Research Statement
Khanh Dao Duc
University of California, Berkeley

My main interest in research is to find and solve problems related to fundamental biological processes, using mathematical and computational methods. For the past years, developments in sequencing, imaging and many other techniques have indeed transformed the field of biology, calling for quantitative approaches. During my PhD I have worked on a broad range of topics in cellular biology and neuroscience, which required solving theoretical problems in stochastic processes, PDE and dynamical systems. After developing expertise in mathematical modeling and analysis, I decided to extend my research to more computational and statistical approaches to tackle challenging biological data. During my postdoc, I consequently focused on the regulation of protein translation and developed statistical methods to infer translation dynamics from deep sequencing data. The results that I obtained led me further to investigate other fields of research in structural biology, evolution and theoretical physics, by studying the biophysical properties and evolution of ribosomes from cryo-EM data, and extending the theory of particle transport process. Following a summary of my main contributions so far, I will detail my proposed research, centered on developing methods for a multi-scale modeling and study of gene expression dynamics in single cell, using heterogeneous and large datasets.

1 Past research accomplishments

Quantifying the impact of stochasticity in biological processes

From diffusion driving molecular motion to mutations impacting evolution, randomness plays an important role in biology at various scales. My PhD was mainly focused on understanding this role, for three types of processes relevant in biological contexts (see Fig.1).

First, I focused on stochastic chemical reactions involving a small number of particles, and derived formulas for the probability and the conditional time to reach a threshold of bindings [4]. I notably applied them to quantify the control of chromosome alignment during cell division [5] and study post-transcriptional regulation of mRNA (Fig.1a)[1].

Secondly, I studied the behavior of a class of neurons called up-and-down state neurons (Fig.1b), involved in working memory and attention [6]. Their electric potential can be modeled by a dynamical system, with noise-induced transition between two attractors [7]. By solving the Dirichlet problem for the Fokker-Planck operator [8, 9], I explained the presence of oscillatory peaks observed in the distribution of residence time in up-state [2]. Using mean-field approaches, I also modeled, in collaboration with Menahem Segal (Weizmann institute) and Nathalie Rouach (College de France) the bursting dynamics of small neuron networks [10], and the dynamics of postassium [11] to study the role of glial cells in neuronal excitability.

Finally, I studied cellular aging, quantified by the number of divisions after which a cell ceases to divide. The process (called senescence) is regulated by the length of telomeres, which are regions of repetitive sequences at each end of chromosomes. Telomeres get shortened during DNA replication and randomly elongated by a complex called telomerase (Fig.1c). In collaboration with Teresa Teixeira (IBPC, Paris), I developed a stochastic model of telomere length, as a drift and jump process. By deriving its first order statistics [12], I showed that the shortest telomere in a cell was the major determinant of senescence [3].
Analyzing protein translation dynamics from deep sequencing data

Proteins are synthesized by ribosomes that translate codons into amino acids, while moving along mRNA transcripts (see Fig.2). Various factors can influence this fundamental process, but the observed heterogeneity of ribosome density along transcript sequences remains only partially explained [13]. To study the determinants of translation dynamics, I introduced a method to infer the position-specific speed from ribosome density information [15]. The method is based on a mathematical model of the ribosome traffic along a given mRNA, derived from a classical particle process called the totally asymmetric exclusion process (TASEP) [14]. After applying this method to experimental data (Fig.2b), I showed that the distribution of codons was not sufficient to explain the observed variation of ribosome dynamics, as claimed previously [16]. By using a statistical model, I found these variations being instead largely explained by biophysical properties of the translated polypeptide, such as the amount of electric charges and the hydrophobicity, in specific windows located around a given site.

We hypothesized that these features were selected because of interactions inside a micro-domain of the ribosome, called the ribosome exit tunnel (illustrated in Fig.2c). Upon extracting the structure of the tunnel from cryo-EM data and expanding the Fick-Jacobs equation [17] to describe the dynamics of a charged particle, I showed that our statistical analysis was in agreement with the estimated impact of the geometry and the electric field created by the ribosome structure.

Developing the theory of exclusion processes, with application to translation

The data used in our study is obtained from a high-throughput sequencing technique called ribosome profiling [18]. One notable issue is that in general, mRNA fragments of specific length are filtered, so the observed density may exclude stacked ribosomes [19]. This may lead to a bias that needs to be corrected when evaluating the ribosome density. To do so, I extended the treatment of the TASEP [20], describing the traffic of particles along a lattice, to
compute the number of collisions and mean distance between particles. Applying these results to experimental data, I found a bias against the detection of nearby ribosomes with gap distance less than 3 codons, in agreement with measurements of larger sequenced fragments [19].

In contrast with classical studies of the TASEP, which mostly consider homogeneous rates, ribosome profiles also indicate large heterogeneity in the dynamics at codon level. In another study [21], I thus generalized the analysis of the TASEP to the heterogeneous case and for particle of arbitrary size, in a continuous limit obtained when the lattice becomes large. This limit (so-called hydrodynamic) gives rise to a second order PDE, which I analyzed to obtain a phase diagram describing the different regimes of particle traffic. I am currently applying these results to improve inference.

2 Future research plans

As highlighted by the Human Cell Atlas or GTEx projects, the emergence of various types of large datasets has given rise to exciting challenges in providing a quantitative description of a single cell [22]. A particular and timely focus is the regulation of gene expression, as recent and massive developments in sequencing and imaging techniques provide an unprecedented opportunity to study both transcription and translation. From my current research, centered on the determinants of translation at a single mRNA level, I now plan to bridge the gap with other scales, and orient my research towards two complementary projects (detailed below): 1) how does the molecular structure of the ribosome impact gene expression, and 2) how to model protein synthesis at the whole-cell level. A better understanding of these problems would notably help to answer fundamental, but yet still open questions: How genetic and molecular variations can differentially influence mRNA vs. protein levels [23]? What is the molecular basis for macrolide antibiotic resistance [24], or sequence recognition that triggers translocation to specialized compartiments of the cell [25]? Answering these questions require a fine understanding of the biophysics involved from atomic to genome-wide scales, and going beyond the use of classical methods in biostatistics and systems biology. I am confident that my background as an applied mathematician, familiar with methods from physics, computational biology and data analysis, will allow me to tackle these problems.
Figure 3: **Comparative analysis of the ribosome tunnel.** Cryo EM data of ribosomes are used to extract the key features of the tunnel, inferring phylogeny, and model the interactions with the translated polypeptide chain during translation. Image from [28].

**Impact of the evolution and molecular structure of the ribosome on gene expression**

The development of cryo-EM, recognized by the last Nobel Prize in chemistry, has revolutionized the field of structural biology. In particular, this technique has led for the past few years to the discovery of an increasing number of ribosome structures at high resolution, unraveling different mechanisms involved in translation, folding and stalling of the polypeptide chain [26]. I plan to take advantage of the availability of these structures to perform a comparative analysis of the ribosome. As a preliminary work [28], I have compiled more than 40 recently obtained structures, accounting for species coming from bacteria, archaea, and eukarya. After extending previously used methods, I isolated from phylogenetic inference the conserved geometric features of the tunnel (see Fig.3). As I previously established the importance of electric charges in the polypeptide, I will also use advanced numerical methods to solve the Poisson Boltzmann-equation satisfied by the electric potential and compare the results obtained between all the structures [29] (in collaboration with Teresa Head-Gordon, UC Berkeley).

Extracting the shape and properties of the tunnel will allow me to properly model translation dynamics at molecular level. Molecular dynamics simulations have revealed important insights on stalling sequences for example, but are limited to nanoseconds [30]. On the other hand, latest coarse grained models work in an oversimplified geometry and environment [31]. My goal is thus to provide a realistic description of the ribosome at a time scale sufficient to predict and analyze the dynamics of a polypeptide, but also other molecules such as macrolide antibiotics, which target the ribosome core. The specificity of the geometry and conditions will give rise to new theoretical problems in polymer translocation [27], formulated as first passage time problems which I am familiar with through my past experience in David Holcman’s lab. In combination with bioinformatic tools, this framework will also be useful for important problems such as the translation of membrane protein genes (involving the recognition of specific motifs at the tunnel exit), deciphering the impact of the tunnel geometry on antibiotic resistance, and finding signals of co-evolution with the proteome.

**Whole-cell modeling of protein synthesis**

At the cell level, the maintenance of the ribosome population form a large part of the protein synthesis machinery, ribosomal RNAs and proteins absorbing some 70% of all transcription
Improved protein homeostasis, implying optimal ribosomal function, is associated with disease resistance and increased lifespan [33]. It is thus essential for a cell to control its ribosome population, through different mechanisms that in return tune protein synthesis. Previous studies have modeled the availability of free ribosome in a whole cell to study the limiting factors of protein synthesis and cell growth [34, 35]. However, these models only offer a partial description of their dynamics, neglecting transcriptional and translational regulating events.

My goal is thus to build a biologically sound mathematical model, able to describe protein synthesis at a whole cell level and integrate transcriptional and ribosome population dynamics. To set the parameters, I will notably use my previous results [15], and other recent studies that quantify the kinetics of ribosome assembly [36] and mRNA degradation [37]. An important application of the model will be to understand how specific perturbations of the ribosomal protein pathways can possibly explain in cancer cells the differential fold changes observed in translation efficiency, affecting transcription and ribosome genes [38]. As the maintenance of ribosomes implies a huge metabolic cost, it will also be interesting to study how the model can achieve some optimality, and the importance of ribosome “recycling” associated with mRNA looping [39], and see how it agrees with theoretical analysis of polymer looping time, notably studied in my former lab [40]. While in a first approach, considering homogeneous deterministic population will allow a better theoretical understanding, I also plan to include spatial and stochastic dynamics, notably associated with cell compartmentalization, diffusion, transcriptional bursting [41] and exclusion processes describing translation and transcription at individual level. To make the model tractable and able to simulate hundreds of thousands of particles, an interesting challenge will be to reduce the state space by mapping the TASEP into an appropriate and simpler queuing process, similar to models I used in previous research [4].

Current and potential collaborations

At the qb3 institute at Berkeley and Yun Song’s lab, I have had the chance to work in a strong interdisciplinary environment, and have several graduate and undergraduate students involved in different parts of my research. In particular, I am currently closely mentoring three of Yun Song’s graduate students (two in mathematics and one in computational biology), who are much helpful in pushing the research towards new interesting directions. My current work has also generated collaborations with Teresa Head-Gordon in computational chemistry (UC Berkeley), and Greg Huber in biophysics (CZ Biohub).

My participation to several conferences for the past years has also allowed me to establish links abroad, like in Germany, as I was recently invited to a summer school organized by a focused research group1. I plan to pursue my efforts in building valuable international connections, by notably organizing (with other researchers from Europe, United States and Canada) a workshop next fall, centered on mathematical methods in single cell, as a part of a 6 weeks program in Pisa, Italy2. The workshop will gather physicists, mathematicians and biologists from prestigious international institutions, and I am hopeful it will generate new research collaborations and approaches to deal with these challenging problems I am interested in.

1DFG-Forscherguppe 1805 : Ribosome dynamics in regulation of speed and accuracy of translation (http://www.dfgerice2017.com/).
2http://www.crm.sns.it/event/425/
References


[3] Z. Xu, K. Dao Duc, D. Holcman, T. Teixeira (2013), The length of the shortest telomere as the major determinant of the onset of replicative senescence, Genetics 194 (4)


[21] D. Erdmann-Pham, K. Dao Duc, Y. S. Song, Hydrodynamic limit of the TASEP with open boundaries, heterogeneous rates and extended particles (submitted)


[38] Arun Wiita (UCSF), unpublished data.

