

# A GPU-ACCELERATED BIOINFORMATICS APPLICATION FOR LARGE-SCALE PROTEIN INTERACTION NETWORKS

Jun Sung Yoon<sup>1</sup>, Won-Hyong Chung<sup>2</sup>

<sup>1</sup>AllegroViva Corporation, California, USA

<sup>2</sup>Korea Research Institute of Bioscience & Biotechnology, Daejeon, Korea



## Introduction

Proteins, nucleic acids, and small molecules form a dense network of molecular interactions in a cell. The architecture of molecular networks can reveal important principles of cellular organization and function, similarly to the way that protein structure tells us about the function and organization of a protein. Protein complexes are groups of proteins that interact with each other at the same time and place, forming a single multimolecular machine. Functional modules, in contrast, consist of proteins that participate in a particular cellular process while binding each other at a different time and place<sup>1</sup>.

A protein-protein interaction network is represented as proteins are nodes and interactions between proteins are edges. Protein complexes and functional modules can be identified as highly interconnected subgraphs and computational methods are now inevitable to detect them from protein interaction data. In addition, High-throughput screening techniques such as yeast two-hybrid screening enable identification of detailed protein-protein interactions map in multiple species. As the interaction dataset increases, the scale of interconnected protein networks increases exponentially so that the increasing complexity of network gives computational challenges to analyze the networks.

Graphics hardware is recently widely used in high-performance computing due to its cost effectiveness. Bioinformatics applications also exploit GPU as a massive parallel multi-core processor to address computational challenges in the many areas such as sequence analysis and protein structure prediction. However, few attempts have been made to analyze biological networks.

We present a fast parallel implementation using commodity graphics hardware based a well-known sequential complex finding algorithm of MCODE<sup>2</sup> to address the computational challenge. Our parallel algorithm is implemented on the NVIDIA parallel computing architecture of CUDA. It is evaluated for a various kinds of large-scale PPI networks. Our GPU accelerated implementation using the latest NVIDIA graphics hardware achieves a speedup of two orders of magnitudes compared to the original MCODE in the latest CPU for lager-scale protein-protein interaction networks.

## Parallel Algorithm

### Parallel MCODE Algorithm

#### 1. Vertex Weighting

Input graph:  $G = (V, E)$

for all  $v$  in  $G$  do *in parallel*

$N_v \leftarrow$  find the subgraph which includes the immediate neighbors of  $v$

$K_v \leftarrow$  Get highest k-core graph from  $N_v$

$k_v \leftarrow$  Get highest k-core number from  $N_v$

$d_v \leftarrow$  Get density of  $K_i$

$W_v \leftarrow k_v \times d_v$

end for

#### 3. Post-processing

$C$ : complex subgraph

$h$ : haircut flag,  $f$ : fluff flag

for all  $c$  in  $C$  do *in parallel*

if  $c$  not 2-core then filter

if  $h$  is TRUE then 2-core complex

if  $f$  is TRUE then fluff complex

end for

#### 2. Molecular Complex Prediction

$d$ : vertex weight percentage

$W_v$ : vertex weight of  $v$

$S_v$ : vertex weight of seed of  $v$

$N_v$ : seed vertex of  $v$

$S_v \leftarrow v$ , for all  $v$

while there is any changes of  $S_v$

for all  $v$  neighbors of  $n$  do *in parallel*

if  $N_v \leftrightarrow N_n$  then

if  $W_v < S_n$  AND  $W_v > (1-d) S_n$  then

$S_v \leftarrow S_n$

$N_v \leftarrow N_n$

else if  $W_v = S_n$  AND  $C_v > C_n$  then

$N_v \leftarrow N_n$

end if

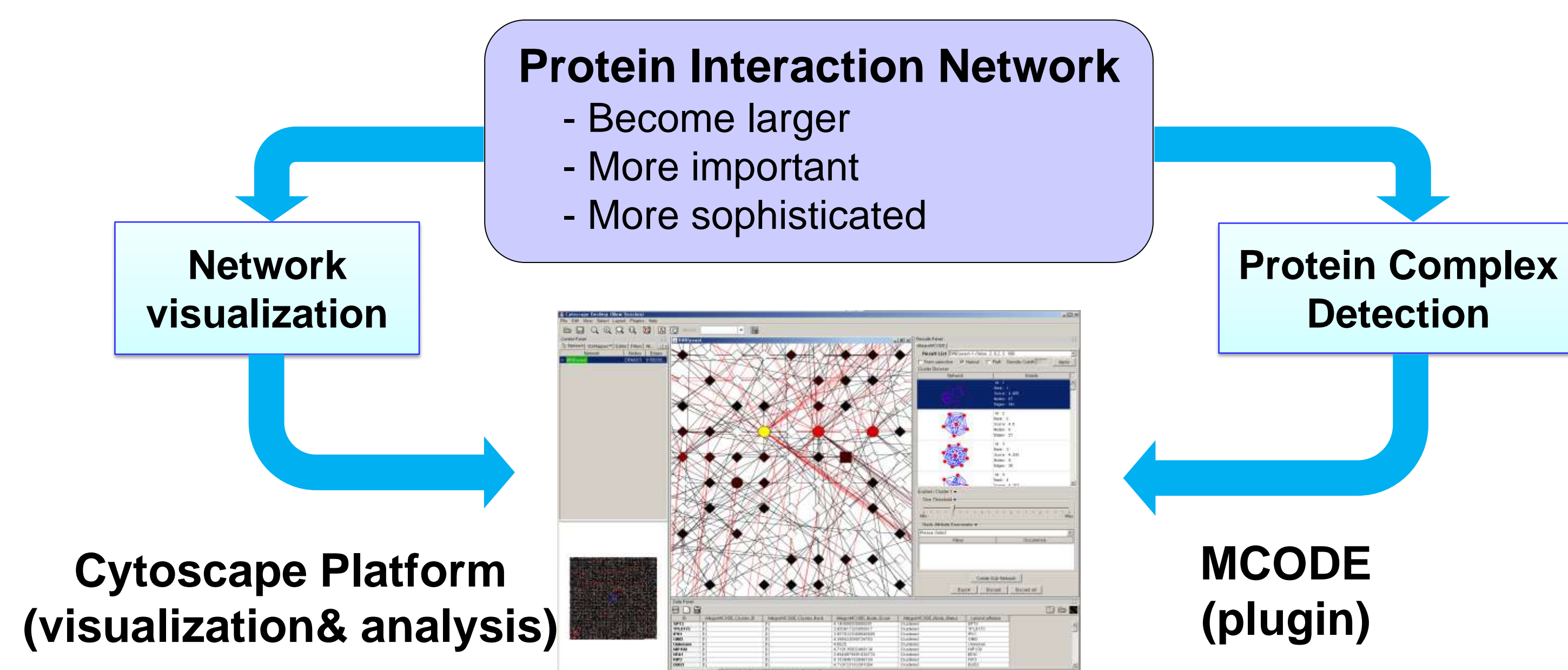
end if

*synthronize all threads*

end while

## Protein Complex Prediction

### Using Cytoscape and MCODE plugin



A well-known molecular complex detection tool of **MCODE** plugin is integrated in the open-source network visualization and analysis platform of **Cytoscape platform**. This architecture has two limitations to handle contemporary large interaction network.

- Serial computation: Can not fully exploit multi-core processors  
→ Long-time waiting to analyze large network interaction
- Standalone system: Its computing power is limited to user's PC hardware spec.  
→ Users need to upgrade their hardware themselves

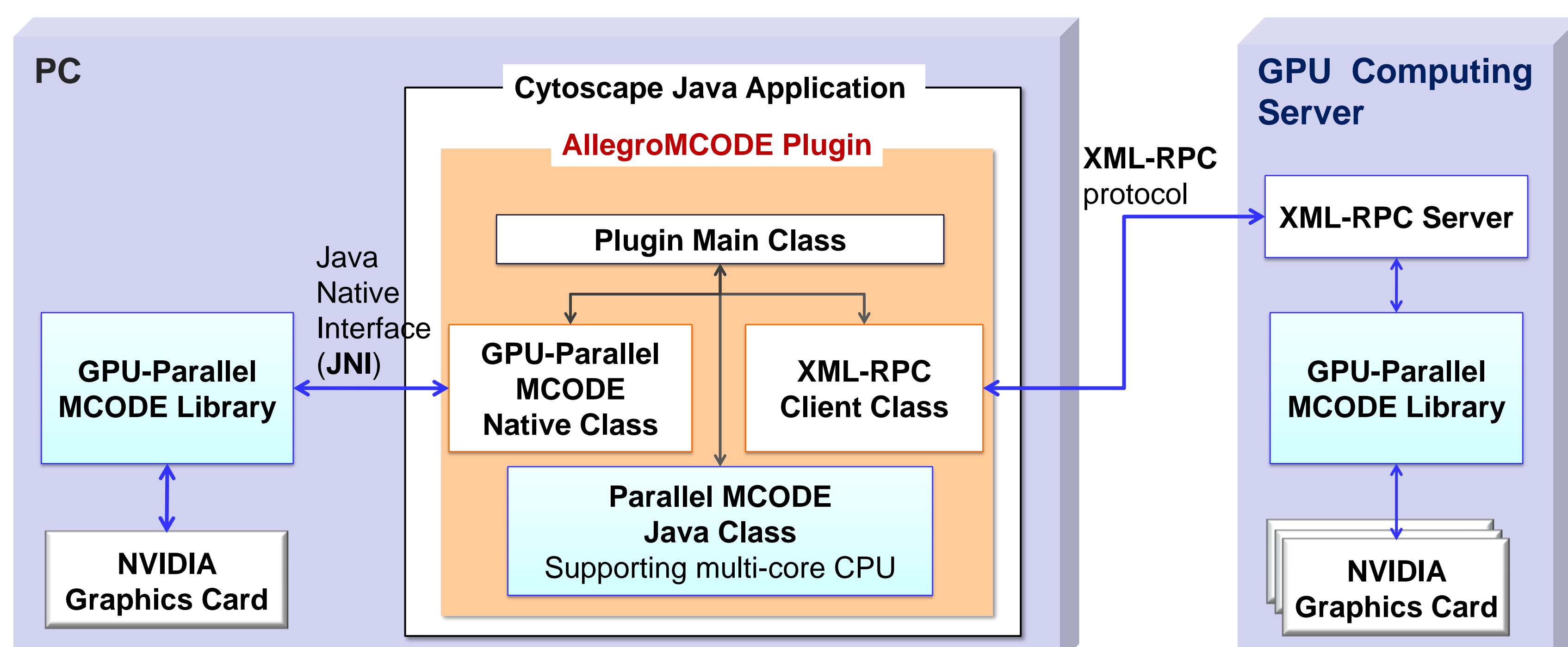
## GPU-accelerated Computing Architecture

### GPU Computing Server

- ✓ Enable you to exploit the GPU acceleration without any special graphics hardware.
- ✓ Provides the remote procedure call via the standard XML-RPC protocol.
- ✓ Various clients implemented in Perl, Python, C, C++, Java and PHP can easily make a request to the server by sending a XML document.

### AllegroMCODE plugin

- ✓ A Cytoscape plugin to help you use the remote GPU Computing Server.
- ✓ Supports the GPU algorithm acceleration to use your graphics hardware by loading the same GPU-Parallel MCODE Library.
- ✓ Includes multi-threaded parallel MCODE Implementation to fully exploit all the cores in a CPU.



## Performance

The processing time is measured by running MCODE Cytoscape plugin and our AllegroMCODE Cytoscape plugin with the same options of the algorithm.

### System Specification

CPU	Main Memory	O/S	GPU
Intel Core i7 920 @ 2.67GHz	6 GB DDR3 RAM	Ubuntu Linux 10.04 LTS	NVIDIA GTX580

### Algorithm Options

Include Loops	Degree Cutoff	Node Score Cutoff	K-core	Max. Depth	Haircut	Fluff
Disabled	2	0.2	2	100	Enabled	Disabled

### Test Networks

Network	Description
A	Protein Interaction Network from BioGRID database
B	Protein Interaction Network from IntAct database
C	Protein Interaction Network from I2D database
D	Protein Interaction Network from DIP database
E	Yeast Protein Interaction Network from DroID database
F	Human Protein Interaction Network from DroID database

### Test Network Statistics

Network	Edges	Nodes
A	317,706	31,215
B	196,631	52,557
C	149,912	6,077
D	86,729	29,405
E	75,937	2,760
F	55,88	4,612

### Processing Time (sec)

Network	MCODE	AllegroMCODE
A	145.91	0.52
B	149.13	0.32
C	66.45	0.19
D	39.45	0.18
E	92.66	0.17
F	47.82	0.11

### Speedup

A	278 x
B	460 x
C	357 x
D	222 x
E	536 x
F	451 x

## Reference

- V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12123–12126, 2003.
- G. D. Bader and C. W. V. Hogue "An automated method for finding molecular complexes in large protein interaction networks", *BMC Bioinformatics*, 4(2), 2003

## Further Information

AllegroMCODE plugin and our GPU computing server are freely available. You can get more information about the installation and usage from [allegroviva.com/allegromcode](http://allegroviva.com/allegromcode). Cytoscape is an open source platform for complex-network analysis and visualization and freely available from [www.cytoscape.org](http://www.cytoscape.org).

Jun Sung Yoon : [jyoon@allegroviva.com](mailto:jyoon@allegroviva.com)

Won-Hyong Chung : [whchung@kribb.re.kr](mailto:whchung@kribb.re.kr)