Interconnected Cost Function Networks (iCFN): an efficient exact algorithm for multistate protein design

MOSTAFA KARIMI[1,2] , YANG SHEN[1,2]

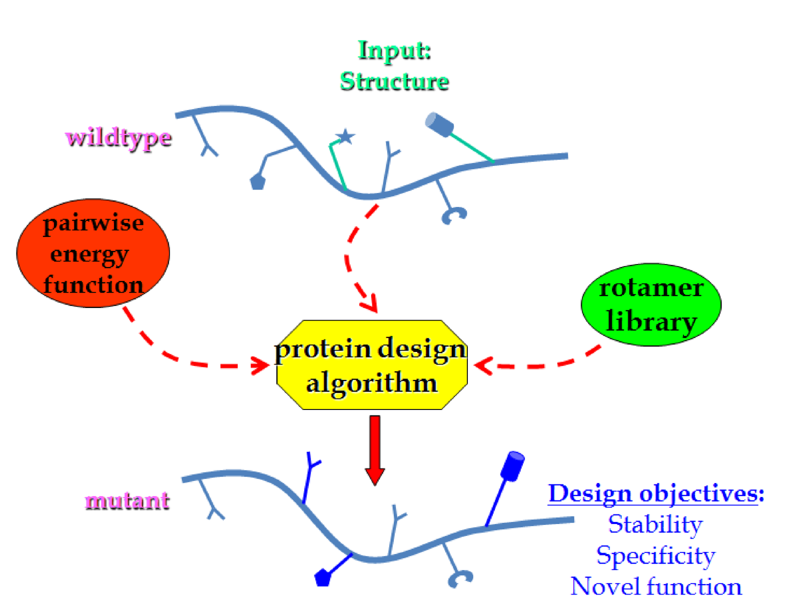
[1]Department of Electrical and Computer Engineering, [2]TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, USA.

Abstract

- We work on **multistate protein design problems** that address both positive and negative **states** and consider an ensemble of biophysical **substates** such as a protein being in various backbone conformers, unbound or bound to a target, or bound to various (off-)targets.
- The **generic formulation** allows for many applications such as stability, affinity, and specificity design.
- iCFN is an **exact algorithm** that guarantees the optimal solutions and near-optimal ensembles thus enable informative interaction with experiments. Its efficiency makes large-scale designs more tractable.
- Its **application** to T-cell receptor (TCR) design for specificity generates experimentally-agreeing results and reveals underlying mechanisms.

Availability: https://shen-lab.github.io/software/iCFN

Problem Definition



Protein design (Figure Credit: Ivelin Georgiev)

Energy model (Assumption: pairwise additive)

$$f(\mathbf{r}) = c + \sum_i E(i_r) + \sum_{i<j} E(i_r, j_s)$$

A generic formulation for multi-state protein design:

$$\mathbf{s}^* = arg\min_{\mathbf{s}\in\mathscr{S}} \left(\min_{p\in\mathscr{P}} \min_{\mathbf{r}\in\mathscr{R}^+(\mathbf{s})} f_p^+(\mathbf{r}) - \min_{q\in\mathscr{Q}} \min_{\mathbf{r}\in\mathscr{R}^-(\mathbf{s})} f_q^-(\mathbf{r}) \right)$$

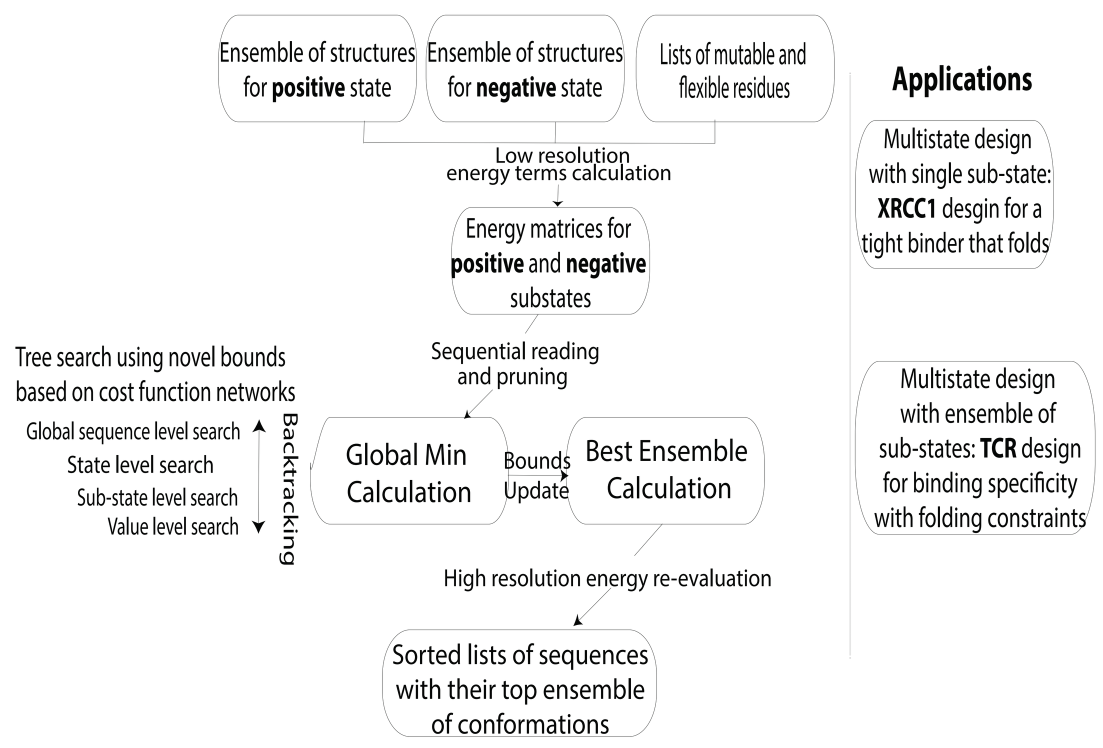
s.t. Constraints on substate functions $f_p^+(\mathbf{r})$ & $f_q^-(\mathbf{r})$

iCFN's approach to solving the NP-hard problem:

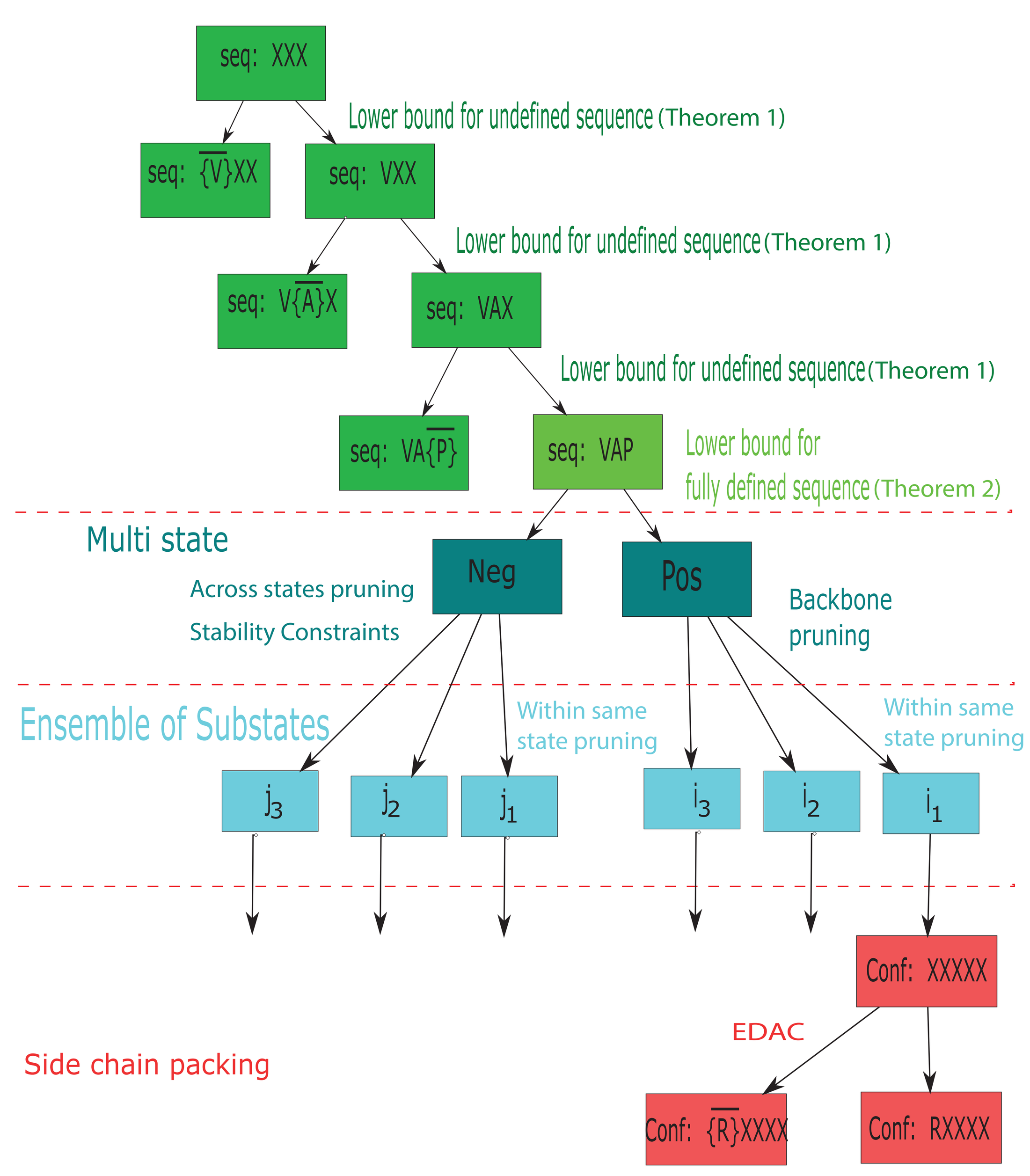
- Each substate design is formulated as a **Weighted Constraint Satisfaction Problem** (WCSP) and modeled by a **Cost Function Network** (CFN) $(\mathscr{X}, \mathscr{D}, \mathscr{C})$;
- The coupled WCSPs are represented as CFNs interconnected over a **tree** of sequences, substates, and rotamers (values);
- Novel lower **bounds** are developed for the sequence (variable) space (Thm. 1,2) and Existential Directional Arc Consistency (EDAC) is exploited for the rotamer (value) space;
- **Depth First Branch and Bound** (DFBB)-based tree search allows positive and negative designs to inform each other and substates within and across states to prune each other.

Methods

Overall scheme



Tree search



Theorem 1. *Lower bound for any undefined sequence S:*

$$\min_{(k,l)} \left(\Delta c_{kl} + \sum_i \min_{a\in S(i)} \min_{(r,r')} \left(\Delta E_{kl}(i_{r,r'}) + \sum_{j>i} \min_{a'\in S(j)} \min_{(s,s')} \Delta E_{l,k}(i_{r,r'}, j_{s,s'})\right)\right)$$

with complexity $O(n^2R^2a^2r)$ (where n is the number of positions, R the average number of rotamers per position, a the average number of substates per state, and r the average number of rotamers per amino acid).

Theorem 2. *Lower bound for any defined sequence S:*

$$\min_{k\in\mathscr{P}} L_k(S) - \min_{l\in\mathscr{Q}} U_l(S)$$

in which $L_k(S)$ is EDAC for sequence S in the k^{th} substate and U_l is LDS for sequence S in the l^{th} substate.

Additional bounds for each substates, across substates of the same state, and across substates of different states.

Results

Multi-state XRCC1 design with a single substate per state:

N_{mut}	d(Å)	N_{flex}	Pre-DEE Size	Post-DEE Size (Ensemble)	ε = 0.5 Kcal/mol COMETS	Reduced iCFN	iCFN	1 kcal/mol COMETS	Reduced iCFN	iCFN	1.5 Kcal/mol COMETS	Reduced iCFN	iCFN
1	3	9	2.88×10^{17}	2.04×10^{8}	4.03	0.02	0.02	4.27	0.03	0.3	4.37	0.03	0.03
1	6	16	5.24×10^{30}	2.75×10^{19}	6.84	0.16	0.12	6.97	0.17	0.12	7.72	0.19	0.13
2	3	10	1×10^{22}	1.73×10^{12}	6.85	0.28	0.15	8.36	0.28	0.16	9.29	0.28	0.16
2	6	19	5.88×10^{37}	1.58×10^{25}	19.46	2.63	1.41	29.07	2.67	1.44	29.27	2.79	1.47
3	3	11	7.94×10^{27}	8.31×10^{18}	M	12.64	6.15	M	12.67	6.51	M	12.65	6.54
3	6	20	1.81×10^{43}	1.44×10^{30}	M	62.17	32.9	M	62.22	33.26	M	62.31	33.47
4	3	14	6.54×10^{36}	8.31×10^{26}	M	493.26	268.28	M	493.31	268.3	M	493.95	268.41
4	6	26	7.94×10^{56}	4.36×10^{38}	M	2060	1458	M	2156	1493	M	2161	1498
5	3	15	3.54×10^{42}	2.34×10^{30}	M	9810	5570	M	9943	6005	M	10046	6040
5	6	26	1.65×10^{61}	1.58×10^{42}	M	49373	37198	M	49978	37223	M	50405	39553

("M" indicates an error for being out of a 20Gb-memory limit whereas iCFN used at most 80Mb)

iCFN outperforms COMETS, the only other exact method for multistate design, in both memory usage and CPU time, which enables large designs in practice.

Multi-state TCR design with ensembles of substates (MD simulated)
(target peptide: AAG; off-target peptide: ELA)
Guaranteed near-optimum ensemble:

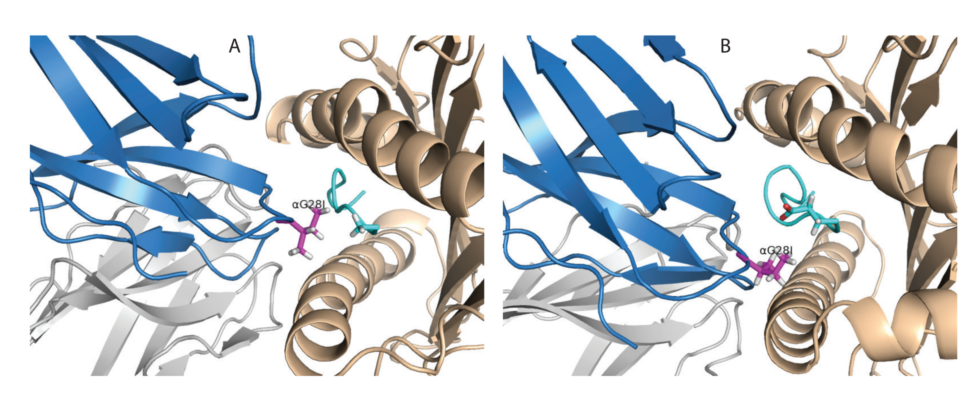
Position	Pre-DEE Size	Post-DEE Size	Reduced iCFN Nodes Expanded	Leaves Visited	Sequences	Time (s)	iCFN Nodes Expanded	Leaves Visited	Sequences	Time (s)
26	10^{61}	10^{54}	6.35×10^{4}	6.20×10^{4}	26	66.69	6.14×10^{3}	6.00×10^{3}	10	21.86
28	10^{66}	10^{61}	5.92×10^{4}	5.70×10^{4}	26	114.22	4.09×10^{3}	4.00×10^{3}	2	23.55
98	10^{58}	10^{55}	5.15×10^{4}	5.00×10^{4}	25	103.29	4.16×10^{4}	4.00×10^{4}	20	43.35
100	10^{84}	10^{77}	7.43×10^{4}	7.11×10^{4}	26	154.51	5.19×10^{3}	5.00×10^{3}	2	23.74
26,28	10^{87}	10^{82}	1.70×10^{6}	1.62×10^{6}	676	7454.93	9.44×10^{3}	9.00×10^{3}	4	1063.89
26,98	10^{119}	10^{111}	1.82×10^{6}	1.73×10^{6}	650	15449.04	2.51×10^{5}	2.38×10^{5}	108	3872.32
26,100	10^{142}	10^{132}	1.89×10^{6}	1.75×10^{6}	676	19780.68	2.62×10^{4}	2.40×10^{4}	10	2226.52
28,98	10^{126}	10^{119}	1.45×10^{6}	1.37×10^{6}	650	23378.51	3.13×10^{4}	3.00×10^{4}	13	2810.31
28,100	10^{141}	10^{132}	1.77×10^{6}	1.60×10^{6}	676	24631.34	4.22×10^{3}	4.00×10^{3}	2	2359.10
98,100	10^{112}	10^{106}	1.60×10^{6}	1.51×10^{6}	650	17303.91	3.98×10^{4}	3.80×10^{4}	19	2056.47
26,28,98	10^{146}	10^{141}	—	—	16900	—	5.86×10^{4}	5.50×10^{4}	27	105343
26,28,100	10^{161}	10^{154}	—	—	17576	—	1.48×10^{4}	1.40×10^{4}	6	99012
26,98,100	10^{169}	10^{161}	—	—	16900	—	6.76×10^{4}	6.00×10^{4}	27	185886
28,98,100	10^{168}	10^{162}	—	—	16900	—	3.73×10^{4}	3.40×10^{4}	12	158995

- iCFN visits on average 6.7 (7.4), 58.8 (110.8), and 455.1 (1397.2) times less sequences for the best single (ensemble of) sequence(s) in single, double, and tripe designs, respectively.
- iCFN runs 3.4 (4.2) and 5.9 (7.8) times faster than reduced iCFN does for global optimum (top ensemble) in an average single and double design, respectively; and it solves tripe designs within 1∼2 CPU days whereas reduced iCFN could not within 1 CPU week.
- iCFN's relative computational gain increases as complexity increases!

Design accuracy:

Method	True Positive (TP)	FP	False Negative (FN)
Rosetta	G28I, G28L, G28Y, F100W	D26Y	D26W, F100Y
Rosetta Min	G28I, G28L, G28Y	N/A	D26W, F100W, F100Y
iCFN	**D26W, G28I, G28L, G28Y, F100W, F100Y**	**D26Y**	**N/A**

Molecular mechanisms of AAG-binding specificity for G28I:



Differential effects of G28I to (A) AAG-binding and (B) ELA-binding revealed in iCFN structural models. Cartoons: DMF5 α; DMF5 β; AAG/ELA peptides; MHC α chain. Stick: αG28I.
Worse vdW packing and continuum electrostatics upon mutation for N-terminal glutamate of ELA but not for N-term alanine of AAG.

References

[1] Mostafa Karimi, Yang Shen. "iCFN: an efficient exact algorithm for multistate protein design", Bioinformatics 34 (17), i811-i820.

Acknowledgements

This project is in part supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM124952 to YS) and the Defense Advanced Research Projects Agency (FA8750-18-2-0027 to ZW). Most of the computing time is provided by the Texas A&M High Performance Research Computing.