M2 INTERNSHIP PROPOSAL

Statistical Protein Design

Laboratory name : Laboratoire Jean Perrin and Center for Interdisciplinary Research in Biology
CNRS identification code: UMR8237 and UMR 7241
Internship director's surnames : Clément Nizak & Olivier Rivoire
e-mail : clement.nizak@upmc.fr, olivier.rivoire@college-de-france.fr Phone : 01 44 27 12 79
Web page : http://statbio.net https://www.labojeanperrin.fr/?jobs71
Internship location : Sorbonne Université, 4 Place Jussieu - Collège de France, 11 place Marcelin Berthelot, 75005 Paris
Thesis possibility after internship : YES

Since the discovery of the genetic code more than half a century ago, understanding the relation between the amino-acid sequence of a protein and its function (specific binding, catalysis, information transmission...) has remained an open problem in biology. The traditional approach to this problem is to combine structural biology (3D protein structures determined by X-ray crystallography or NMR) and computer modeling. This approach has recently cracked the 'sequence-3D structure' problem [1]. The 'sequence-function' problem remains, however, open. As a consequence, we cannot design new synthetic protein sequences with arbitrary function. We are indeed limited by the intractability of computational models at the ms-scale, the timescale of many protein functions, and our theoretical understanding of how these ms-scale motions translate into function.

A completely different approach is to look at the problem from the standpoint of Evolution, the dynamical process by which natural proteins are formed, and to ask **how evolution encodes function into protein sequences**. This approach has been taken to infer statistical models of the sequence-function relation from large datasets of protein sequences using tools from statistical physics (for instance Potts models). These statistical models, which are in sequence space rather than physical space, are generative and can produce new synthetic protein sequences, so-called **statistical protein design**. A first demonstration following this new route has recently been published [2].

We apply this approach to study enzymes, which are proteins that catalyze biochemical reactions. Enzymes display remarkable catalytic properties that are out of reach of current molecular engineering approaches, such as exquisite substrate specificity and $>10^{15}$ acceleration rates. These properties likely emerge from long-range cooperative effects between amino-acids that are captured by statistical models. The main novelty of our approach is to supply the statistical models with unprecedented quantitative data obtained from controlled experiments where we measure at high-throughput multiple functional properties.

Our experiments consist in constructing libraries of thousands to millions of enzyme variants that we encapsulate one by one in mono-disperse picoL droplets using microfluidic devices (see for instance [3]). The encapsulated protein variants are expressed and assayed in each droplet for enzymatic function in a one-day experiment. High-throughput sequencing read-out yields quantitative information to learn statistically the sequence-function relation.

We are looking for a candidate with a strong background in either physics, mathematics or computer science and a strong interest for biological problems. Prior experience with molecular biology and microfluidics are not required but the candidate should be ready to learn these techniques. The **M2 internship** will focus on learning the experimental workflow. The **PhD work** will combine experiments and data analysis and/or theoretical modeling.

The project will take place in an interdisciplinary team of physicists and biologists, theoreticians and experimentalists. The projects also involves collaborations with theoretical physicists at ENS Paris and with experimentalists at the University of Chicago.

Keywords: quantitative biology; statistical physics; machine learning; protein evolution; microfluidics

References:

1. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016;537(7620):320–7.
2. Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, Hilvert D, Monasson R, Cocco S, Weigt M, Ranganathan R. An evolution-based model for designing chorismate mutase enzymes. Science. 2020 Jul 24;369(6502):440–5.
3. Fallah-Araghi A, Baret J-C, Ryckelynck M, Griffiths AD. A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. Lab Chip. The Royal Society of Chemistry; 2012 Mar 7;12(5):882–91.