An Efficient Parallel Algorithm for Accelerating Computational Protein Design

Yichao Zhou (周奕超)[1], Wei Xu (徐葳)[1], Bruce R. Donald[2], and Jianyang Zeng (曾坚阳)[1,*]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University,
[2]Department of Computer Science and Department of Biochemistry, Duke University

清華大學 Tsinghua University

Abstract

Structure-based computational protein design is an important topic in protein engineering. Under the assumption of a rigid backbone and a finite set of discrete conformations of side-chains, various methods have been proposed to address this problem. A popular method is to combine the dead-end elimination (DEE) and A* tree search algorithms, which provably finds the Global Minimum Energy Conformation (GMEC) solution.

In our work, we present a variant of A* algorithm in which the search process can be performed on a single GPU in a massively parallel fashion. In addition, we made some efforts to address the memory exceeding problem in A* search. As a result, our enhancements can achieve a significant speedup of the A*-based protein design algorithm by four orders of magnitude. We also show that our parallel A* search algorithm could be successfully combined with iMinDEE, an state-of-the-art DEE criterion for rotamer pruning to further improve structure-based computational protein design with the consideration of continuous side-chain flexibility. Our software is available and distributed open-source under the GNU Lesser General License 2.1.

Problem Description

Given a desired 3-dimensional protein structure. What is a sequence that will fold to that structure?

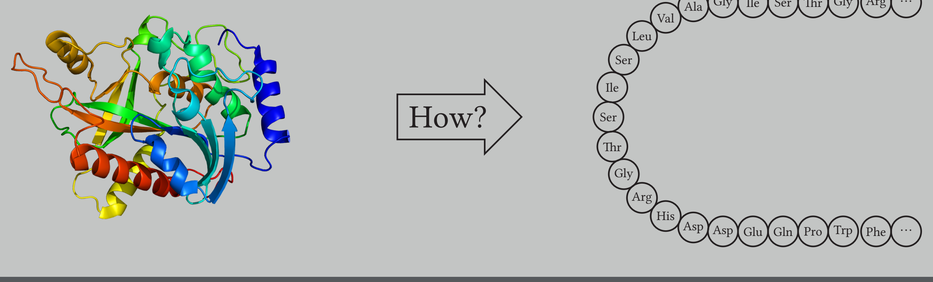
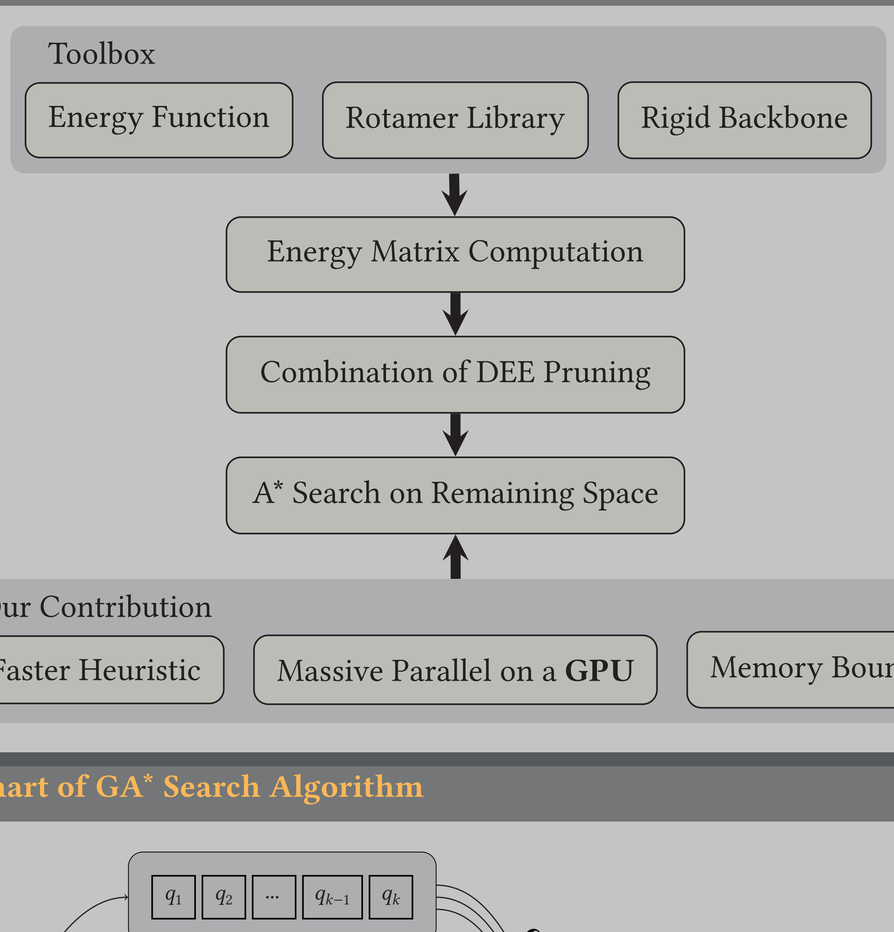


Diagram of DEE/A* Method



Flow chart of GA* Search Algorithm

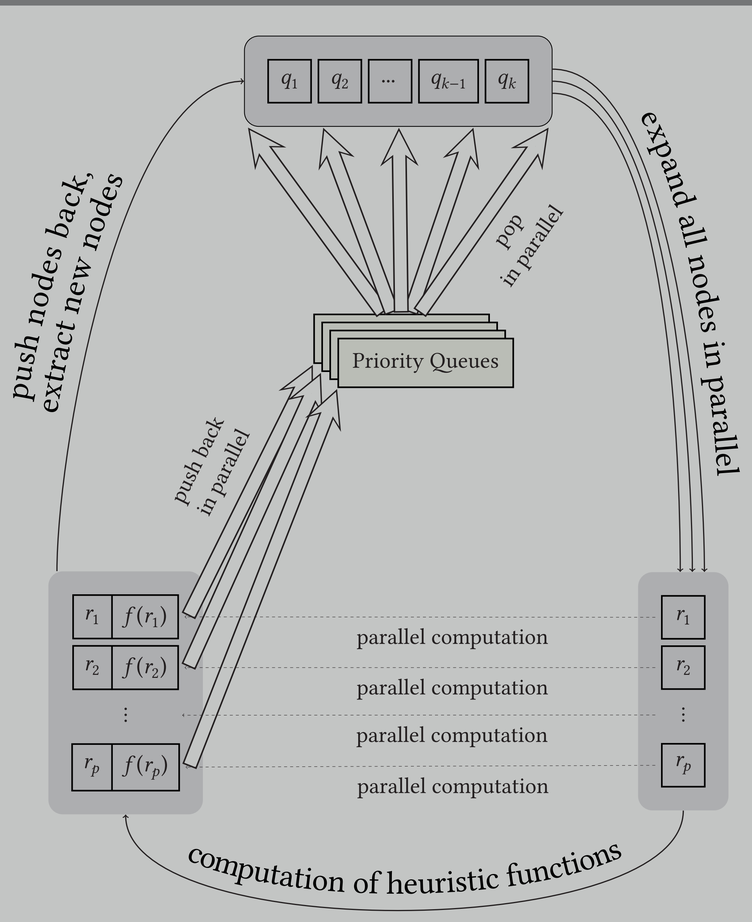


Figure: Flow chart of our GA* search algorithm for accelerating protein design. Symbols r_i represent all parallel expanded rotamers, and p is the total number of expanded nodes. A shaded rounded square represents a global state, which can be regarded as a global synchronization point. The directional black edges mean that the procedure needs to be done between two synchronization points. The dashed arrows and the double bold arrows represent the data flow among different states and the priority queues, respectively. A group of similar arrows means that the operations are performed in parallel.

Experiment: Environment

- Redesign the cores of wild proteins
- Test the speed and memory of GA*
- Test the *correctness* of GSMA*
- CPU: Intel Xeon™ E5-1620 3.6GHz
- GPU: NVIDIA Tesla K20C
  - 4.8G global memory
  - 2496 CUDA cores

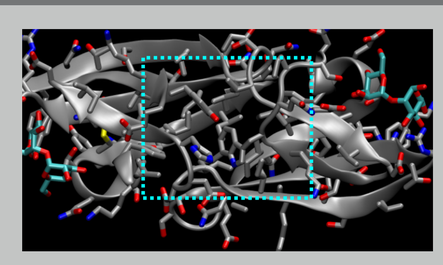


Figure: The core keeps the structure stable

Results: Performance of GA*

PDB	Space	OSPREY	A*	GA*768	GA*4992
2QCP	$2 \cdot 10^{17}$	21551916	51091	3075	1146
1XMK	$2 \cdot 10^{14}$	247585	2990	296	121
1X6I	$7 \cdot 10^{13}$	96990	1406	138	73
1UCS	$6 \cdot 10^{12}$	88135	1771	182	79
1CC8	$3 \cdot 10^{14}$	77614	1078	99	53
2CS7	$8 \cdot 10^{12}$	64187	1154	149	57
2BWF	$9 \cdot 10^{13}$	18457	307	33	24
1I27	$7 \cdot 10^{11}$	8151	88	18	16
1T8K	$2 \cdot 10^{13}$	6806	89	18	15
1R6J	$2 \cdot 10^{14}$	6018	107	18	21

Results: Performance of GSMA*

PDB		1OAI	1U2H	1ZZK	2CS7	2DSX	3D3B
# of Mutable Residues		16	18	14	15	15	15
Conformation Space		$2{\cdot}10^{22}$	$2{\cdot}10^{20}$	$2{\cdot}10^{15}$	$2{\cdot}10^{23}$	$3{\cdot}10^{20}$	$6{\cdot}18^{18}$
GA*768 Search Space		$4{\cdot}10^{7}$	$8{\cdot}10^{6}$	$8{\cdot}10^{6}$	$4{\cdot}10^{7}$	$4{\cdot}10^{7}$	$3{\cdot}10^{7}$
$3{\cdot}10^{4}$ limit	Scan Count	252	104	99	202	182	109
	GMEC Gotten	NO	YES	YES	NO	YES	NO
	GMEC Assured	NO	NO	NO	NO	NO	NO
	Correctness	4%	100%	20%	12%	32%	6%
	Recover Ratio	62%	75%	85%	48%	46%	48%
$3{\cdot}10^{5}$ limit	Scan Count	139	43	36	103	97	55
	GMEC Gotten	YES	YES	YES	YES	YES	YES
	GMEC Assured	NO	YES	YES	NO	NO	NO
	Correctness	100%	100%	100%	100%	100%	44%
	Recover Ratio	74%	75%	87%	46%	48%	54%
$3{\cdot}10^{6}$ limit	Scan Count	22	3	3	24	22	18
	GMEC Gotten	YES	YES	YES	YES	YES	YES
	GMEC Assured	YES	YES	YES	YES	YES	YES
	Correctness	100%	100%	100%	100%	100%	100%
	Recover Ratio	74%	75%	87%	46%	48%	53%
$3{\cdot}10^{7}$ limit	Scan Count	1	0	0	1	1	1
	GMEC Gotten	YES	YES	YES	YES	YES	YES
	GMEC Assured	YES	YES	YES	YES	YES	YES
	Correctness	100%	100%	100%	100%	100%	100%
	Recovery Ratio	74%	75%	87%	46%	48%	53%

Conclusions and Future Work

We have developed an innovative method to improve the A* algorithm for computational protein design, which reduces running time by up to *four orders of magnitude*.

It is interesting to know whether it can achieve similar performance on an affordable GPU card. Also we can port our algorithm to the existing large clusters of CPUs and GPUs, where we may solve a larger design problem than ever before.

Reference

Yichao Zhou, Wei Xu, Bruce R. Donald, Jianyang Zeng*. **An Efficient Parallel Algorithm for Accelerating Computational Protein Design**. *Proceedings of the 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2014)*. Boston, Massachusetts, USA, July 2014. *Bioinformatics*. To Appear.

Funding

This work is supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003. This work is supported by a grant to B.R.D. from the National Institutes of Health (R01 GM-78031).