INTERNSHIP / PhD PROPOSAL

Statistical Protein Design

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 scale 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 a very well characterized enzyme called trypsin that has resisted attempts to solve the sequence→function problem despite many efforts. 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 experimental workflow consists in constructing libraries of 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 in each droplet, assayed for enzymatic function using fluorescent assays and sorted into bins that correspond to each level of enzymatic activity. High-throughput sequencing of protein-encoding genes in each bin yields quantitative information from which we aim at learning statistically the sequence-function relation. Thousands to millions of variants are assayed in a 1-day experiment.

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 experiences 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, located at Collège de France in Paris. 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 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; 201 2;12(5):882–91.

Contacts: clement.nizak@espci.fr , olivier.rivoire@college-de-france.fr

Location: Center for Interdisciplinary Research in Biology (CIRB) Collège de France, 11 place Marcelin Berthelot, 75005 Paris
Webpage: www.statbio.net