a Python3 package for distributed computing of very large batches of short jobs. I am currently maintaining and updating the code.

Lucy Forrest (supervisor), Antoniya Aleksandrova (implementation and testing), Srujan Ganta, and Amar Yavatkar (web development) have collaborated to this project.

3. Originality and difficulty

EncoMPASS is the first database to compare the structure and symmetry of any membrane protein of known structure. The design of the pipeline required adaptive solutions ranging from solving inconsistencies between sources of information, to the design of specific analysis algorithms and protocols, to the optimization of the computational-intensive pipeline.

4. Validation and impact

The Encompass website is the first systematic benchmark for membrane protein structure similarity, and can be used to find homologous proteins serving as structural models for a wide range of studies. EncoMPASS has ~200 unique users per month and is becoming a reference database in the membrane protein community.

5. Dissemination

One of my contributions to AlignMe will be described in [J₉]. EncoMPASS is described in [J₃, J₁₀, C₁, C₂, C₃, P₅, P₆]. The filtering algorithm has been presented in [P₃].

Form 3: Predicting the evolution and interactions of SARS-CoV-2 proteins

1. Description of the contribution

Viruses evolve much faster than any other organism. Extreme variation is indeed a clever defensive strategy for organisms living under constant evolutionary pressure – that is, under constant attack from other organisms aiming at destroying them. SARS-CoV-2 makes no exception: in one year of pandemics, we have already witnessed the emergence of multiple variants defying or reducing the effect of vaccines, and this trend is bound to continue (as for the flu or the common cold, which for the same reason are impossible to eradicate). It thus becomes of outmost importance to try and predict which evolutionary trajectories the virus is more likely to take.

The Analytical Genomics (AG) group lead by Alessandra Carbone recently developed GEMME, an algorithm that uses multiple sequence alignments (MSAs) in order to identify the phylogenetic predisposition to mutations of each amino acid of a query sequence. In addition, AG also developed BIS2, a methodology that, from the same input as GEMME, identifies clusters of amino acids having similar evolutionary histories (i.e., present coevolutionary traits). The aim of my present research is to combine the two algorithms together for obtaining a new and more powerful approach. On one side, I am working on the design of a variational implementation of GEMME that, starting from two similar MSAs, is able to correctly identify the relevant score differences which would be connected to mutation-prone sequence sites. On the other, I want to exploit the information coming from the BIS2 analysis in order to train an attention algorithm.

Thanks to an unprecedented, coordinated effort, extensive sequencing data about SARS-CoV-2 are available: the GISAID database already contains several hundred thousand sequence reads precisely annotated with time and geographical references. These data will be used to systematically test and train the new mutation prediction algorithm.

2. Personal contribution of the applicant

Whereas GEMME and BIS2 had already been developed prior to my arrival in AG, I am now completely in charge of the project, its implementation and testing. Alessandra Carbone and Elodie Laine are collaborating with interesting discussions and ideas, as well as with their expertise in the development and application of the two algorithms.

Since a correct understanding of both the methodological framework and the rapidly growing available data is central to the success of this project, I have also conducted preliminary works on the study of SARS-CoV-2 spike mutations with BIS2, with which I have been able to pinpoint the main evolutionary events the community agrees upon.

3. Originality and difficulty

The originality of this approach lies in the combination of phylogenetic information with statistical correlations. The computational methods involved have a complexity not superior to $O(mn^2)$, where m is the number of sequences in the MSA and n is the number of amino acids of the query sequence.

4. Validation and impact

Predicting future mutations of the SARS-CoV-2 virus is obviously a very anticipated and difficult objective. Moreover, the methodology is completely general and can be applied to any protein, provided a MSA of homologous sequences can be provided. Most importantly, the study is also shedding light on the information contained in very common bioinformatics data structures, such as phylogenetic trees and MSAs, and will lead to original analysis algorithms that will be implemented and made available in the context of the algorithm repository of the laboratory of quantitative and computational biology (LCQB).

5. Dissemination

The project is still ongoing, but preliminary results and analyses have already been presented in [C4].

Geometrical and statistical algorithms for viral protein evolution prediction

Abstract

Viruses constantly mutate in order to defy the defenses of their hosts: they explore the so far uncharted sequence space for maximizing their efficiency. My research project proposes a computational framework for characterizing it and designing stochastic processes for reproducing and thus predicting viral evolution.

1. Impact

[Predicting the next viral strain] Viruses are one of the oldest, most diverse and mysterious groups of biological entities. In order to replicate, they exploit the environment and the molecular processes of the host organism, that they are able to navigate undisturbed thanks to their formidable adaptability: their genome mutates $\sim 10^4$ times faster than any other organism, and new strains defying the host's immune response keep appearing. Everyone has recently become aware of the importance of studying the evolution of viral proteins, which on the one hand makes it possible to elucidate their origins and on the other to predict the new mutations that could lead away from the host's immune system. Recent technical advances in the analysis of biological macromolecules have made available an enormous amount of information about the genomes of viruses and the structure of their proteins, thus opening unprecedented possibilities for developing broad computational approaches for the understanding, prediction and prevention of mutational patterns in viral proteins.

[Characterizing and exploring data spaces of variable dimension] Any protein's structure and function depend on its chemical sequence, a ~ 50 - to >2000 -character word using an alphabet of 20 symbols (amino acids). All life relies on the mutation of these sequences (encoded as DNA genes): diversity allows organisms to adapt, generation after generation, to the ever-changing environment. Thus, punctual modifications (character insertions, deletions, and replacements) are proposed every time the organism replicates, and by natural selection, the most efficient resulting set of proteins will have better chances to be passed along to the offspring and propagate throughout the species.

Due to their variable length, protein sequences cannot be represented in terms of coordinates, yet the space itself can be described by a set of pairwise distances according to a metric of choice. We can think of evolution as a stochastic process on this space guided by a potential called fitness, which is unknown but for which local approximations can be made. Although the problems of geometric characterization of complex datasets, stochastic processes on unknown potentials, and efficient exploration of large energy landscapes have all been tackled separately, current methodologies are hardly compatible with each other and an integrated approach at the characterization and exploration of sequence spaces has not yet been proposed.

The reconstruction of scalar fields on spaces of variable dimension is thus an ambitious and anticipated goal with important repercussions in computer science and the understanding of a variety of physical processes.

2. Challenges

Four main challenges can be identified, each corresponding to a research objective:

- A. The geometrical and topological characterization of the protein sequence space
- B. The realization of a stochastic process guided by an approximation of the unknown fitness potential
- C. The estimation of the true fitness potential for large regions of the sequence space and the prediction of viral protein evolution
- D. The application of this model within the scope of viral membrane proteins in order to study their function

3. Strategy

[Presentation of the project] The aim of this project is two-fold: on one side it proposes to extend and deepen the studies on the evolution and the function of viral proteins that I started in the Analytical Genomics team, on the other it will develop original methodologies based on the combination of my expertise in computational physics, bioinformatics and statistical mechanics and the skills and algorithmic resources of ABS. Four main objectives will be achieved:

- I. The characterization of the effective sequence space, requiring a new on-the-fly algorithmic approach for calculating the (variable) intrinsic dimension of the dataset
- II. The design of a position-dependent local approximation of the true fitness potential, based on statistical mechanics and artificial intelligence approaches
- III. The exploration of large regions of the space via suitably adapted enhanced sampling techniques able to escape potential basins while providing a thermodynamically rigorous description.

IV. The validation of each step of this methodology by means of several existing experimental datasets, and its application to relevant biological problems.

In order to carry out these objectives I will also develop an extensive library containing a diverse set of algorithms to efficiently solve relevant problems pertaining this scope, such as a generalized Wang-Landau approach, and a tool for geometrical and topological analysis. Low-level packages corresponding to the generic algorithms will be combined with a new integrated application for evolutionary dynamics (Metaevolution), which will be designed during the project.

[Background] The following methodologies will be fundamental to the project:

- Direct coupling analysis (DCA): estimation of the correlation between the mutations occurring at any two sites of a protein (described via a multiple sequence alignment, or MSA) through the calculation of their unmediated mutual information, expressed as the coupling term in a Potts Hamiltonian. DCA has been used as a generative model of sequences of a given family: its potential is a reliable measure of the fitness of a protein sequence.
- Wang-Landau algorithm: Monte Carlo importance sampling method designed to estimate the density of states (volume in phase space) of a system. The method performs a non-Markovian random path to construct the density of states by rapidly visiting the entire available energy spectrum. It can be applied to any system characterized by a cost (or energy) function.

Objective I - Characterization of the sequence space

If we were to randomly propose sequences of any length with a 20-letter alphabet we would end up sampling the whole very high-dimensional space of all possible sequences. If we instead restrict to natural protein sequences only, the space dimensionality will be much lower, but due to the complex topology of the set of sequences, they will still lie on an even lower-dimensional subspace. Attempts at calculating the intrinsic dimension (ID) of such sets of points have been made for limited datasets, for which a constant ID could be assumed: these studies have found IDs in the range of 5-15. Yet, in order to explore vast regions of this space, the assumption of a constant dimensionality must be neglected. The following steps are then to be taken:

- I.a) A reliable filtration (computational topology) of the sampling, its global dimension and its Betti numbers will be calculated using a topological data analysis method. Several implementations have already been designed, such as the GUDHI library developed at INRIA Sophia Antipolis, which integrates perfectly with the SBL tools.
- I.b) For a given point of the space, the intrinsic dimension relative to the neighborhood must be calculated on-the-fly. The recent nonlinear method TWO-NN will be chosen for a new C++ efficient and robust implementation, as it only needs a restricted set of pairwise distances (whose calculation constitutes the main bottleneck). Local neighborhoods and other precautions inspired by molecular dynamics engines will be employed.
- I.test) An artificial dataset generated by a mixture of overlapping distributions of variable dimensions will be created for testing purposes. In parallel, the ID of some big protein sequence families (e.g. viral fusion proteins, retrotranscriptases, etc.) will be assessed and compared with other state-of-the-art methods.

Objective II - Creation of a potential for guiding the evolutionary stochastic process

Markov Chain Monte Carlo (MCMC, here described in its Metropolis-Hastings formulation) can be used as the fundamental stochastic process for exploring the sequence space. From a starting sequence s , a new sequence s' is drawn from a probability distribution $g(s'|s)$, and accepted with probability 1 if $P(s')g(s|s') > P(s)g(s'|s)$, or with probability $P(s')g(s|s)/P(s)g(s'|s)$ otherwise. For symmetric proposing distributions $g(s'|s) = g(s|s')$, the decision becomes independent of g (yet g is still used for proposing the new step). Additionally, we can imagine that the probability distribution $P(s)$ is generated by a potential $H(s)$ (called fitness) such as $P(s) \sim \exp\{-H(s)\}$. In order to perform MCMC we thus need to model $g(s'|s)$ and $H(s)$, or at least some distribution $g_s(s'|s)$, $H_s(s')$ that approximate the two real distributions in a neighborhood of s . In order to design a faithful mathematical model for the explored sequence space, we follow three steps:

- II.a) The stochastic process must propose sequences lying on the natural sequence space characterized in Objective I. Lacking a proper system of coordinates, the position of a sequence is defined by the distance from at least $d+1$ other sequences, where d is the local intrinsic dimension calculated in I.b. $g_s(s'|s)$ will thus be a uniform distribution over a set of spatial constraints which can be implemented using Cayley-Menger determinants (distance geometry).
- II.b) Whenever the process explores a region of the space populated by sequences of a given protein family, the DCA score can be used to generate new sequences, on which the method had previously been trained. Thus $H_s(s') = H_{[fam(s)]}(s')$. If we generate a DCA model for each cluster of natural sequences, we can combine the potentials into a weighted sum $H_s(s') = \text{SUM}_{fam}\{\exp[-d(s', \text{center}(fam))/d_0] H_{fam}(s')\}$ where

each weight depending on the distance between the current location and the center of a family vanishes with increasing distance.

Test) The algorithmic precision of the proposed stochastic process can be validated on artificial datasets of strings of variable length, since DCA is a very generic framework that does not make assumptions on the origin of the input sequences. In addition, we can use the many deep mutational scanning datasets, which report, for a target sequence, the experimental fitness upon mutating any single or pair of amino acids.

Objective III - Exploration of fitness pools and evolutionary prediction

Once a stochastic process capable of reproducing the features of the sequence space has been designed, importance sampling methods are needed in order to explore larger regions of the space, and thereby calculate the density of states needed for the reconstruction of the missing sequence data. Unlike umbrella sampling methods, Wang-Landau has the advantage of not having to define a set of reaction coordinates (i.e., preferential directions along which the stochastic process will develop). Current implementations of the Wang-Landau algorithm can be quickly adapted to the exploration of sequence spaces. The algorithm has been successfully tested in spaces with up to 60 dimensions, which is a compatible upper limit for sequence spaces. The new version of Wang-Landau will be devised in the following steps:

III.a) As in the stochastic process, the quality of the proposal in the Wang-Landau algorithm is central for a correct exploration of the space. The main challenge of this objective will be to devise a new adaptive proposal mechanism accounting for the properties of the sequence space characterized in Objective I.

III.test) The evolutionary version of the Wang-Landau algorithm can be tested on well-studied cases of divergent evolution in protein families. One such case is the metallo-beta-lactamase (MBL) protein family, whose sequence distribution has been recently studied and described by Tokuriki's group. Viral examples can include SARS-CoV, HIV-1 and influenza strains.

Viral evolution constitutes an ideal and very hard validation set for the described theoretical framework: the annotated sequencing of a large variety of viral proteins can be found in comprehensive online resources such as GISAID (for SARS-CoV-2) or GenBank. For each viral protein sequence, a large set of reads performed at different times and locations is available, allowing to accurately reconstruct the protein's evolutionary trajectory. Tests on the prediction of SARS-CoV-2 variants will be performed using the data that I am already curating within the scope of my postdoctoral project in the LCQB Analytical Genomics group at Sorbonne Université.

Objective IV – Application: functional classification of viral membrane proteins (VMPs)

Viruses are in constant interplay with membrane environments. Integral viral membrane proteins (VMPs) are thus essential in viral genetic repertoires throughout any ecosystem, yet their functions and mechanisms remain so far poorly understood.

Mutational data can be combined with analyses of the structure and dynamics of different proteins by using already existing frameworks: the structures of viral membrane proteins are already collected, related and analyzed in the EncoMPASS online database (<https://encompass.ninds.nih.gov/>), of which I am one of the two authors and maintainers. The EncoMPASS database provides a network of structure and sequence relationships between membrane proteins, and is thus a suitable framework for the study of viral membrane proteins. The following steps will be taken:

IV.a) An extensive search throughout literature and existing databases for specific families or subclasses of VMPs as well as the use of state-of-the-art structural prediction algorithms will be followed by several structural quality checks with the renowned MolProbity tool as well as other more recent or specific approaches such as MAIDEN and ProteinGCN.

IV.b) Since the collected data will be of heterogeneous quality and, if merged with the rest of EncoMPASS, would artificially overrepresent a subset of structures, I will design EncoMPASS-IT (EncoMPASS Interactive Terminal), a downloadable package allowing each user to locally run the EncoMPASS analyses on any coordinate file describing membrane protein conformations. The package will then relate and contextualize the user's models with the experimentally verified entries contained in the public database: the user's data will thus remain local and private, while still being able to exploit the online resources of EncoMPASS.

IV.c) The structural and evolutionary knowledge about the characterization of VMPs accumulated throughout the project can be combined in order to collect insights on functional characterization. A new relational GNC (R-GCN) could be based on the VMP database for labelling each VMP family according to any existing protein functional classification.

IV.test) The performance of the R-GCN can be assessed on the current functional annotations contained in NCBI Virus, and in databases curating more specific sets of viral families.

4. International collaborations and experimental validation

Throughout the development of this project I will also continue some of the scientific collaborations I have already established at national and international level: my ongoing collaboration with the Center for Research in Infectious Diseases (CIRI) on the study of the protein-protein interaction network of SARS-CoV-2 could be extended to the study of VMPs interactions, and could also open very interesting perspectives in the field of drug discovery. Moreover, my ongoing collaboration with Dr. Lucy Forrest's laboratory at NINDS (NIH) will be a key aspect of the realization of EncoMPASS-IT.