Statistical Design of Serine Protease Substrate Specificity

- **Context:**

We study the problem of **engineering enzyme substrate specificity using S1A proteases as a model system**. These enzymes belong to a large family including trypsin, chymotrypsin, elastase, that spans a wide array of host species and are involved in many processes. More than 150,000 sequences of the corresponding Pfam database entry are now available. These enzymes share the same global fold and catalytic mechanism which relies on a triad of three conserved amino-acids in their catalytic site. Their structure and biochemical properties have been characterized in detail (*1*).

These proteases catalyze the hydrolysis of a peptide bond between two successive amino-acids in proteins, with generally a very high substrate specificity. They cut indeed peptide chains at very precise motifs, almost uniquely defined by the amino-acid immediately preceding the cut bond. For instance, trypsin cuts very efficiently peptide bonds that are downstream of the basic amino-acids arginine and lysine (nearly regardless of flanking amnio-acids), and cuts typically 10^6 times less efficiently bonds downstream of other amino-acids. For our purpose, **engineering their exquisite substrate specificity has proven challenging in spite of the wealth of data available on their structure and catalytic mechanism and of many efforts towards this goal for the past decades (*1*).**

- **State of the art:**

The achievements and limits of the structural approach for substrate specificity engineering.

Mutating amino-acids of the substrate binding pocket, which is generally considered as structurally distinct from the catalytic site, seemed a promising route for enzyme substrate specificity engineering. Yet, in S1A proteases and other related classes of serine proteases, **mutations aiming at modifying the substrate binding pocket in order to fit a new substrate are most often strongly deleterious for catalytic activity towards all tested substrates**, including the initial substrate (*2–4*). In the rare cases where targeted substrate specificity conversion was achieved, the underlying mechanism may not be the expected one (*1*, *5*). Namely, the binding affinities of the initial and the engineered enzyme for the new substrate are found to be similar, even though structure-guided engineering aimed at increasing enzyme binding affinity to the new substrate.

The remarkable success of the conversion of trypsin's specificity into that of chymotrypsin required to swap residues located in surface loops distal from the active site, in addition to more structurally obvious residues located in the substrate binding pocket (*5*). These distal sites were not predicted by structural analysis but rather through a statistical comparison of the handful of sequences of trypsin vs chymotrypsin across phylogeny available at the time. However, this remarkable achievement has remained largely limited in scope, as all attempts to engineer reciprocally trypsin specificity in chymotrypsin by performing the reverse mutational swap failed (*6*).

These observations motivate our approach that does not depend on structural information but rather consists in a systematic scan of sequence space.

Statistical analysis and modeling of protein sequences for protein design.

On the computational side, **our project builds on an approach inspired by statistical physics that considers the sequence-function relationship in sequence space** rather than in the physical space of protein structures. Developed by MW and others, **statistical models based on abundant genomic data (homologous protein families) yield impressive predictions on protein structure and function from sequence data alone** *(7)*. Namely, MW's Direct Coupling Analysis (DCA) predicts amino-acids in contact in the tridimensional structure from sequence data alone, and the effect of amino-acid mutations. A recent achievement by MW's group is the computational generation of sequences of artificial proteins that rescue *in vivo* the deletion of an essential metabolic enzyme gene *(*8*)*, an example of the **emerging paradigm of statistical protein design** that we propose here to extend to the design of specificity (see Fig. 1). For this project, MW's group provides new efficient techniques for generative model learning *(*9*)*.

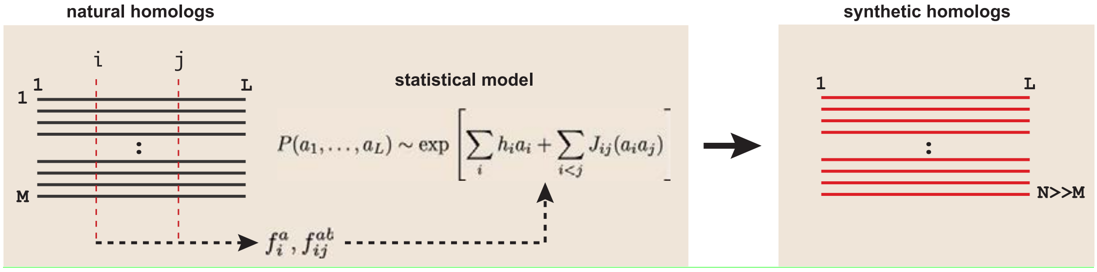
natural homologs
i
j
1
1
L
M
statistical model
$P(a_1,\ldots,a_L) \sim \exp\left[\sum_i h_i a_i + \sum_{i<j} J_{ij}(a_i a_j)\right]$
f_i^a, f_{ij}^{ab}
synthetic homologs
1
L
N>>M

Figure 1: From statistical sequence models to protein design* (8*). *A multiple sequence alignment of homologous proteins is used to estimate 1-point statistics (amino-acid conservation at each position) and 2-point statistics (correlated amino-acid usage in pairs of positions), and to infer a DCA model whose parameters describe specific amino-acid biases (h_i) and two-site couplings (J_{ij}). This generative model can be sampled to create synthetic sequences of proteins with functional features identical to those of the input natural proteins, as demonstrated in the case of chorismate mutase enzymes (*8*)*.

Large-scale enzyme sequence-function experimental data.

Our physics-biology interdisciplinary experiments combine droplet microfluidics and molecular biology to measure the enzymatic activity of *hundreds* of *protease* enzymes. Mutational scans rely on saturated mutagenesis, a functional assay and high-throughput DNA sequencing read-out to map the effect of every point mutation in a protein sequence on its function *(*10*)*. **We propose to extend this approach to a large-scale functional exploration of sequence space of the S1A protease family** (see Fig. 2A) instead of focusing on the neighborhood of a particular sequence, in order to obtain optimal datasets for statistical approaches that have been developed mainly on genomic data.

Our experiments will harness **commercially available large-scale gene synthesis**, which can produce hundreds of genes of arbitrary sequence and was already put to use in the enzyme design work by MW and colleagues *(*8*)*.

Contrary to most functional assays for enzymes performed *in vivo* (8), **droplet microfluidics technology provides high throughput *in vitro* enzymatic assays that directly probe enzymatic properties *(*11*)***. Droplet microfluidics relies on the encapsulation of individual enzyme genes from a library in micron-size (picoL) droplets. The enzyme library genes are then expressed in droplets and their functional properties are assessed by addition of fluorogenic enzymatic substrates by pico-injection *(*12*)*, incubation in delay lines, and a fluorescence readout. Droplets are then sorted according to this readout by dielectrophoresis *(*13*)*. Encapsulation, pico-injection, readout and sorting occur at up to kHz rates, allowing the screening of libraries of >10^6 genes in 1 day *(*11*)*. **CN's group has developed a droplet microfluidic setup to quantitatively assay thousands of protease enzymes against two substrates in a single experiment (see preliminary results in Fig. 2B&C).**

High-throughput DNA sequencing technologies have been rapidly evolving over the past 15 years, with currently two main categories available: 10^7-10^{10} reads of 50-200bp at a 10^{-4}bp^{-1} error rate (provided by Illumina) vs 10^6 reads of <10kb with a 10^{-2} error rate (provided by Nanopore). Our workflow makes use of both types of sequencing to analyze the output of microfluidics enzyme library sorting (see preliminary results in Fig. 2B&C).

- **Objectives**:

Our interdisciplinary project will run through 3 successive objectives:

Objective 1: The PhD candidate will use the existing workflow in CN's group (Fig. 2) **to measure the catalytic activity of 200 natural S1A proteases** identified with MW to cover as widely and uniformly as possible sequence diversity in this enzyme family, against 4 peptide substrates differing by a single amino-acid and a protein substrate (FITC-casein), to probe respectively **specific vs non-specific protease activity**. This objective is **low risk** considering current know-how in CN's and MW's groups.

Objective 2: From the experimental data, the PhD candidate will learn under MW's guidance, and in close collaboration with his team members, **how specificity is encoded in the sequence through statistical modeling**. The generative model, built upon a generalization of Trinquier et al (*9*) to the case of **semi-supervised learning** using data partially annotated in Objective 1, will be tested by proposing **100 computationally designed synthetic sequences** with predicted substrate specificity.

Objective 3: The PhD candidate will measure under CN's supervision the catalytic activity and specificity profile of the 100 computationally designed enzymes to **validate the model predictions**.

We stress that the experimental data obtained from Objective 1 can be exploited in many other ways besides Objectives 2 and 3, which provides considerable flexibility over the course of the PhD thesis.

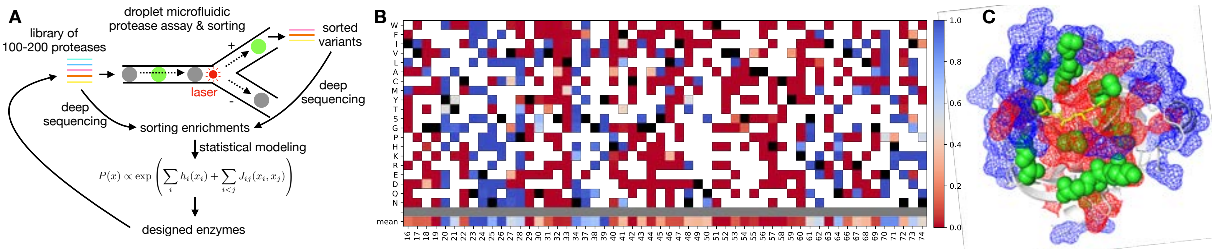

Figure 2. Workflow and preliminary results. A*. In our project****, 200 proteases (natural homologs)*** *will be expressed and assayed individually in* ***microfluidic droplets*** *with a fluorescence readout and* ***sorted according to detected activity.*** *Deep sequencing of input and output genes yields sorting enrichments that are used to infer statistical models and computationally design proteases. Activity of 100 designed proteases will be tested with the same experimental workflow.* ***B****. Preliminary results on the enrichment of >4000 1-point mutants of rat trypsin upon droplet microfluidic sorting for protease activity towards a Lys substrate (x-axis=sequence [shown=Nter], y-axis=mutation, red=deleterious, blue=neutral/beneficial).* ***C****. Sensitive residues (red) are mainly in the structural core and close to the substrate (yellow), robust residues (blue) are in the periphery. Residues that bias activity towards Arg vs Lys (trypsin natural P1 substrates) are located in-between (green), either close or distal to the substrate.* ***Our results are consistent with and extend beyond the literature, and validate the experimental workflow for this project.***

- **Candidate profile**:

Our project requires a **PhD candidate with a physics/engineering background** for droplet microfluidics experiments and statistical physics models (hence the relevance of EDPIF), together with a strong motivation to learn concepts and techniques of experimental and computational biology. The PhD candidate will mainly carry out microfluidic/molecular biology experiments in CN's group, and take part to modeling with the support of other members of MW's group.

- **References**:

1. L. Hedstrom, *Chemical reviews*. 102, 4501–4524 (2002).
2. C. S. Craik *et al.*, *Science*. 228, 291–297 (1985).
3. L. Gráf *et al.*, *Biochemistry*. 26, 2616–2623 (1987).
4. L. B. Evnin, J. R. Vasquez, C. S. Craik, *PNAS*. 87, 6659–6663 (1990).
5. L. Hedstrom, L. Szilagyi, W. Rutter, *Science*. 255, 1249–1253 (1992).
6. I. Venekei, L. Szilagyi, L. Gráf, W. J. Rutter, *FEBS letters*. 379, 143–147 (1996).
7. S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, M. Weigt, *Reports on progress in physics.* 81, 032601 (2018).
8. W. P. Russ *et al.*, *Science.* 369, 440–445 (2020).
9. J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, M. Weigt, *Nat Commun*. 12, 5800 (2021).
10. D. M. Fowler, *Nature methods*. 11, 801–807 (2014).
11. J. J. Agresti *et al.*, *PNAS*. 107, 4004–4009 (2010).
12. A. R. Abate, T. Hung, P. Mary, J. J. Agresti, D. A. Weitz, *PNAS*. 107, 19163–19166 (2010).
13. J.-C. Baret *et al.*, *Lab on a chip*. 9, 1850–1858 (2009).