# REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)
03-07-2018

2. REPORT TYPE
Final Report

3. DATES COVERED (From - To)
25-May-2016 - 29-Apr-2017

4. TITLE AND SUBTITLE
Final Report: 11.1 STIR: Multi-level Hidden Markov Model for Co-translational Protein Targeting

5a. CONTRACT NUMBER
W911NF-16-1-0286

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Harvard University
Office for Sponsored Programs
1033 Massachusetts Ave 5th Floor
Cambridge, MA 02138 -5369

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
67762-MA-II.2

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

| a. REPORT | b. ABSTRACT | c. THIS PAGE |
|---|---|---|
| UU | UU | UU |

17. LIMITATION OF ABSTRACT
UU

15. NUMBER OF PAGES

19a. NAME OF RESPONSIBLE PERSON
Shingchang Kou

19b. TELEPHONE NUMBER
617-496-8423

# RPPR Final Report
as of 05-Jul-2018

Agency Code:

Proposal Number: 67762MAII **Agreement Number: W911NF-16-1-0286**

**INVESTIGATOR(S):**

**Name:** Shingchang Samuel Kou
**Email:** kou@stat.harvard.edu
**Phone Number:** 6174968423
**Principal:** Y

Organization: **Harvard University**
Address: Office for Sponsored Programs, Cambridge, MA 021385369
Country: USA

| | |
|---|---|
| DUNS Number: 082359691 | EIN: 042103580N |
| **Report Date:** 29-Jul-2017 | Date Received: 03-Jul-2018 |

**Final Report** for Period Beginning 25-May-2016 and Ending 29-Apr-2017
**Title:** 11.1 STIR: Multi-level Hidden Markov Model for Co-translational Protein Targeting

| | |
|---|---|
| **Begin Performance Period:** 25-May-2016 | **End Performance Period:** 29-Apr-2017 |
| **Report Term:** 0-Other | |
| Submitted By: Shingchang Kou | Email: kou@stat.harvard.edu<br>Phone: (617) 496-8423 |

**Distribution Statement:** 1-Approved for public release; distribution is unlimited.

**STEM Degrees:** 1 **STEM Participants:** 1

**Major Goals:** The primary research goal of the project is to build mathematical models to investigate the unsolved questions regarding the molecular mechanism of the co-translational protein targeting process. The specific objectives are as follows. (a) Build a novel multi-level hidden Markov model for the conformational dynamics of the molecular complexes that govern the protein targeting process. This multi-level mathematical modeling allows to combine and extract information from multiple proteins for a deeper understanding of the molecular mechanism. (b) Employ the multi-level hidden Markov model to account for the stochastic heterogeneity among the multiple proteins, which is necessary for detailed molecular understanding of the co-translational protein targeting process. (c) Apply the multi-level hidden Markov model to recent single-molecule experimental data on bacterial protein targeting systems. Direct fitting to the biological data not only gives the experimental assessment of the model, but also allows to extract experimental information in combination with our mathematical model to elucidate the detailed molecular mechanism of protein targeting.

**Accomplishments:** A major mathematical challenge that makes it difficult to model and analyze the experimental data, which consist of hundreds of data trajectories, each of which is a time series that corresponds to the conformational change of one protein complex, is that the data trajectories are probabilistically heterogeneous, meaning that they cannot be regarded as coming from the same stochastic process.

Owing to this difficulty, one cannot use the conventional likelihood theory (including maximum likelihood and traditional Bayesian statistical methods), which is built on the assumption that the trajectories are identically distributed, to study the heterogeneous experimental data. To bypass the difficulty, researchers previously have tried to analyze the data trajectories one by one, and then overlay the results on top of each other (e.g., through histograms) to put them together. But this is very inefficient, because as each data trajectory is handled in solo it does not allow any information sharing between the different data trajectories. We introduced a novel multi-level hidden Markov model (HMM), and the multi-level structure enables us to account for the stochastic heterogeneity of individual proteins (as revealed in their respective data trajectories) as well as to efficiently combine and extract all the information from the multiple proteins. Specifically, in our hierarchical HMM, the global stochastic characteristics at the top layer generates the individual stochastic characteristics of each HMM that models each protein at the middle layer. The individual stochastic characteristics at the middle layer controls the evolution of the respective Markov chain at the lower layer, which in turn generates the observable through the HMM observational equation. The individual HMMs at the middle layer are linked at the top layer. This hierarchical structure enables us to combine all the information from all the protein complexes to accurately estimate/infer the global stochastic characteristics, the estimate of which, in turn, provides better estimate of the individual characteristics. This way of conducting inference for the individual characteristics is much more efficient in that we are able to combine the information from all the protein complexes, instead of doing the inference on the individual protein complexes one by one.

Before our multi-level hidden Markov model was introduced, researchers have tried to analyze the data trajectories one by one, owing to the inherent stochastic heterogeneity (as the conventional likelihood theory cannot be applied to combine the data trajectories). Such method is very inefficient, as each data trajectory is handled in solo. Our multi-level structure enables us to account for the stochastic heterogeneity of individual data trajectories and at the same time to efficiently combine and extract all the information from the hundreds of data trajectories. Consequently, we are able to improve the estimation efficiency by up to two orders of magnitude.

Applying our multi-level hidden Markov model as diagrammed above to the single-molecule experimental data on the prokaryotic (bacterial) system, we are able to delineate the detailed mechanistic steps of protein targeting.

Step 1. SRP (Signal Reconition Particle) recognizes the signal sequence on RNC (ribosome-nascent-chain complex) and binds it. The RNC is delivered to the target membrane where the SR (Signal Receptor) can localize to.
Step 2. When the SRP-SR complex is initially formed, the FtsY complex binds at the RNA capped end near the ribosome exit site, blocking the site from translocon binding.
Step 3. As the RNC initiates contact with the translocon, the latter actively facilitates the conformation change of SRP-SR complex and drives the FtsY complex from the capped end to the distal end of RNA.
Step 4. GTP-hydrolysis is initiated at the RNA distal end to disassemble the SRP and SR. Meanwhile, the nascent chain is released from the Ffh M-domain to the translocon on the membrane.

**Training Opportunities:** A Ph. D. student Ms. Yang Chen worked on and was funded by the project. Ms. Yang Chen graduated from Harvard University in May 2017. A significant part of her thesis is based on the project. She is currently an assistant professor of statistics at the University of Michigan.

**Results Dissemination:** Nothing to Report

**Honors and Awards:** A Ph. D. student Ms. Yang Chen working on and funded by the project got the “NESS IBM Student Award,” which was given to the best student papers at NESS (New England Statistical Symposium). Yang Chen graduated from Harvard University in 2017; a significant part of her thesis is based on the project. She is currently an assistant professor of statistics at the University of Michigan. Her job talk primarily focused on the result of this project.

**Protocol Activity Status:**

**Technology Transfer:** Nothing to Report

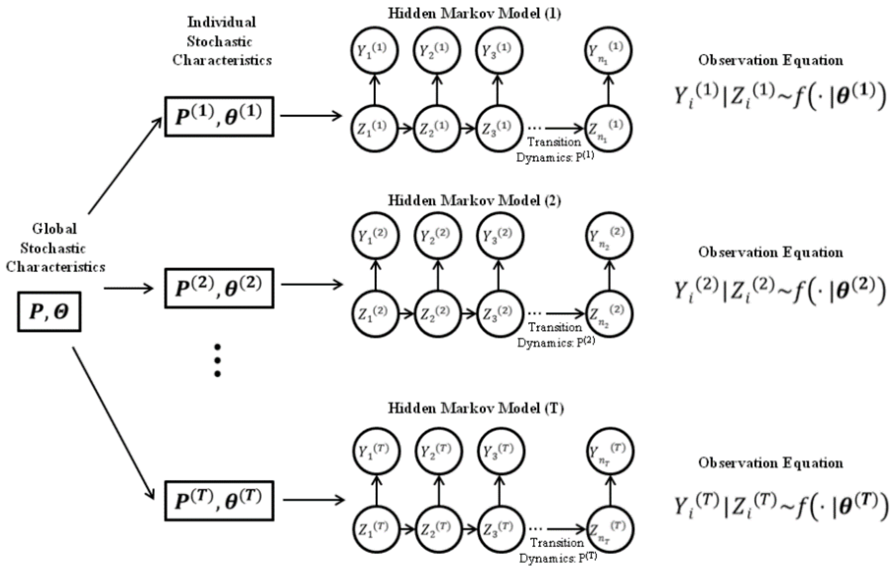
**PARTICIPANTS:**

**Participant Type:** Graduate Student (research assistant)
**Participant:** Yang Chen
**Person Months Worked:** 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

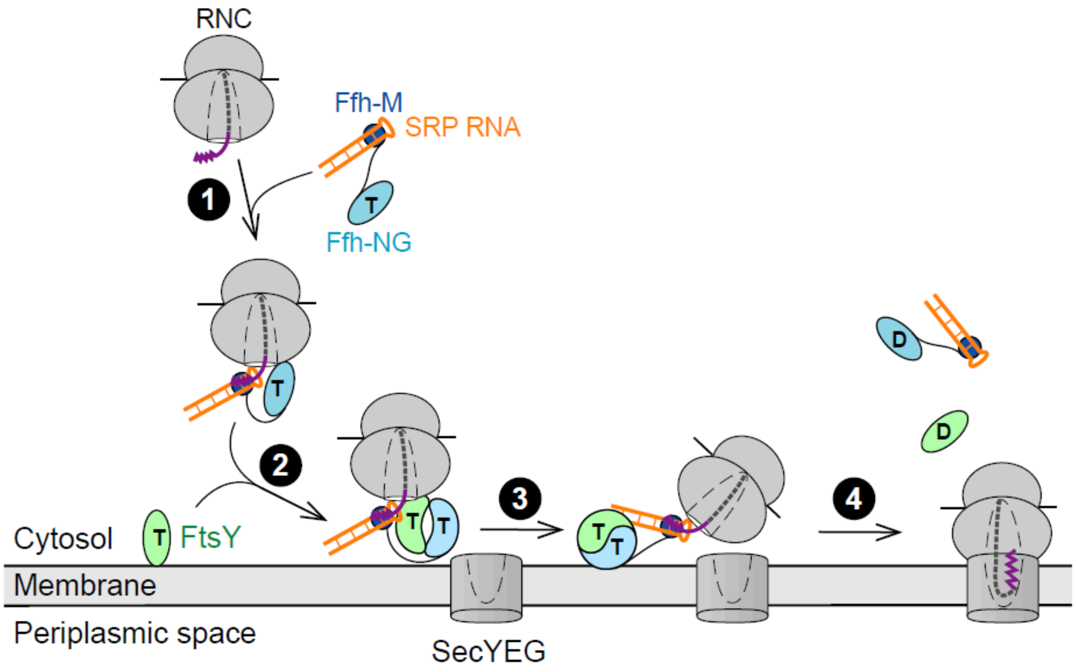
**Participant Type:** PD/PI
**Participant:** Samuel Kou
**Person Months Worked:** 9.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N

Other Collaborators:

1. An illustration of our hierarchical hidden Markov model (HMM):


Individual Stochastic Characteristics
Global Stochastic Characteristics
$\boldsymbol{P}, \boldsymbol{\theta}$
$\boldsymbol{P}^{(1)}, \boldsymbol{\theta}^{(1)}$
$\boldsymbol{P}^{(2)}, \boldsymbol{\theta}^{(2)}$
$\boldsymbol{P}^{(T)}, \boldsymbol{\theta}^{(T)}$
Hidden Markov Model (1)
$Y_1^{(1)}$ $Y_2^{(1)}$ $Y_3^{(1)}$ $Y_{n_1}^{(1)}$
$Z_1^{(1)}$ $Z_2^{(1)}$ $Z_3^{(1)}$ $Z_{n_1}^{(1)}$
Transition Dynamics: $P^{(1)}$
Observation Equation
$Y_i^{(1)} | Z_i^{(1)} \sim f(\cdot | \boldsymbol{\theta}^{(1)})$
Hidden Markov Model (2)
$Y_1^{(2)}$ $Y_2^{(2)}$ $Y_3^{(2)}$ $Y_{n_2}^{(2)}$
$Z_1^{(2)}$ $Z_2^{(2)}$ $Z_3^{(2)}$ $Z_{n_2}^{(2)}$
Transition Dynamics: $P^{(2)}$
Observation Equation
$Y_i^{(2)} | Z_i^{(2)} \sim f(\cdot | \boldsymbol{\theta}^{(2)})$
Hidden Markov Model (T)
$Y_1^{(T)}$ $Y_2^{(T)}$ $Y_3^{(T)}$ $Y_{n_T}^{(T)}$
$Z_1^{(T)}$ $Z_2^{(T)}$ $Z_3^{(T)}$ $Z_{n_T}^{(T)}$
Transition Dynamics: $P^{(T)}$
Observation Equation
$Y_i^{(T)} | Z_i^{(T)} \sim f(\cdot | \boldsymbol{\theta}^{(T)})$

2. Applying our multi-level hidden Markov model as diagrammed above to the single-molecule experimental data on the prokaryotic (bacterial) system, we are able to delineate the detailed mechanism of protein targeting, as showing in the following illustration:


RNC
Ffh-M
SRP RNA
T
Ffh-NG
1
T
2
T
T
3
T
T
4
D
D
Cytosol
T
FtsY
Membrane
Periplasmic space
SecYEG