hp120131

タンパク質間相互作用ネットワークの推定とその応用に関する研究

Large-scale protein-protein interaction network prediction by an exhaustive rigid docking system MEGADOCK

松崎 由理

Yuri Matsuzaki

東京工業大学　情報生命博士教育院

matsuzaki@acls.titech.ac.jp

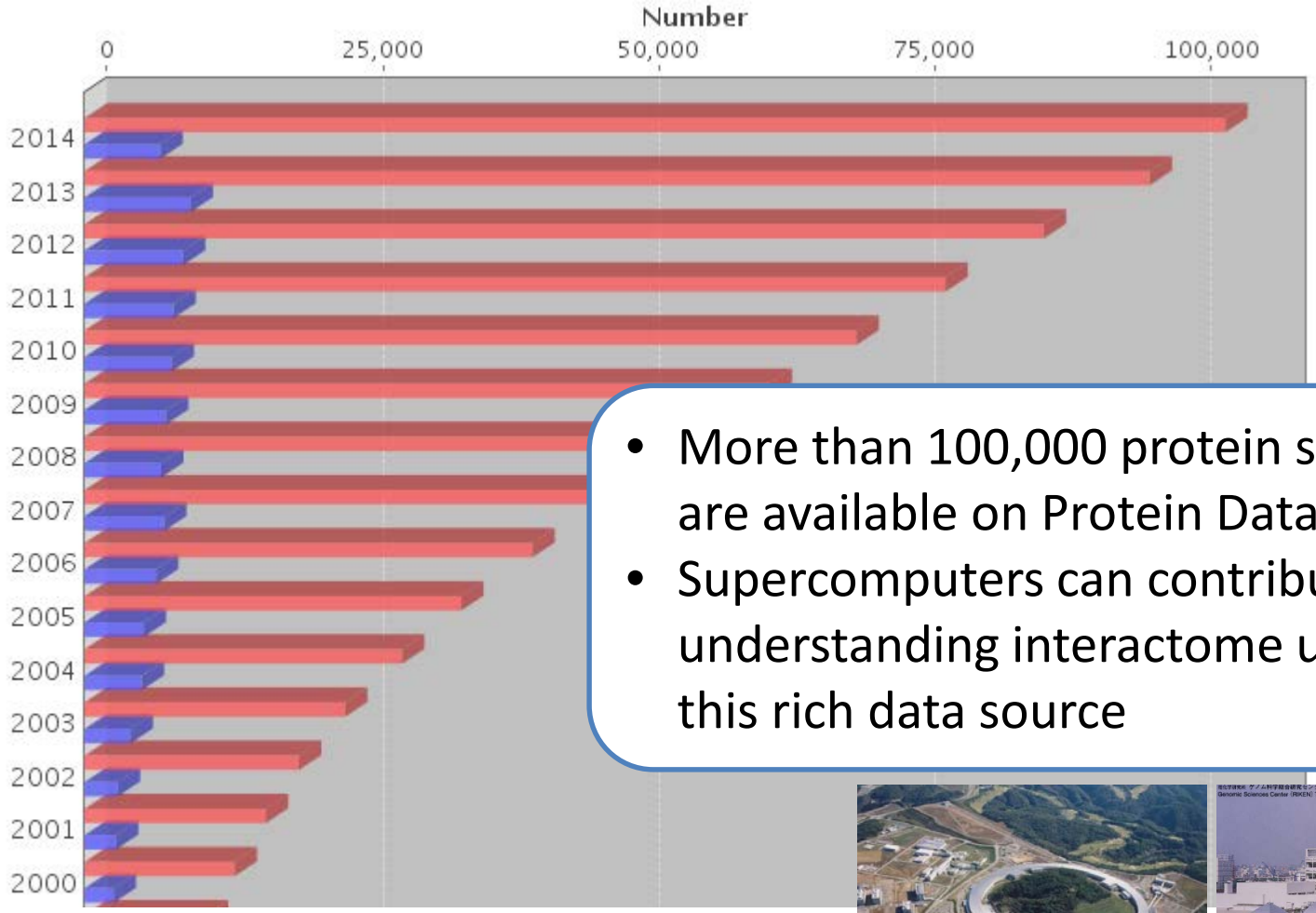
■課題代表者　秋山 泰（東工大）

Yutaka Akiyama (Tokyo Institute of Technology)

■参加者　松崎由理，石田貴士，大上雅史（東工大），内古閑伸之（中央大）

Yuri Matsuzaki, Takashi Ishida, Masahito Ohue (Tokyo Institute of Technology), Nobuyuki Uchikoga (Chuo University)

The number of available protein tertiary structures is growing rapidly



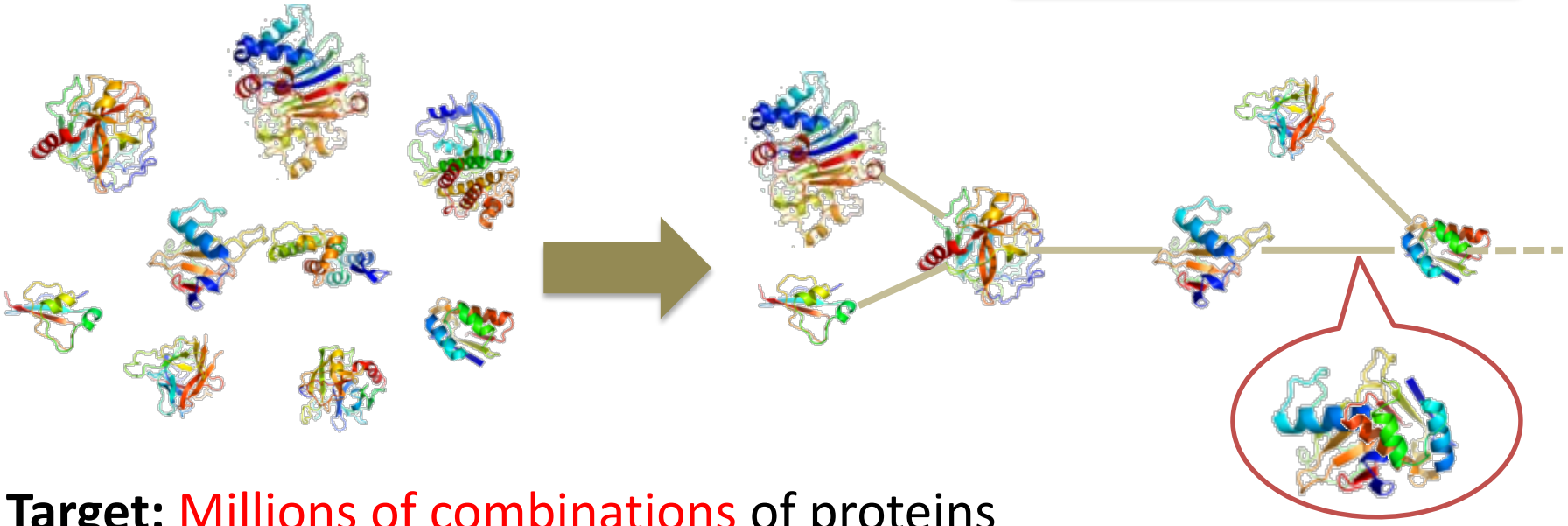
- More than 100,000 protein structures are available on Protein Data Bank
- Supercomputers can contribute understanding interactome utilizing this rich data source

RCSB PDB
(as of 2014-09-22)

What we do: Protein-protein interaction (PPI) network prediction using proteins tertiary structure data

Input: Protein structures

Output: PPI networks



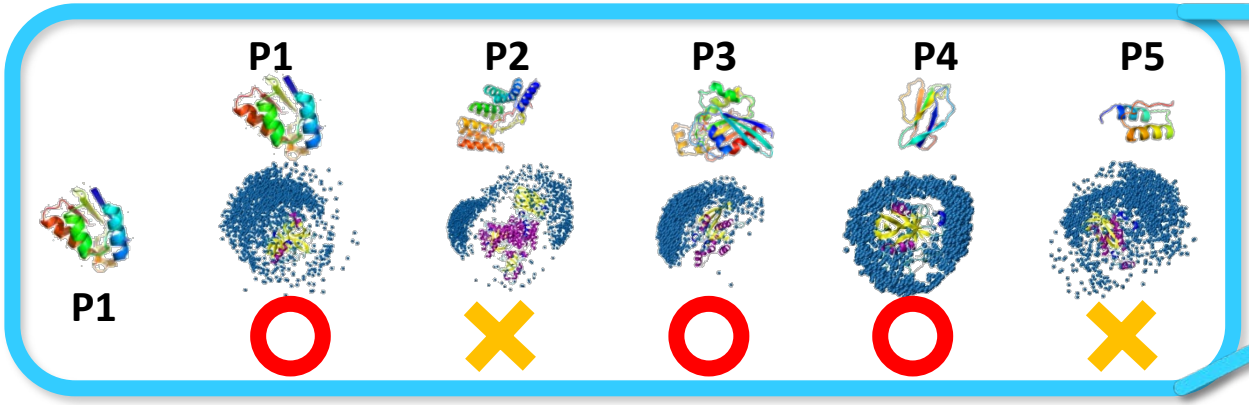
Target: Millions of combinations of proteins

Strategy: Exhaustive rigid docking among target proteins

- Feasible assuming massively parallel computing environments
- Prediction of complex structures can also be provided

MEGADOCK system: PPI network prediction by an ultra-fast rigid-docking tool

Exhaustive docking & post-docking analysis



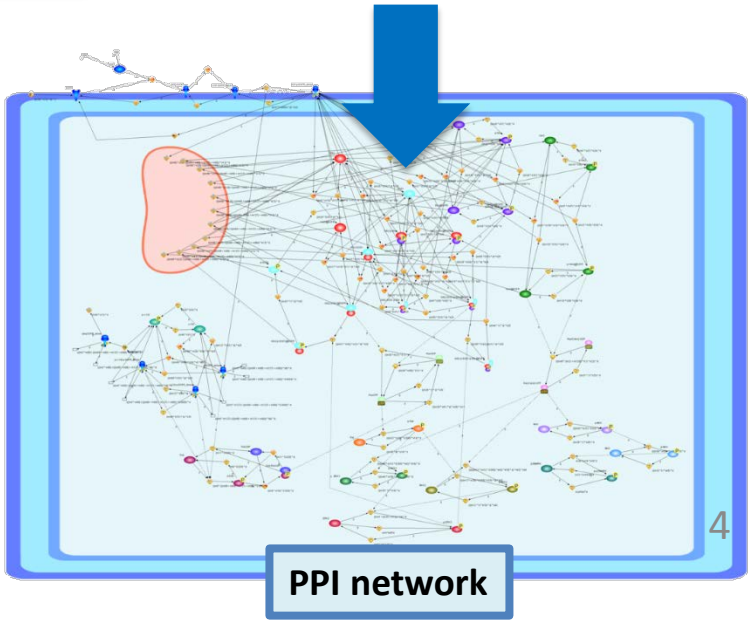
Prediction of binders

	P1	P2	P3	P4	P5
P1	○	×	○	○	×
P2	×	×	×	×	×
P3	○	×	×	○	×
P4	○	×	○	×	○
P5	×	×	×	○	×

MEGADOCK

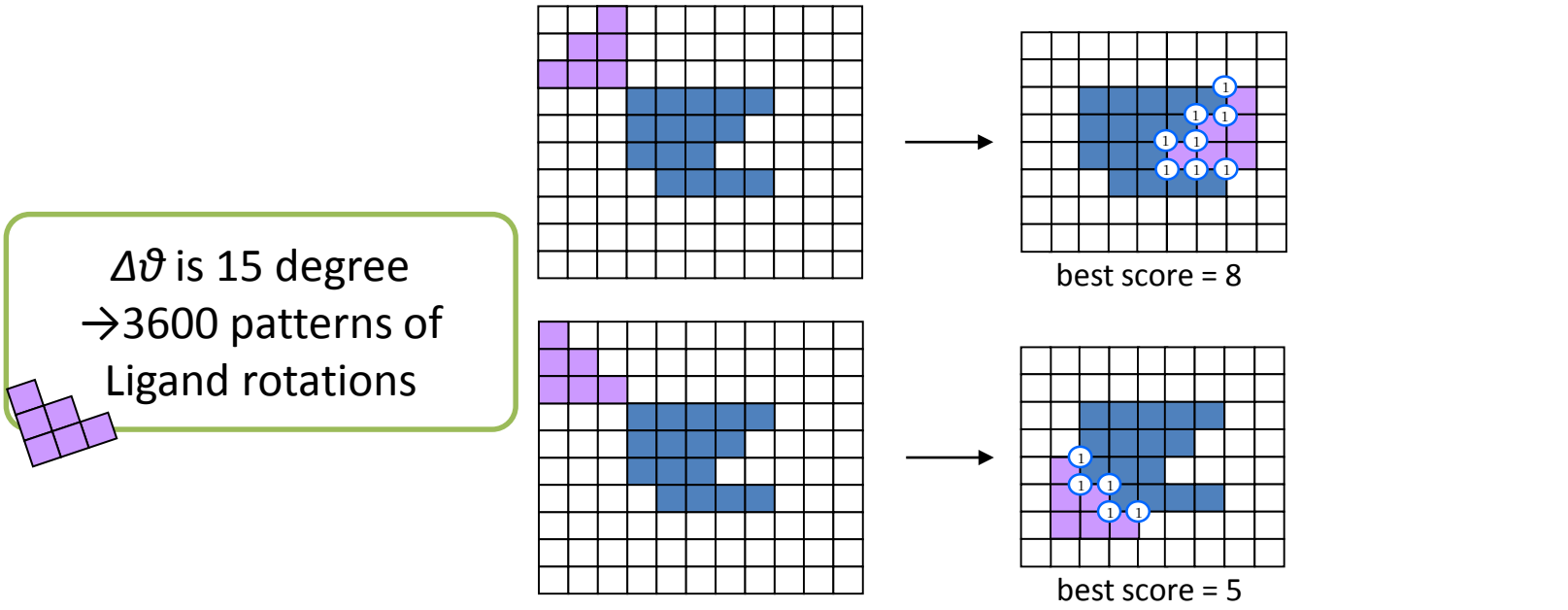
- Rigid-docking software designed for exhaustive docking studies
- Suitable for running on supercomputers
  - K, SCLS, TSUBAME2.5 (CPU/GPU version)
- Open-source
  http://www.bi.cs.titech.ac.jp/megadock/

Ohue, *et al.*, *Bioinformatics*, accepted.



Rigid Docking

- Evaluates docking scores mainly based on the complementarity of protein tertiary structure



Katchalski-Katzir E, *et al*. *PNAS*, 1992.

A compact score function of MEGADOCK

Compress three terms into **one complex number**:

- **Shape complementarity**
- **Hydrophobic interaction**
- **Electrostatic interaction**

$$\boldsymbol{S}(\alpha,\beta,\gamma) = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\boldsymbol{R}(l,m,n)\boldsymbol{L}(l+\alpha,m+\beta,n+\gamma)\right]$$

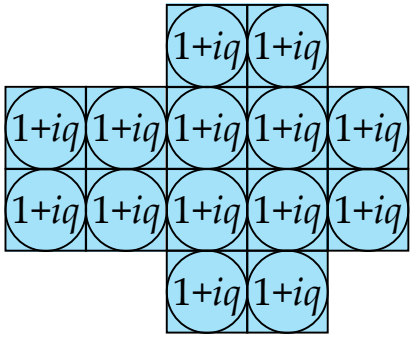
$$\boldsymbol{R}(l,m,n) = G_R(l,m,n) + w_h H(l,m,n) + i\phi(l,m,n)$$

$$\boldsymbol{L}(l,m,n) = G_L(l,m,n) + i w_e q(l,m,n)$$

$$\boldsymbol{S}(\alpha,\beta,\gamma) = \Re\left[\mathrm{IFT}\left[\mathrm{DFT}\left[\boldsymbol{R}(l,m,n)\right]^{*}\mathrm{DFT}\left[\boldsymbol{L}(l,m,n)\right]\right]\right]$$

1+H +$i\phi$	2+H +$i\phi$	3+H +$i\phi$	3+H +$i\phi$	3+H +$i\phi$	2+H +$i\phi$	1+H +$i\phi$
2+H +$i\phi$	-45	-45	-45	-45	-45	2+H +$i\phi$
3+H +$i\phi$	-45	-45	-45	-45	-45	2+H +$i\phi$
3+H +$i\phi$	-45	-45	-45	5+H +$i\phi$	2+H +$i\phi$	1+H +$i\phi$
3+H +$i\phi$	-45	-45	-45	5+H +$i\phi$	2+H +$i\phi$	1+H +$i\phi$
3+H +$i\phi$	-45	-45	-45	-45	-45	2+H +$i\phi$
2+H +$i\phi$	-45	-45	-45	-45	-45	2+H +$i\phi$
1+H +$i\phi$	2+H +$i\phi$	3+H +$i\phi$	3+H +$i\phi$	3+H +$i\phi$	2+H +$i\phi$	1+H +$i\phi$

Receptor protein

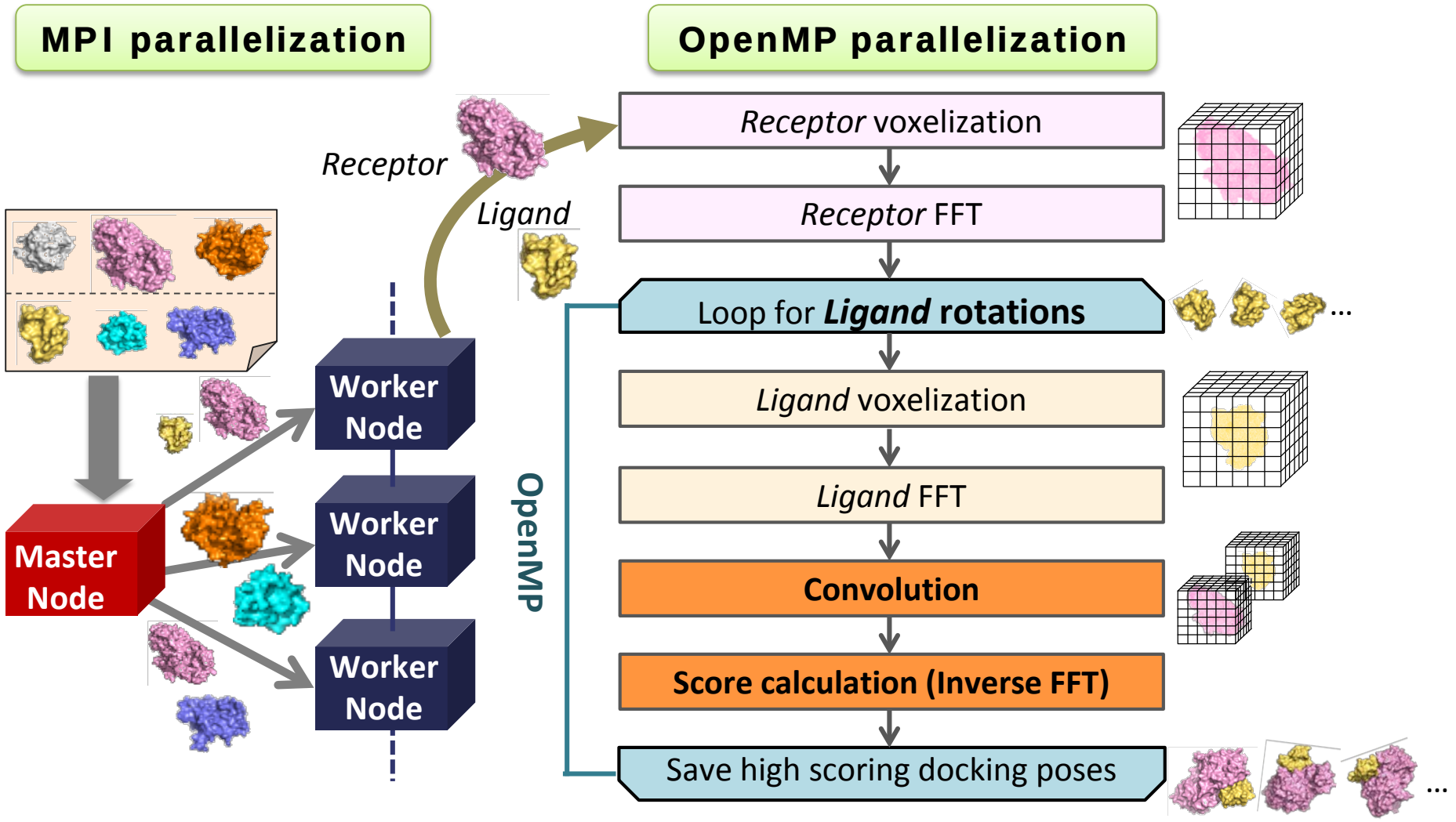


Ligand protein

Convolution can be calculated fast by FFT (Katchalski-Katzir model)

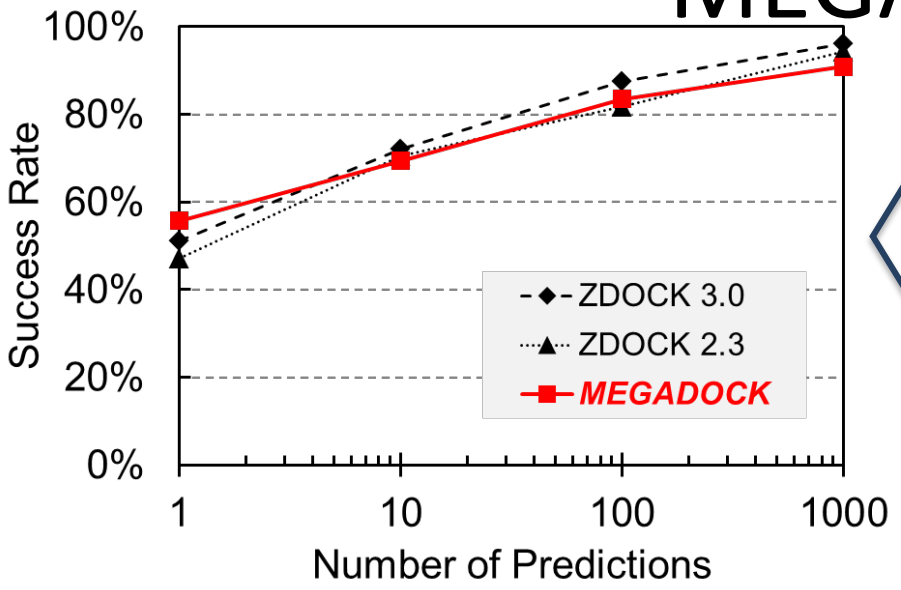
Ohue *et al.*, *Lecture Note in Bioinformatics*, 2012.

Implementation by hybrid parallelization



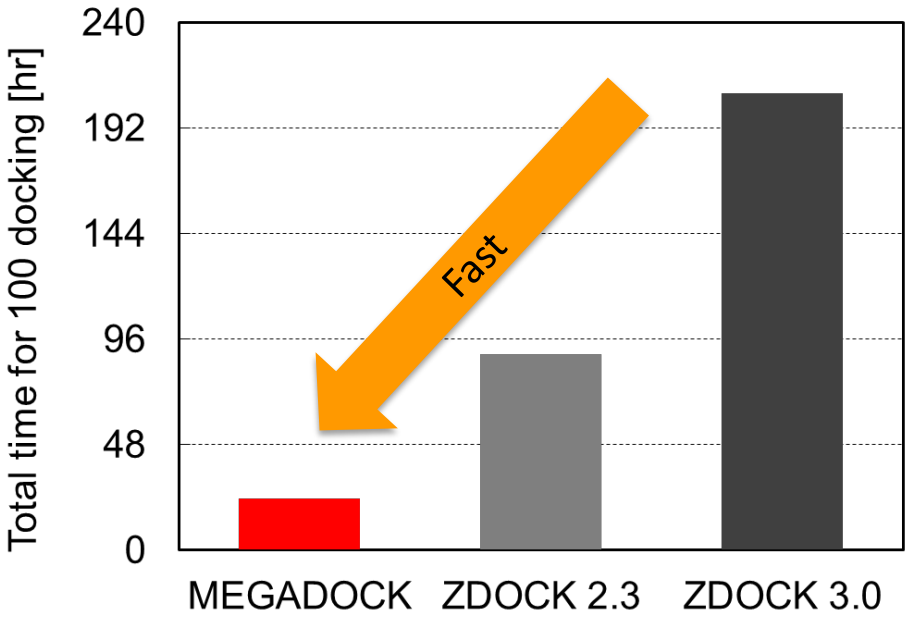
Matsuzaki, *et al.*, *Source Code Biol Med*, 2013.

Performance of protein docking by MEGADOCK



Docking accuracy was **comparable** to conventional tools
(Benchmark 4.0, 176 bound complexes)

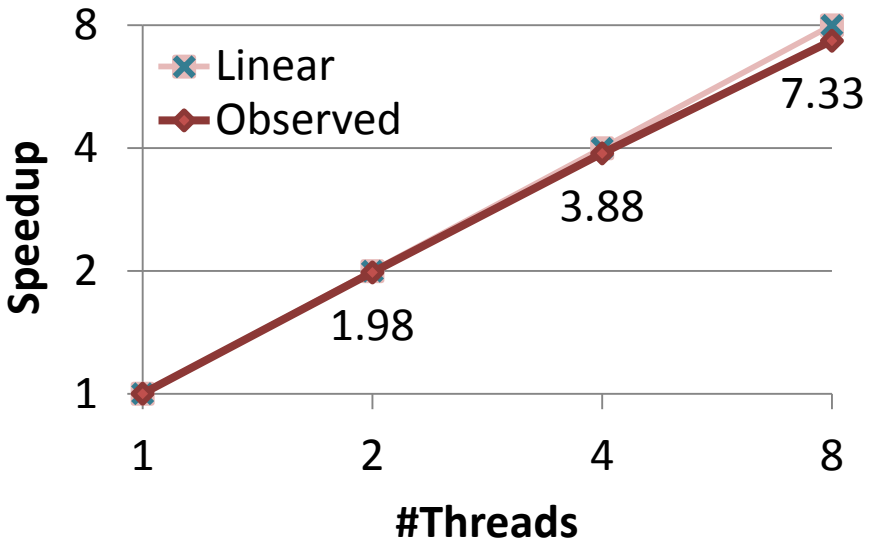
8.8 times speedup
compared to ZDOCK 3.0
(single node, single thread)



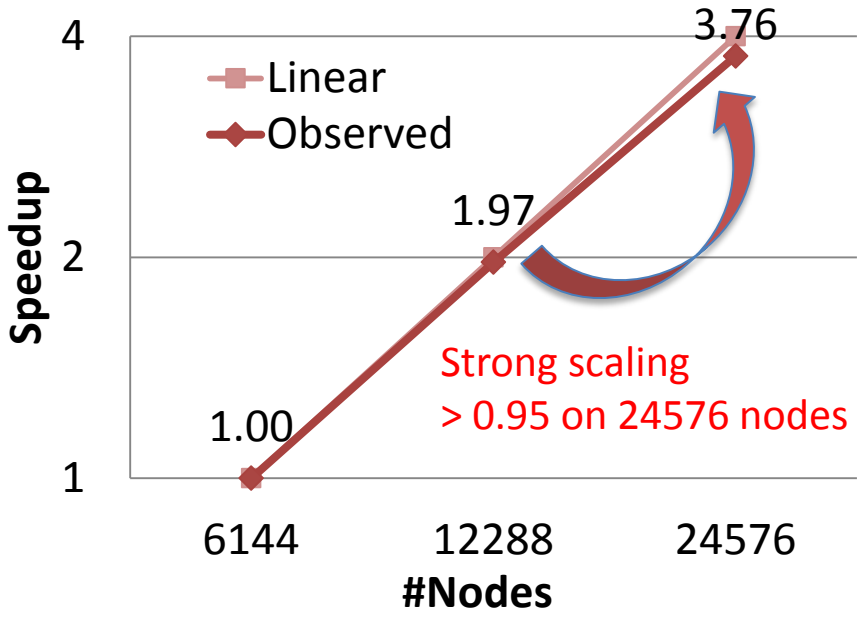
Ohue, *et al.*, *Lecture Note in Bioinformatics*, 2012.

Scalability on K computer

OpenMP parallelization



MPI parallelization

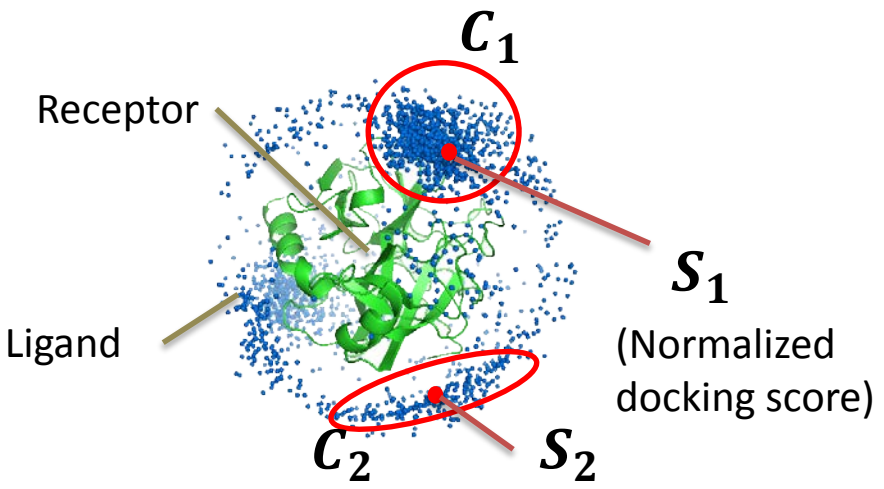


- Less but sufficient scalability (strong scaling 0.91) was observed with 82,944 nodes.

Matsuzaki, *et al.*, *Source Code Biol Med*, 2013.

Prediction method 1: clustering based

- Get 2,000 high scoring models by docking of each protein pair
- Conduct clustering based on structure similarity
- Define the highest docking score (normalized) of the data included in the cluster C_i : s_i
- Define cluster population (normalized) : m_i
- Select populated clusters C' with threshold m^* of population of the cluster $C' = \{C_i \mid m_i > m^*\}$
- Decide PPI score E
- Evaluate each pair of protein combination as interacting if E is higher then the threshold E^*

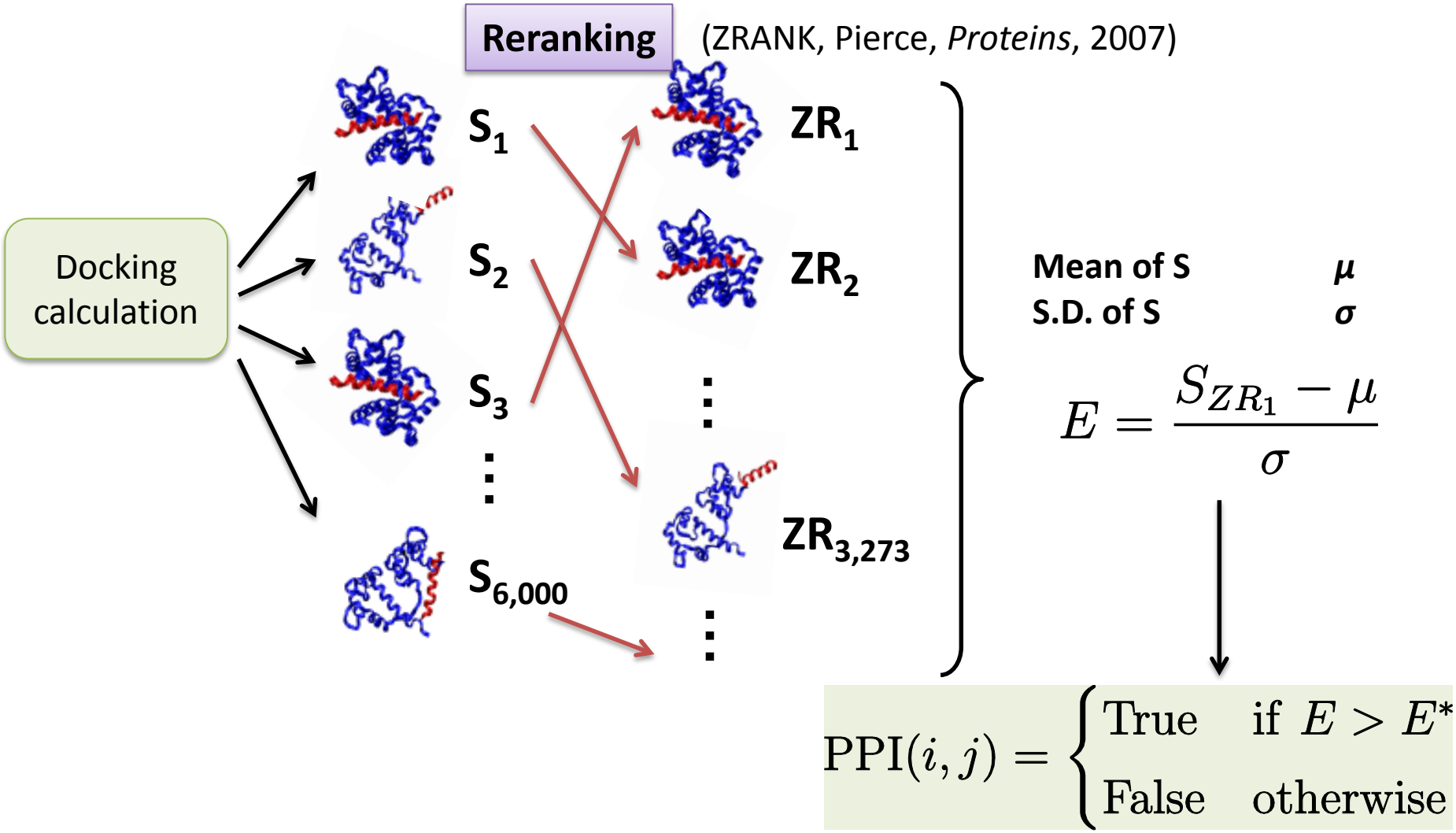


$$E = \begin{cases} max\ s_i,\ i \in C' & \text{if } C' \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

$$interaction = \begin{cases} true & \text{if } E > E^* \\ false & \text{otherwise} \end{cases}$$

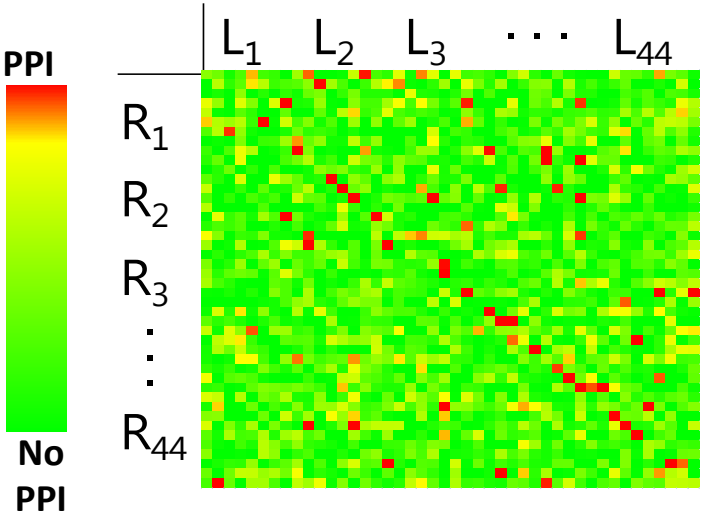
Matsuzaki, *et al., J Bioinform Comput Biol, 2009.*

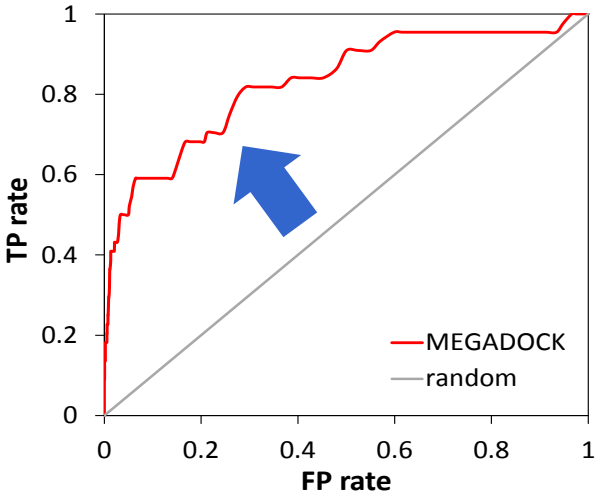
Prediction method 2: reranking based

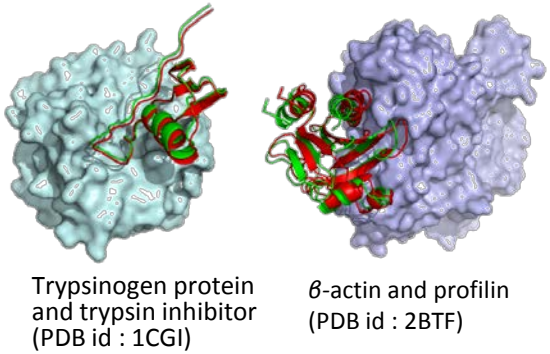


Ohue, *et al.*, *Protein Pept Lett*, in press.

PPI prediction using general benchmark data







Binding partner prediction from 44x44=1936 dockings and post-docking (Diagonal: interacting pairs)

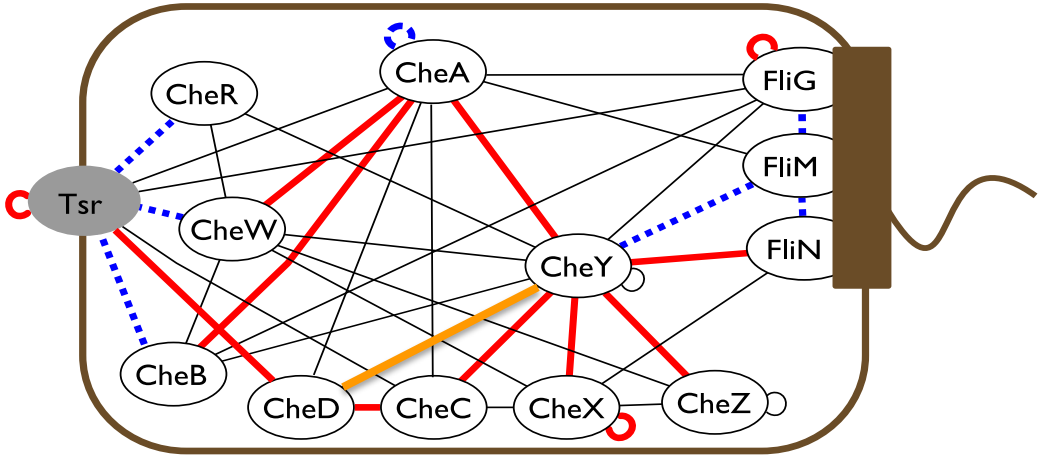
F-measure : 0.42

$$F\text{-measure} = \frac{2 \cdot \mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})+(\mathrm{TP}+\mathrm{FN})}$$

PPI prediction by MEGADOCK achieved better than random performance on general benchmark dataset (monomer pair from protein-protein docking benchmark 2.0, Mintseris *et al*, *Proteins*, 2005.)

Matsuzaki, *et al.*, *J Bioinform Comput Biol*, 2009.
Ohue, *et al.*, *Protein Pept Lett*, *2014*.

Application to bacterial chemotaxis pathway

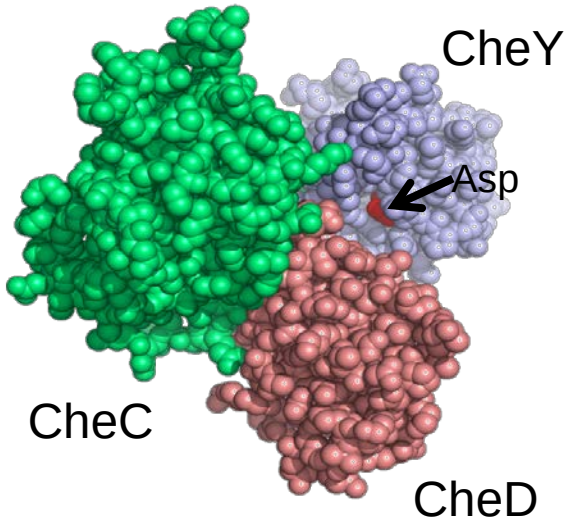


: True Positive
: False Negative
: False Positive

13 proteins
(101 structures)

F-measure : 0.44

Acceptable performance was shown on a real biology pathway reconstruction problem.

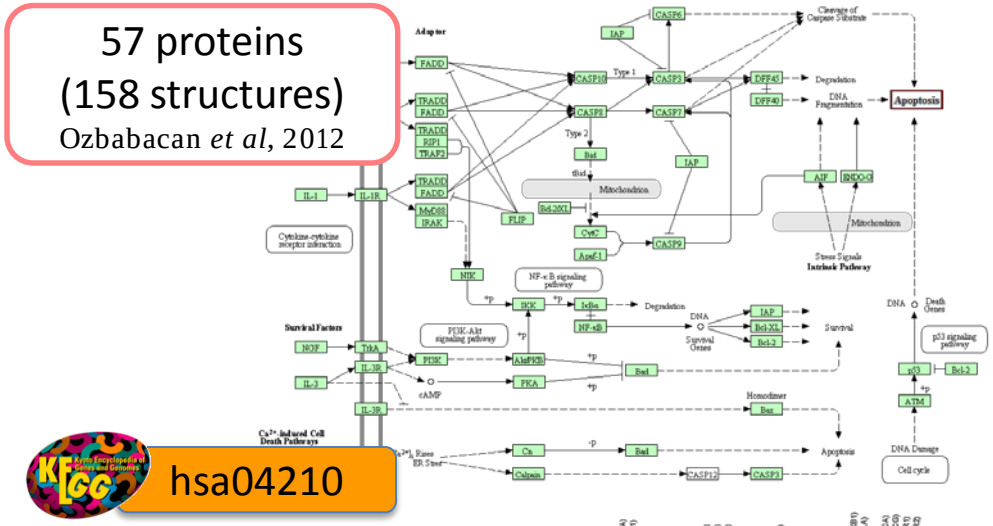


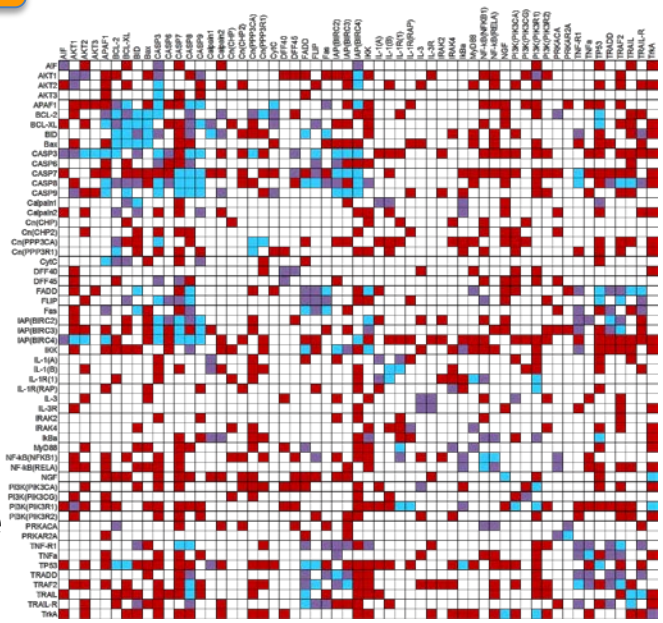
A suggestion of CheY-CheC interaction by using "False-positive" pair CheY-CheD as a mediator

Matsuzaki, *et al.*, *J Bioinform Comput Biol*, 2009.

Application to human apoptosis pathway

57 proteins
(158 structures)
Ozbabacan *et al*, 2012





True Positive

False Positive

False Negative

F-measure : 0.28

PPI prediction by docking without any other knowledge showed comparable results to template-based search of interaction partners (F-measure 0.30, Ozbabacan *et al.*, *J Struct Biol*, 2012).

Prediction	Interacting	No Interaction
Positive	88	364
Negative	96	1105

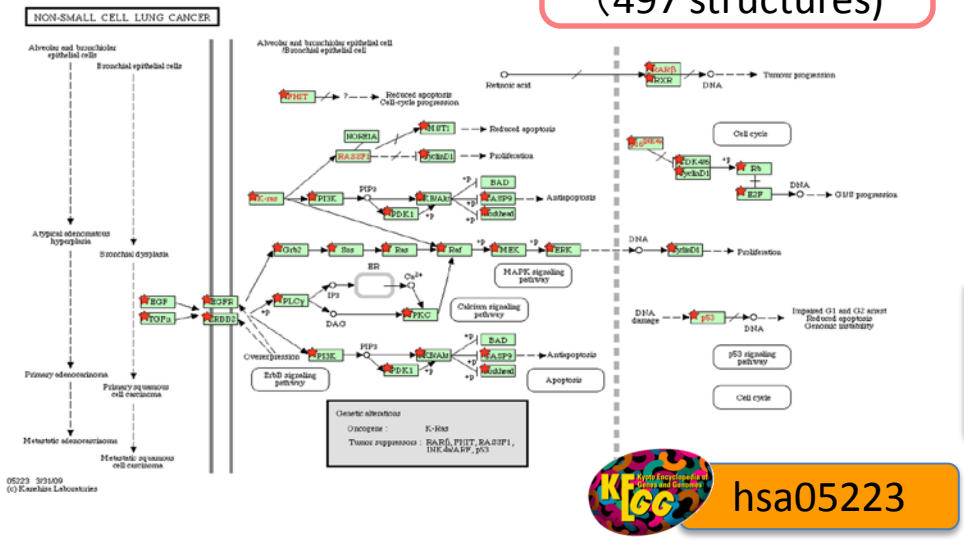
Ohue, *et al.*, *BMC Proc.*, *2013*.

Application to non-small cell lung cancer pathway

- Completed large-scale exhaustive docking
  - 497 structures, all-to-all docking
    ＝247,009 structure pairs
- Achieved high PPI prediction performance
  - Precision 0.29
  - Recall 0.47
  - F-measure 0.36

PPI prediction of about 250 thousand structure pairs showed comparable performance to the application to bacterial chemotaxis (10 thousand pairs).

44 proteins
(497 structures)



hsa05223

Prediction	Interacting	No Interaction
Positive	53	131
Negative	59	747

Counts are based on protein species

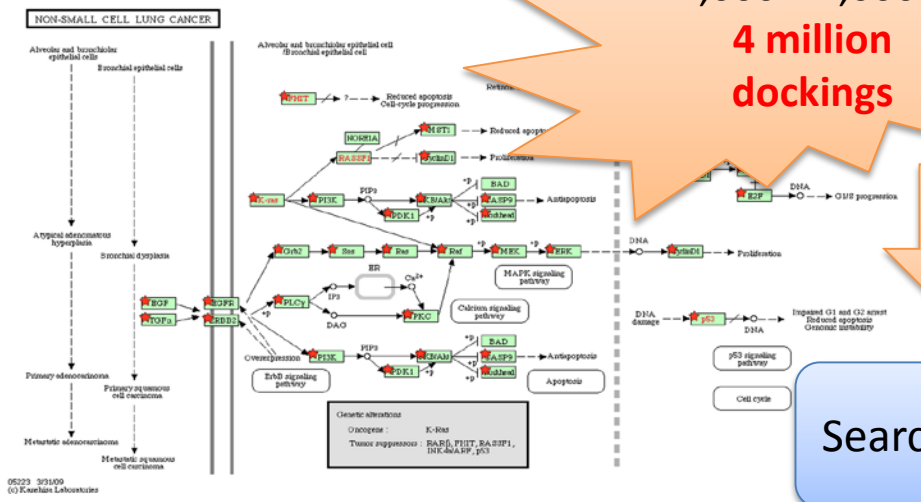
Application to lung-cancer drug related proteins

EGFR pathway related to non-small cell lung cancer

+

Proteins related to Gefitinib estimated by Miyano lab., The Univ. of Tokyo, from microarray analysis

44 proteins
(497 structures)



2,000 x 2,000 =
4 million dockings

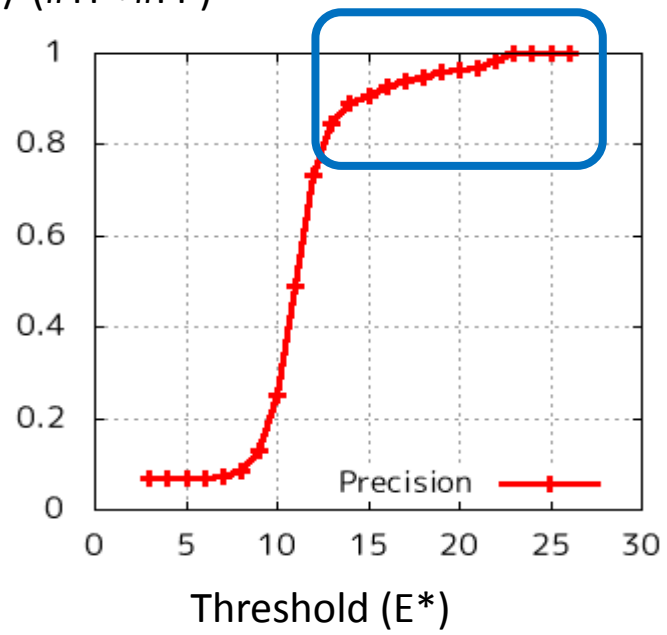
294 proteins
(1424 structures)

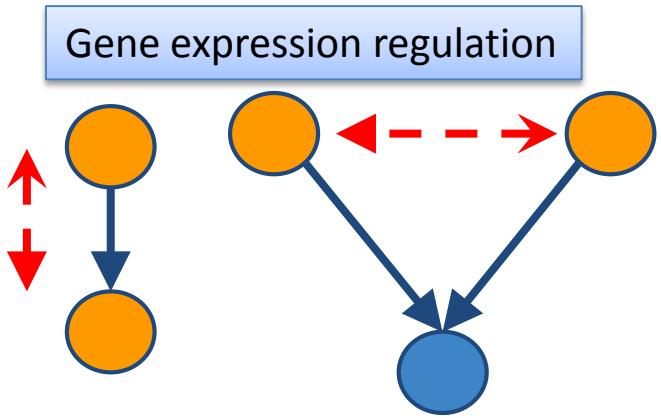
#TP / (#TP+#FP)

Search novel cancer related PPIs

PPI prediction result

Precision
#TP / (#TP+#FP)





- Using threshold of E* = 13.0
  - 3873 structure pairs
  - 175 protein pairs
  - Evaluated the prediction by 6 public databases (MIPS, DIP, IntAct, HPRD, BioGRID, MINT)
  - Undefined positives
    35 pairs
- Looked up these pairs on cancer gene regulatory networks derived by correlation of transcription data
  - Selected highly correlated pairs
  - Obtained 11 pairs

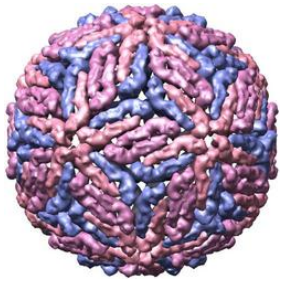
Evaluation of 7 potential PPIs by SPR

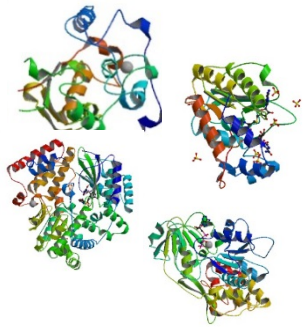
- 7 pairs were sent to assay using surface plasmon resonance (SPR) spectroscopy
  - Reference Biolabs Inc., Korea
  - Device: Reichert SR7500DC
- Binding affinities were measured except from 1 pair

Reference Biolobs Inc.

[Ongoing] Virus-human PPI prediction

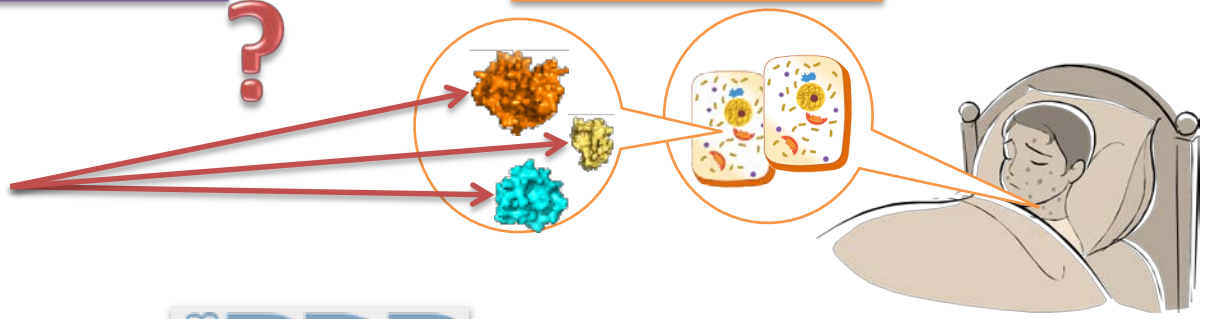
4 major enzymes of Dengue virus





?

Human proteins



Collected human proteins data:

- ✓ >25 residues
- ✓ X-ray resolution better than 3.25 Å
- ✓ No mutation

Enzyme
Protease
Methyltransferase
Polymerase
Helicase

#Proteins (UniProt IDs)	3,353
#Structure data (PDB-chains)	30,544

(2013-06-15)

$4 \times 30{,}544 = 122{,}176$ dockings
Suggested 28 potential interactions

References

(Papers contributed by MEGADOCK group)

- Ohue M, Shimoda T, Suzuki S, Matsuzaki Y, Ishida T, Akiyama Y, MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers., *Bioinformatics,* accepted.
- Matsuzaki Y, Ohue M, Uchikoga N, Akiyama Y, Protein-protein interaction network prediction by using rigid-body docking tools: application to bacterial chemotaxis., *Protein and Peptide Letters, **21**:790-798, 2014*.
- Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y, MEGADOCK: An all-to-all protein-protein interaction prediction system using tertiary structure data., *Protein and Peptide Letters, **21**:766-778, 2014*.
- Matsuzaki Y, Uchikoga N, Ohue M, Shimoda T, Sato T, Ishida T, Akiyama Y, MEGADOCK3.0: A high-performance protein-protein interaction prediction software using hybrid parallel computing for petascale supercomputing environments., *Source Code for Biology and Medicine,* **8**:18, 2013.
- Uchikoga N, Matsuzaki Y, Ohue M, Hirokawa T, Akiyama Y, Improved post-processing of protein-protein docking data using profiles of interaction fingerprints., *PLoS ONE, **8**:e69365, 2013.*
- Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y, Highly precise protein-protein interaction prediction based on consensus between template-based and *de novo* docking methods., *BMC Proceedings, **7***(Suppl. 7):S6, 2013.
- Ohue M, Matsuzaki Y, Ishida T, Akiyama Y, Improvement of the protein-protein docking prediction by introducing a simple hydrophobic interaction model: an application to interaction pathway analysis., *Lecture Note in Bioinformatics,* **7632**:178-187, 2012.
- Ohue M, Matsuzaki Y, Akiyama Y, Docking-calculation-based method for predicting protein-RNA interactions*., Genome Informatics*, **25**:25-39, 2011.
- Fleishman SJ, *et al*., Community-wide assessment of protein-interface modeling suggests improvements to design methodology., *Journal of Molecular Biology*, **414**:289-302, 2011.
- Uchikoga N, Hirokawa T, Analysis of protein-protein docking decoys using interaction fingerprints: application to the reconstruction of CaM-ligand complexes., *BMC Bioinformatics*, **11**:236, 2010.
- Ohue M, Matsuzaki Y, Matsuzaki Y, Sato T, Akiyama Y, MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data and its application to systems biology study., *IPSJ Transactions on Mathematical Modeling and Its Applications*, **3**: 91-106, 2010.
- Matsuzaki Y, Matsuzaki Y, Sato T, Akiyama Y, *In silico* screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis., *Journal of Bioinformatics and Computational Biology,* **7**:991-1012, 2009.