Advanced Core Technologies for Big Data Integration

Foundations of Innovative Algorithms for Big Data

Naoki Katoh (Kwansei Gakuin Univ.)

DATAIA – JST INTERNATIONAL SYMPOSIUM ON DATA SCIENCE AND AI

July 11th, 2018

- For a huge size of big data, polynomial time algorithm paradigm becomes obsolete.

- So, we propose a new paradigm

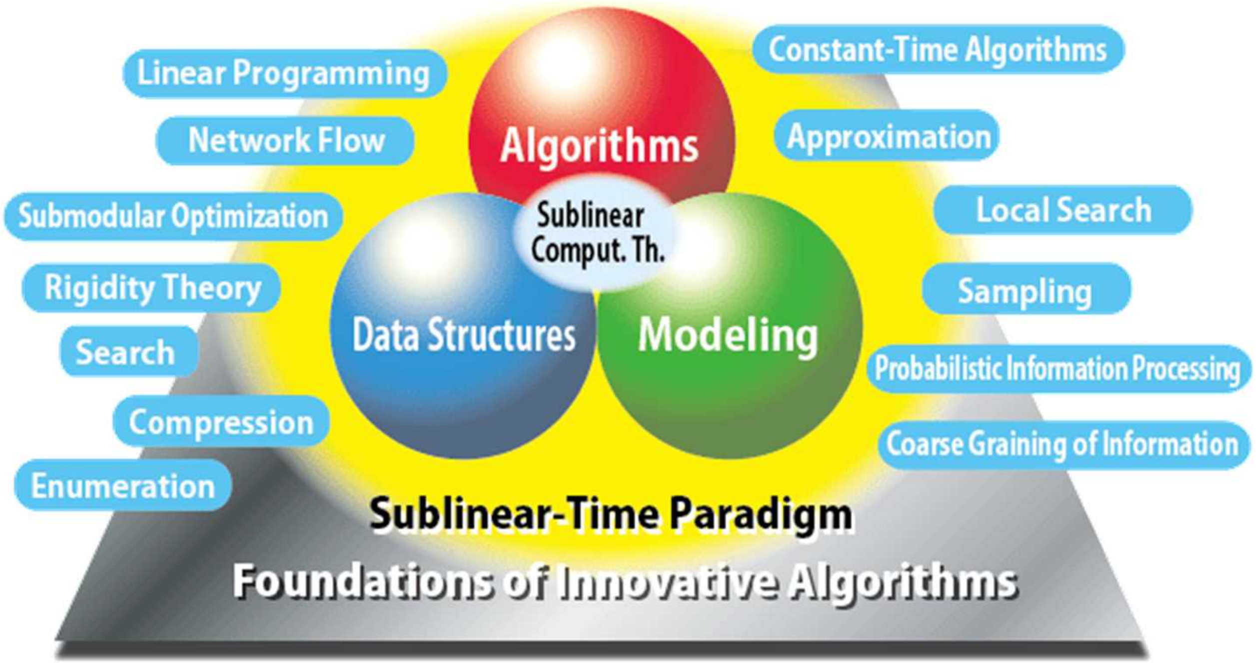
Sublinear Time Paradigm

Run time / Data size	$\log n$ Binary search	$\sqrt{n}$	n Linear search	$n \log n$ Sorting	n^2 Shortest path	n^3 Max flow
$n=1000$	10	30	1000	1000	10^6	10^9
$n=10^6$	20	1000	10^6	2×10^7	10^{12}	10^{18} ≒317 years
$n=10^9$(# of Web servers)	30	30000	10^9	3×10^{10} ≒300 Sec.	10^{18} ≒317 years	
$n=10^{12}$	40	10^6	10^{12} =10000 Sec.	4×10^{13} ≒111 hours		
$n=10^{15}$	50	3×10^7	10^{15} 10^7 Sec.≒110days	5×10^{16} ≒5500 days		
$n=10^{18}$	60	10^9 ≒10 Sec.	10^{10} Sec ≒317years.	6×10^{19} ≒19000 years		

(※)Assumption: 10^8 calculations need 1 second.

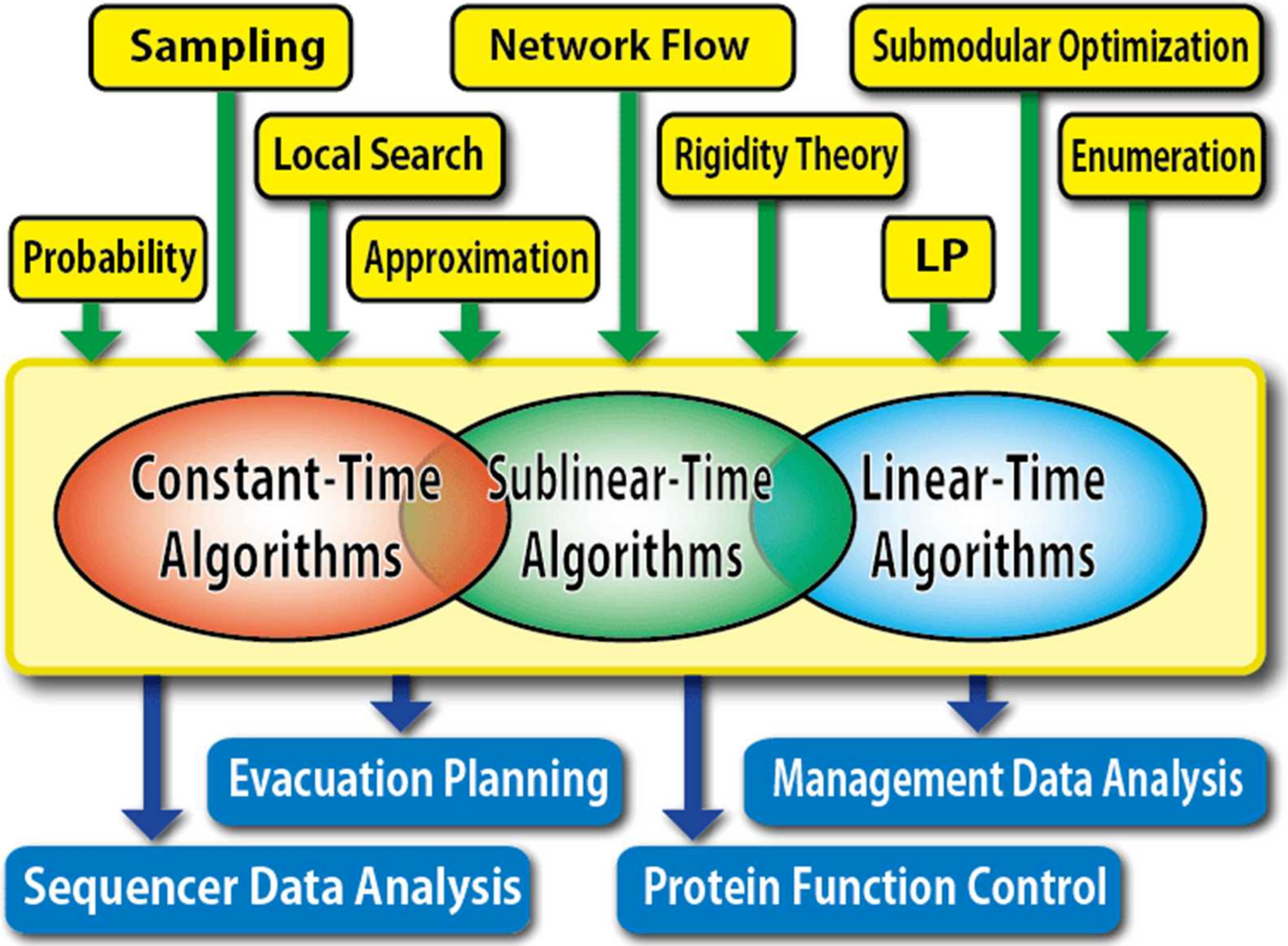
Project Overview

Foundations of Algorithm Theory for BIG DATA



1. Foundations of Sublinear Time Algorithm (Katoh group)
2. Foundations of Sublinear Data Structures (Shibuya group)
3. Foundations of Sublinear Modeling (Tanaka group)

Team A: Sublinear Time Algorithm approach (Katoh group)



Major results of Team A (Katoh Group)

A1. Constant time algorithm for complex network

A2. Protein function analysis by combinatorial rigidity theory

What is a constant-time algorithm?

How do we evaluate BIG DATA, e.g. web graphs?

Web Graph

Read only **O(1)** data

Approximate properties or parameters (e.g., connectivity, cluster coefficient, average node-to-node distance, page rank, …)

The base technology for the big-data era!

Two paradigms in the big-data era: Constant-time and polynomial-time

Physics:

- Newtonian mechanics (17st --): necessary for normal physical calculation.
- Theory of relativity (20st--): necessary for the ultrahigh-speed situation.

Both are necessary

Analogy

Computing:

- **Polynomial-time algorithms (1950's --): necessary for normal computing.**
- **Constant-time algorithms (1990' --): necessary for the big-data era.**

Both are necessary

Constant-time algorithms for complex networks: background and main result

sampling

A traditional technique for treating big-data
(e.g., statistics, machine learning)

Apply it to complex problems, e.g. graphs

Constant-Time Algorithms

Appeared on 1990's
[Rubinfeld & Sudan, SODA92][Goldreich et. al, FOCS95]

- Evaluate properties of given graphs very fast through sampling
- Efficiency is proven theoretically

Constant-time algorithms are developed rapidly in 21st. But...
We have been able to treat dense or bounded degree graphs only.

In our project, we extend the testable graph class to
a wider class including **complex networks** [Ito, ESA2016]

Major Results of Team A

Algorithms

A1. Constant Algorithms for Complex Networks

A2. Protein function analysis by combinatorial rigidity theory

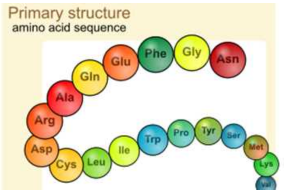
Protein Function Analysis by Combinatorial Rigidity Theory

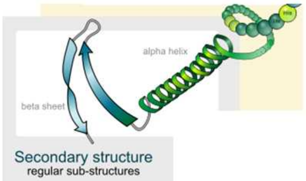
Why study protein flexibility and dynamics?

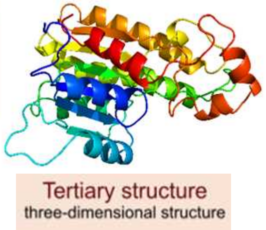
Sequence → Structure → Flexibility and Dynamics → Function

EASY

VERY HARD







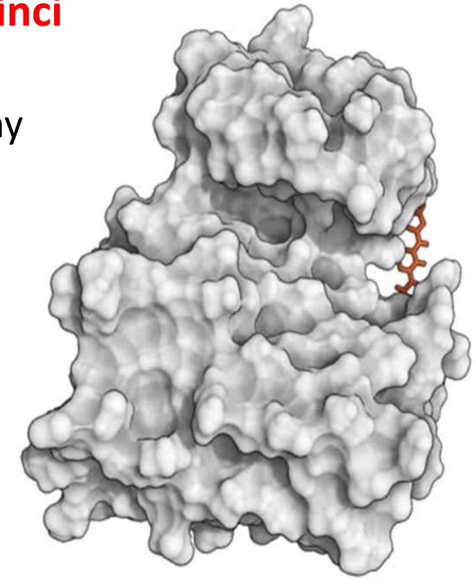
"Movement is the cause of all life."

Leonardo da Vinci

RCSB PDB PROTEIN DATA BANK

X-ray Crystallography

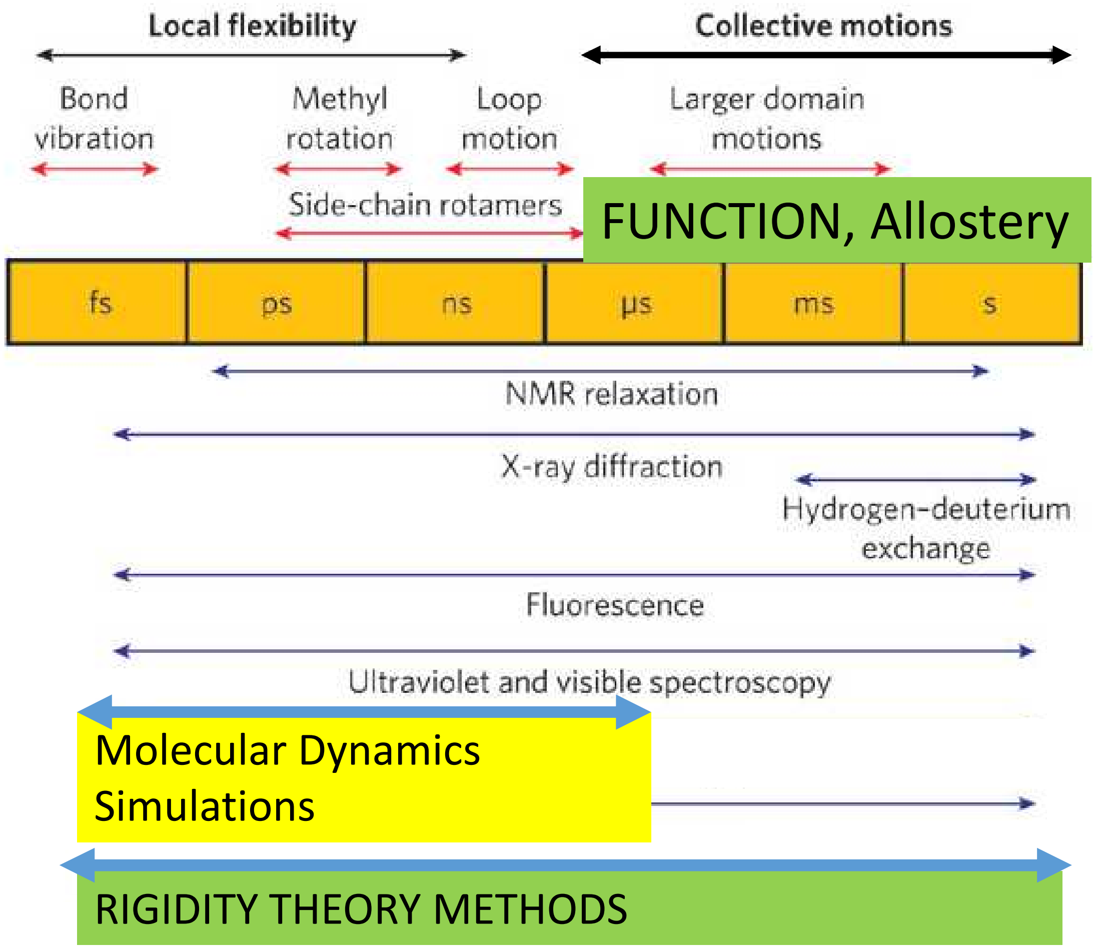
NMR

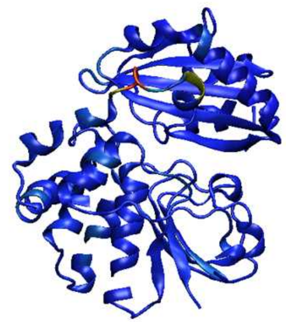


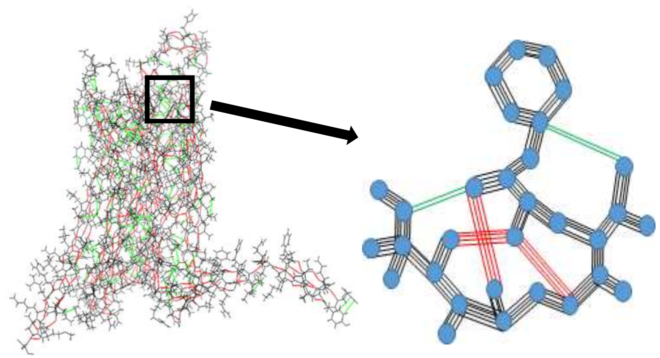
- Flexibility/rigidity critical for protein function
- Experiments are expensive, slow and give limited information
- Molecular Dynamics is too slow and often impractical
- **Rigidity theoretical approach fastest, can also be used to speed up simulations**

(Plos One Sljoka et al 2015)

Key difficulties in studying protein flexibility: Proteins motions occur on many time scales







Major Results

1. The Role of Dimer Asymmetry and Protomer Dynamics in Enzyme Catalysis (Science 2017)

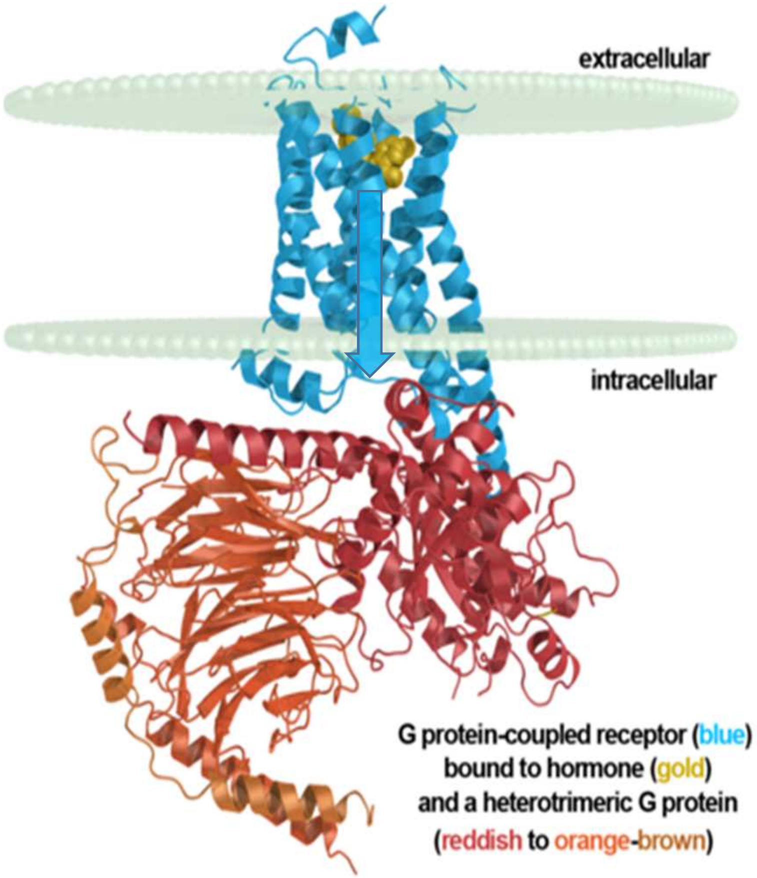
2. Mechanistic insights into allosteric regulation of the A2A adenosine GPCR (G protein-coupled receptor) by physiological cations (Nature Communication 2018)

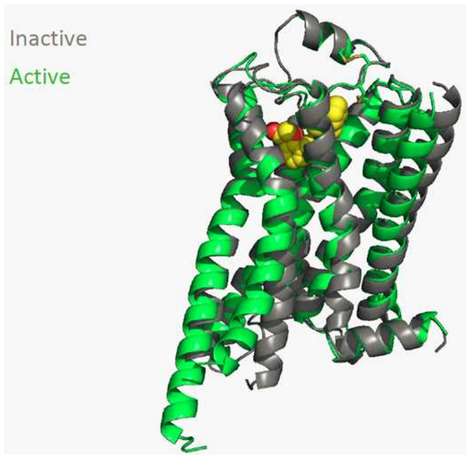
３. Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification (Frontiers in Immunology 2018)

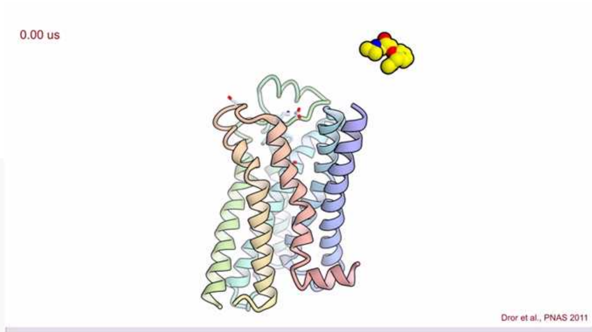
How cells achieve signaling.
A story of GPCRs

Allosteric communication in GPCRs by mechanical propagation in rigidity / transmission in DOF

- Largest group of receptors, respond to drugs, hormones, neurotransmitters, …
- Humans have over 800 GPCRs
- Naturally allosteric but allosteric mechanism not well understood



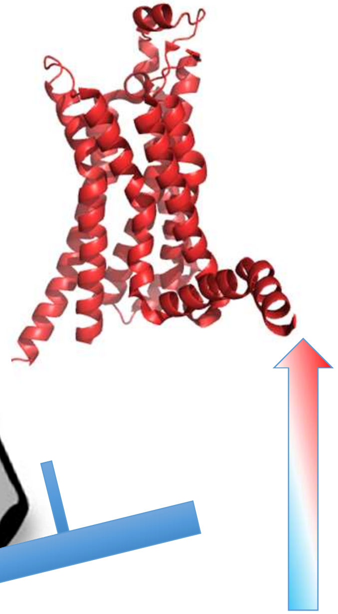


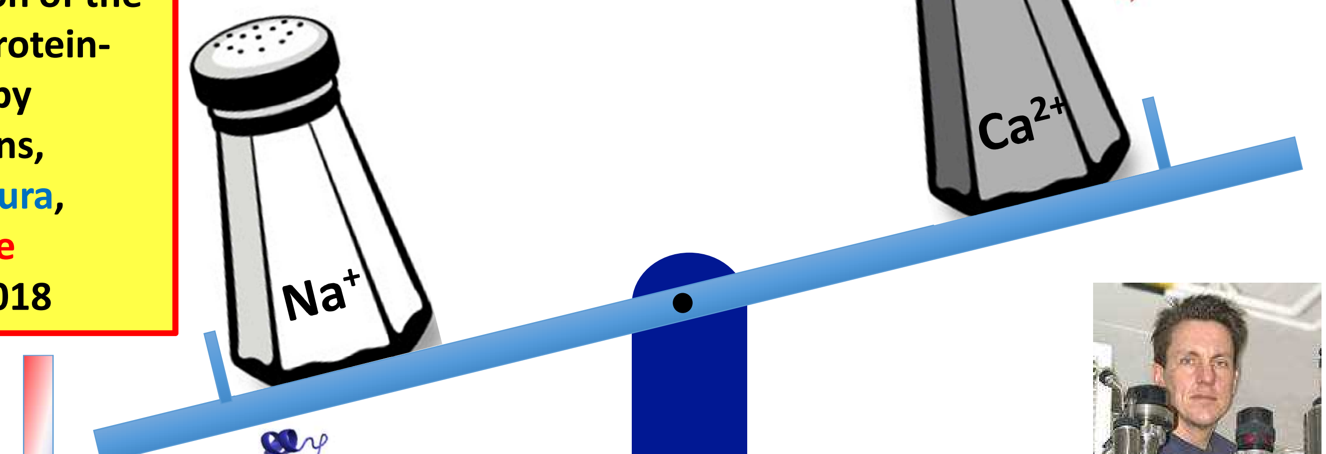


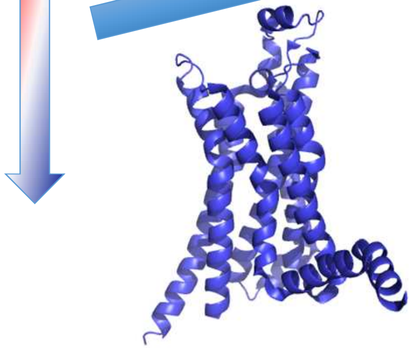
Novel application: Cations (calcium, magnesium, sodium) are key allosteric modulators of GPCR signaling and likely play a role in serving as switches in the activation process

Mechanistic Insights into Allosteric Regulation of the A_{2A} Adenosine G-Protein-Coupled Receptor by Physiological Cations,
Ye, Sljoka, Tsuchimura, Prosser et al Nature Communication, 2018

Activation of A_{2A} adenosine G-protein-coupled receptor by Ca^{2+} and Mg^{2+}



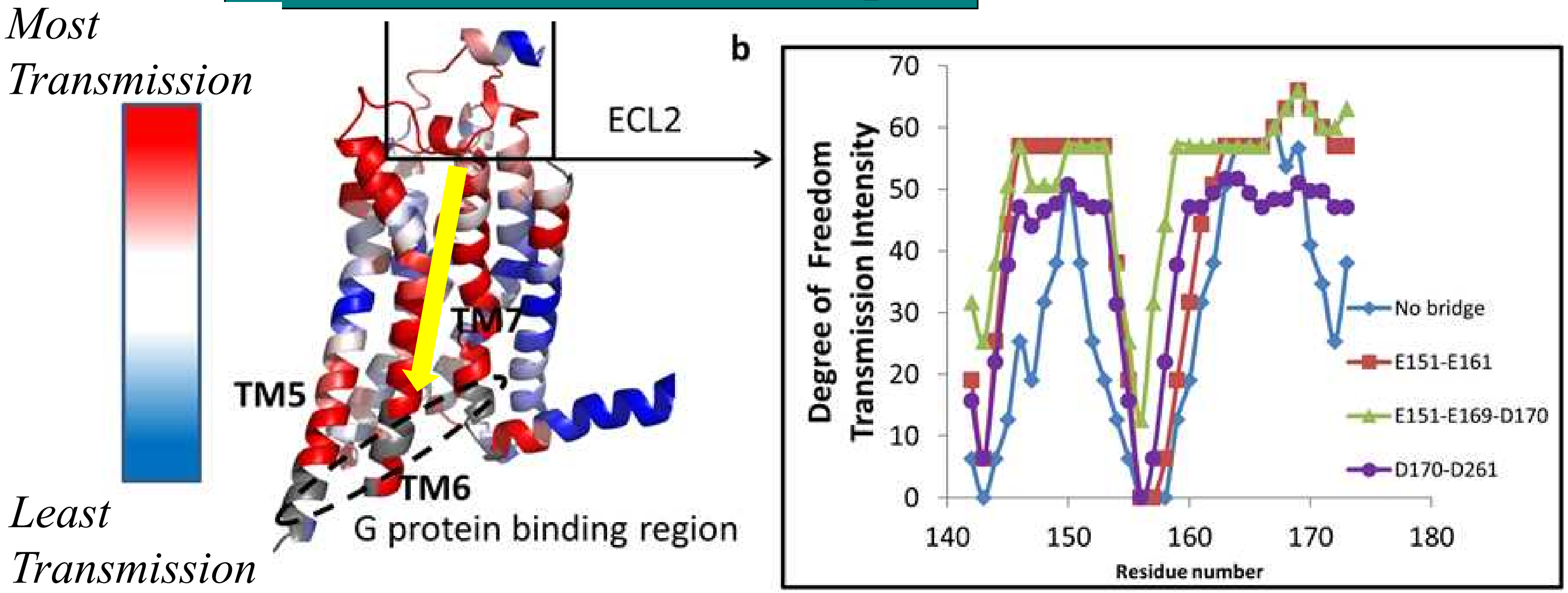




Joint work with Scott Prosser, University of Toronto

Center for Nonlinear Studies, Los Alamos
Department of Pharmacology, UCSD

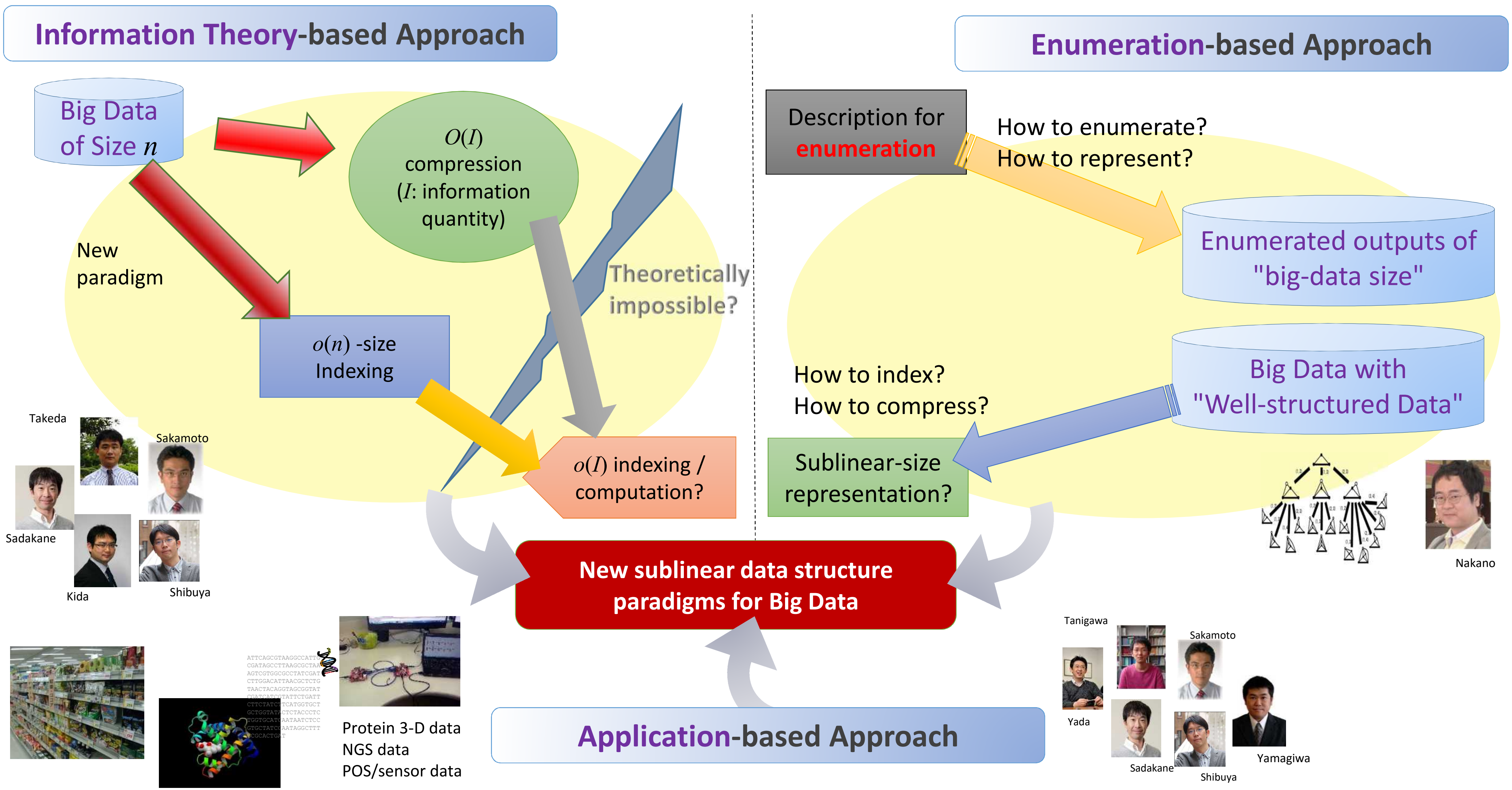
Detection of allosteric hotspots



Positive allosteric modulation of $A_{2A}R$ by Mg^{2+} and Ca^{2+} cations

Mechanistic Insights into Allosteric Regulation of the A_{2A} Adenosine G-Protein-Coupled Receptor by Physiological Cations,
Ye, Sljoka, Tsuchimura, Prosser et al Nature Communication, 2018

Team D: Sublinear Data Structure Research from 3 Approaches



Achievements

- Deep Theories for Compression
  - Small memory compression methods (Sakamoto 2015, 2016, 2017, Kida 2016, 2017, 2018, Sadakane 2016, 2017)
- Big-Data Applications
  - Security data structure designed for massive data
    - Succinct ORAM (Oblivious Random Access Memory) (Onodera, Shibuya, STACS 2018)
  - IoT/Big-data Communication
    - Optimal-space fully-online grammar compression (Takabatake, I, Sakamoto, ESA 2017)
    - Real-time compression/decompression on FPGA for IoT communications (Yamagiwa, Marumo, Sakamoto, VLDB/BPOE 2016) Best Paper Award
  - Network Algorithms
    - Succinct Index for connectivity query on dynamic graphs (Nakamura, ISAAC 2017) Best Student Paper
  - Bioinformatics
    - Protein structure matching/indexing (Shibuya, 2015, 2016) IPSJ Yamashita Award
    - NGS data analysis (Sadakane, Shibuya, 2015)

Application to FPGA-based Low-Cost Communication

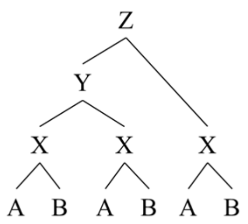
[Yamagiwa, Marumo, Sakamoto, VLDB/BPOE 2016, Best Paper Award]

- **Based on small-memory online self-index**
  - High performance compression
  - **Small FPGA memory space**
  - Online construction
  - Supports search and partial extraction

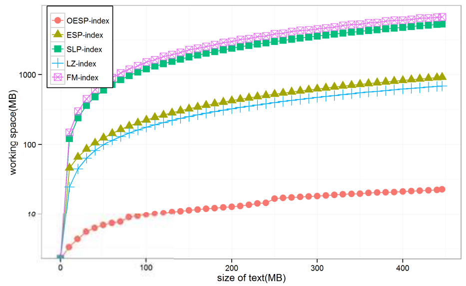
- Hardware implementation
  - Very low cost FPGA
    - Small circuit size / 1CPU time compression

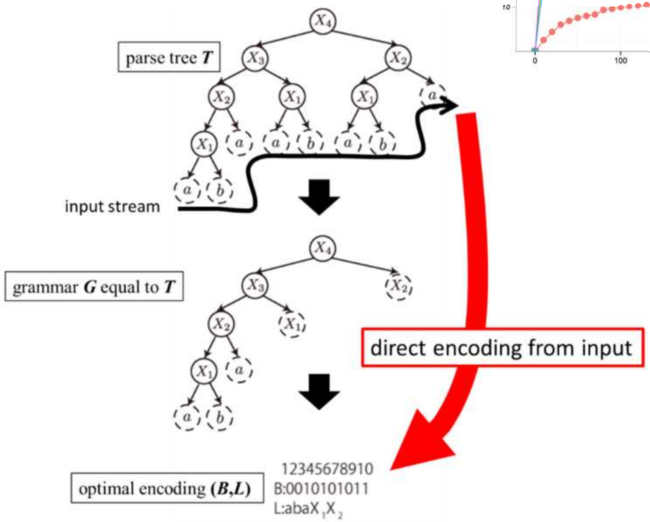
Embedded Technology 2015 Special Award

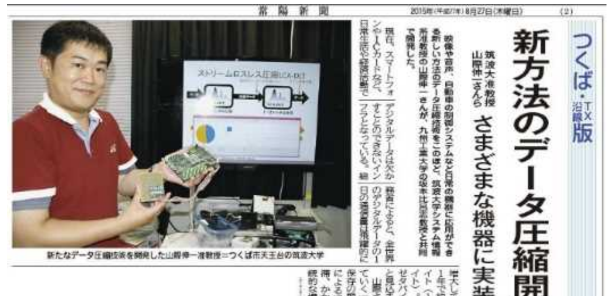
VLDB BPOE workshop best-paper award

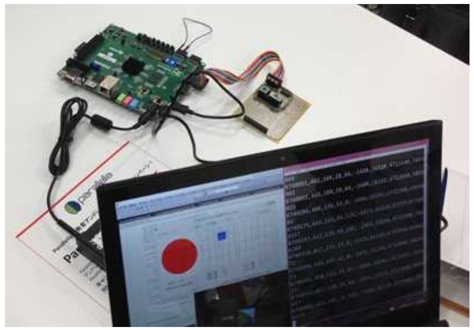


100~1000 times scalability







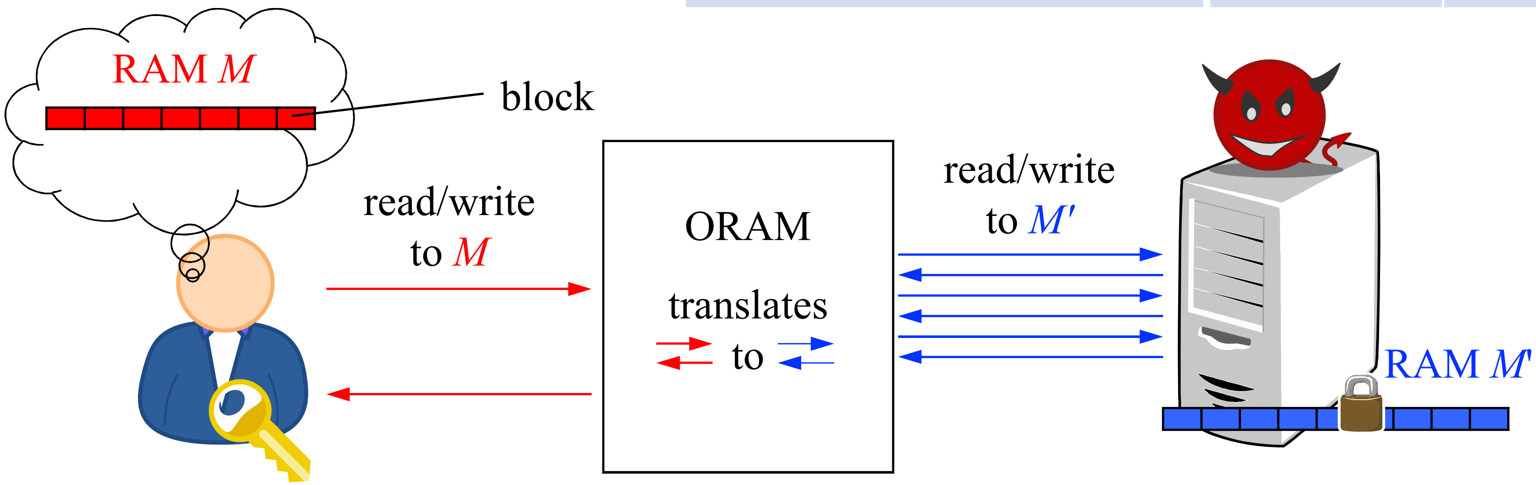


Platform of stream computing in IoT Era

Succinct Oblivious RAM [Onodera, Shibuya STACS 2018]

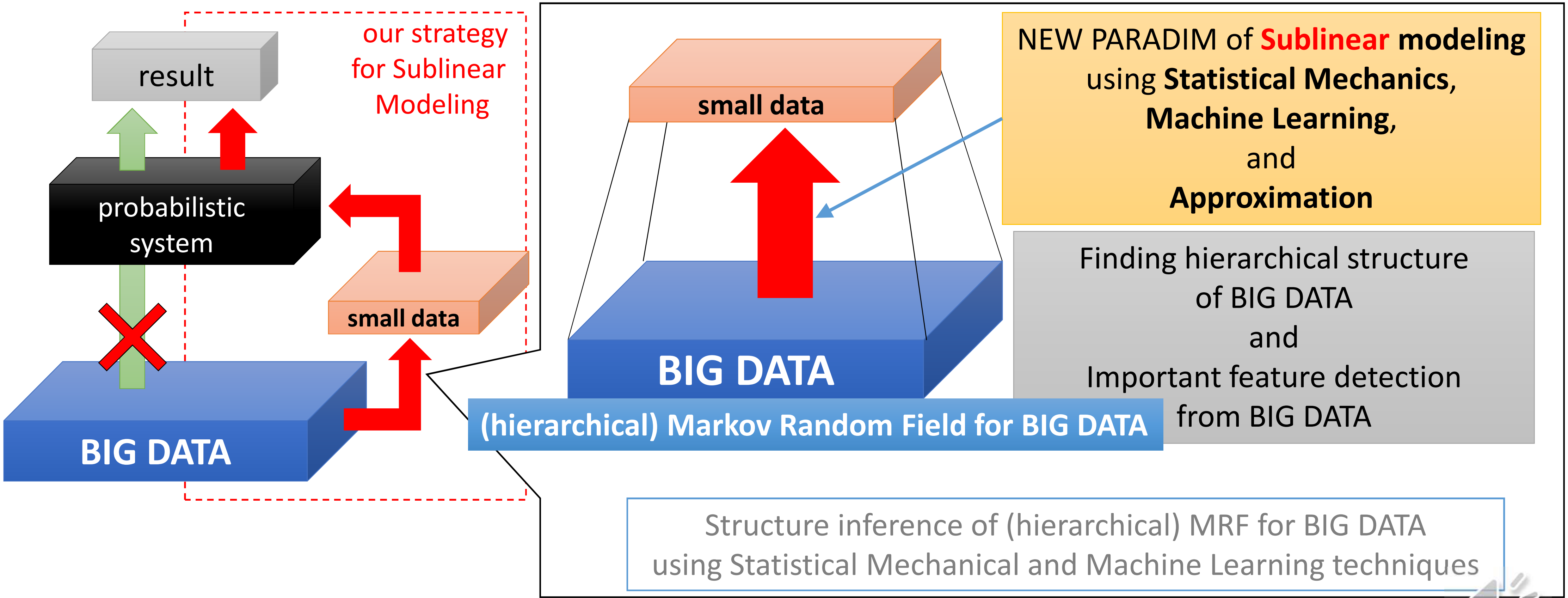
- The first ORAM with $o(\sqrt{n})$ access time and sublinear storage overhead
- Practical performance
  - Needs only 1/100 - 1/4 of the path ORAM/Ring ORAM storage overhead

	Access	Storage overhead
Square Root ORAM Goldreich. STOC87	$O(\sqrt{n})$ amortized	$O(\sqrt{n})$
Path ORAM Stefanov et al. CCS13	$O(\log^2 N)$	$>10N$
Ours	$O(\log^2 N)$	$O\left(\frac{N\log\log N}{\log^{1.4} N}\right)$



Team M: Sublinear Modeling from *Statistical Mechanics*

- **Coarse graining** of Information based on statistical mechanics and machine learning
- Developing efficient approximate algorithms by combining algorithm theory and statistical mechanics



→ Create a new scheme of **sublinear** modeling for Big Data

Sublinear Time Inference using Loopy Belief Propagation for Markov Random Fields

Muneki Yasuda, Shun Kataoka and Kazuyuki Tanaka: Statistical analysis of loopy belief propagation in random fields, *Phys. Rev. E*, Vol.92, pp.042120, 2015.

Numerical Experiments by Loopy Belief Propagation

Source vector

noise

Data vector 1 → Inference of Source Vector 1

Data vector 2 → Inference of Source Vector 2

Data vector 3 → Inference of Source Vector 3

⋮

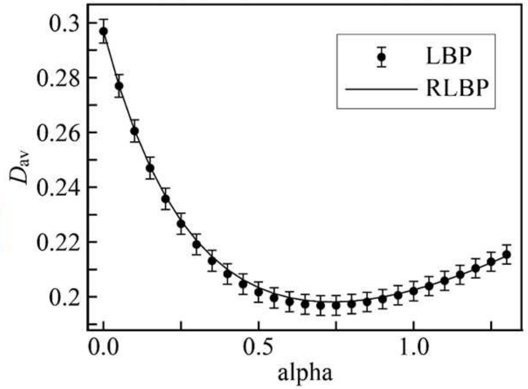
Data vector *M* → Inference of Source Vector *M*

Input

Output

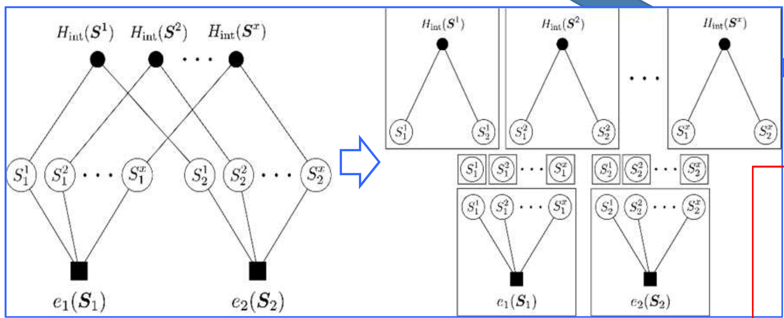
sample average of qualities of outputs

Average Performance of System

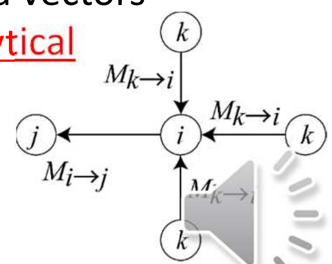


$$O(M) \rightarrow O(1)$$

Analytical Estimation by Replica Loopy Belief Propagation



Sample average for *M* data vectors has been reduced to analytical calculation for statistical average of performance



Approach from Quantum Computation to Big Data Analysis

Query

Response

Digital computer

Quantum computer

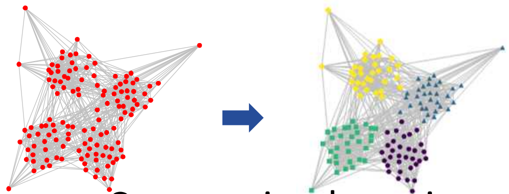
- Quantum annealing machine is a special purpose device for combinatorial optimization problems
- High power saving performance
- Short computational time is expected

Issues

- Is quantum annealing useful for real-world problems?
- What type of problems can be solved efficiently?

How about problems in big data analysis

Approach



Community detection

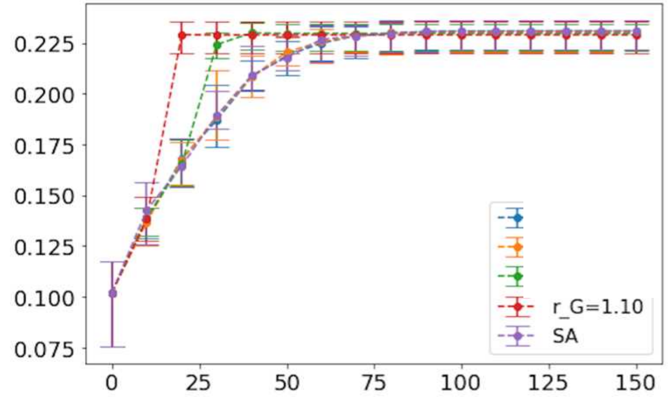
$$X = \begin{pmatrix} 0.7 & 1.4 & 1.7 & ? \\ 1.9 & ? & 0.6 & 1.2 \\ ? & 0.4 & ? & 0.9 \\ ? & 0.0 & 0.4 & ? \end{pmatrix}$$

Matrix interpolation

- Framework of quantum annealing is applied to community detection and matrix interpolation
- Compare the performance of simulated quantum annealing method (SQA) and the conventional simulated annealing method (SA).

Results

- SQA outperforms SA.
- This indicates a positive proof that if we can use D-Wave we may have a better result



Statistical-Mechanical Analysis of Compressed Sensing for Hamiltonian Estimation of Ising Spin Glass

Chako Takahashi[*], Masayuki Ohzeki[*], Shuntaro Okada[*†], Masayoshi Terabe[†], Shinichiro Taguchi[†], Kazuyuki Tanaka[*]

[*] Graduate School of Information Sciences, Tohoku University
[†] DENSO CORPORATION

Background and Contribution

Machines dedicated to solving combinatorial optimization problems by utilizing quantum fluctuation (e.g. D-Wave 2000Q) have recently appeared

- The structure of the machines is sparse Ising Hamiltonian (sparse cost function)
- The unknown parameters of the Ising Hamiltonian must be determined to input problems into the machines – this is a nontrivial task!
- A general-purpose method that expresses real-world problems as sparse Ising Hamiltonian is needed

We propose the Hamiltonian estimation as the L_1-norm minimization

and give the theoretical guarantee of the performance of the L_1-norm minimization

Problem Setting and Formulation

Ising Hamiltonian (cost function)

$$H(\boldsymbol{\sigma}) = -\frac{1}{N}\sum_{i<j} J_{ij}\sigma_i\sigma_j, \ \ \sigma_i \in \{-1,+1\}$$

energy value

$$\boldsymbol{E} = -\frac{1}{N} S \boldsymbol{J}^0$$

observed data

$$\boldsymbol{E} = \left(E^{(1)}, E^{(2)}, \cdots, E^{(M)}\right)^{\mathrm{T}}$$

$$S = \begin{pmatrix} \sigma_1^{(1)}\sigma_2^{(1)} & \cdots & \sigma_1^{(1)}\sigma_N^{(1)} & \sigma_2^{(1)}\sigma_3^{(1)} & \cdots & \sigma_{N-1}^{(1)}\sigma_N^{(1)} \\ \sigma_1^{(2)}\sigma_2^{(2)} & \cdots & \sigma_1^{(2)}\sigma_N^{(2)} & \sigma_2^{(2)}\sigma_3^{(2)} & \cdots & \sigma_{N-1}^{(2)}\sigma_N^{(2)} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \sigma_1^{(M)}\sigma_2^{(M)} & \cdots & \sigma_1^{(M)}\sigma_N^{(M)} & \sigma_2^{(M)}\sigma_3^{(M)} & \cdots & \sigma_{N-1}^{(M)}\sigma_N^{(M)} \end{pmatrix}$$

true coupling constants
(**unknown**)

$$\boldsymbol{J}^0 = \left(J_{12}^0, \cdots, J_{1N}^0, J_{23}^0, \cdots, J_{N-1,N}^0\right)^{\mathrm{T}}$$

$$P(J_{ij}^0) = (1-\rho)\,\delta(J_{ij}^0) + \rho\,\mathcal{N}(0,1)$$

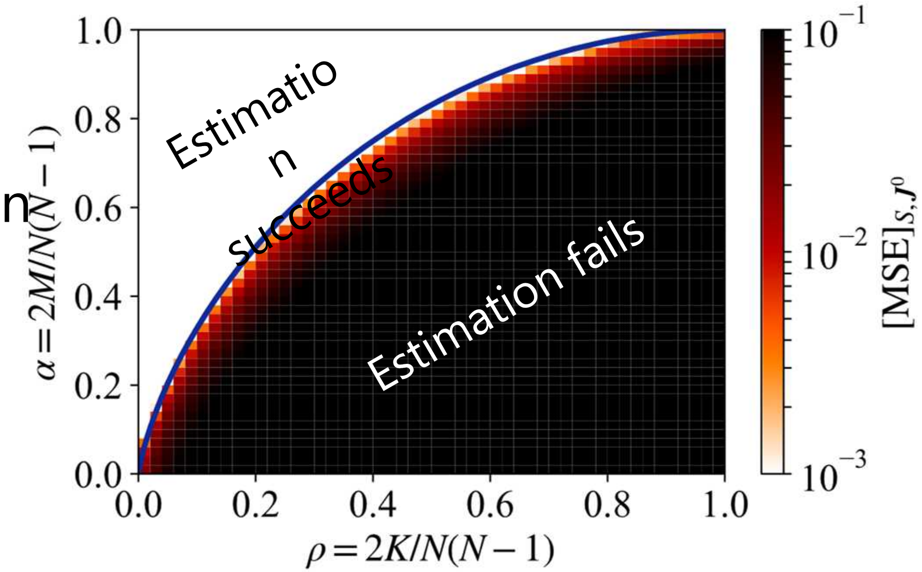
Hamiltonian estimation

$$\min_{\boldsymbol{J}} \|\boldsymbol{J}\|_1 \ \ \text{subject to} \ \ \boldsymbol{E} = \left(-\frac{1}{N} S \boldsymbol{J}^0\right) = -\frac{1}{N} S \boldsymbol{J}$$

Theoretical Analysis of The Estimation

When does the L_1-norm minimization gives “good” solutions?

good: low mean squared error (MSE)



We analyze the behavior of the estimation via replica method [Mezard et al., 1987]

$$[\mathrm{MSE}]_{S,\boldsymbol{J}^0} = \frac{2}{N(N-1)}\left[\langle\|\boldsymbol{J}-\boldsymbol{J}^0\|_2^2\rangle_{\boldsymbol{J}|\boldsymbol{E}}^{\beta\to\infty}\right]_{S,\boldsymbol{J}^0}$$

$$= \rho - 2m + Q$$

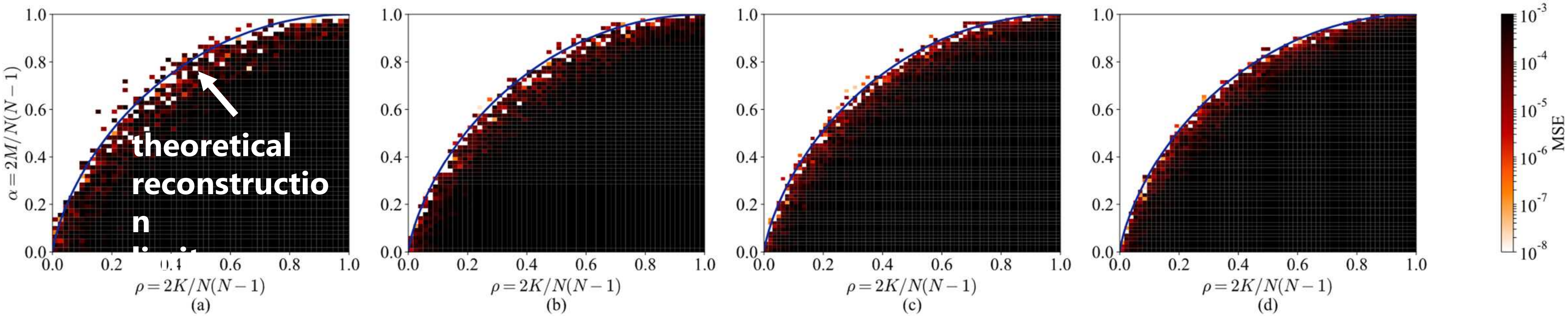
The performance evaluation of the estimation can be done in **sublinear time**!

Numerical Verification

We solve the L_1-norm minimization **quantitatively**

via alternating direction methods of multipliers [Boyd et al., 2011]

$$\mathrm{MSE} = \frac{2}{N(N-1)} \|\boldsymbol{J} - \boldsymbol{J}^0\|_2^2$$



Our theoretical analysis can be considered to be valid!