Boosting AND/OR-Based Computational Protein Design: Dynamic Heuristics and Generalizable UFO

Bobak Pezeshki (pezeshkb@uci.edu), Radu Marinescu (radu.marinescu@ie.ibm.com), Alexander Ihler (ihler@ics.uci.edu), Rina Dechter (dechter@ics.uci.edu)

Overview

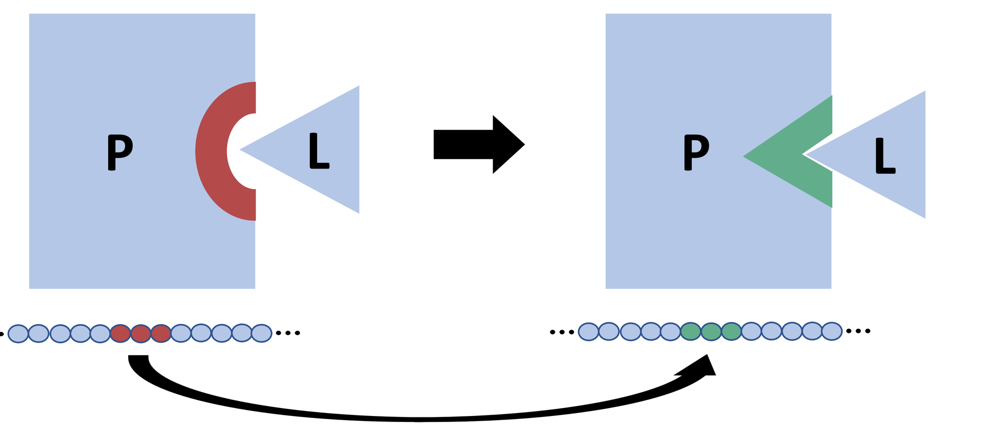
Recently a new protein re-design algorithm, AOBB-K*, was introduced and was competitive with state-of-the-art BBK* on small protein re-design problems. However, AOBB-K* did not scale well. In this work, we focus on modifications to AOBB-K* that significantly enhance scalability.

Contributions:

1. **AOBB-K*-b (boosted):** AOBB-K* with stronger wMBE-K* heuristic and modifications to search
2. **AOBB-K*-DH:** AOBB-K* with dynamic heuristics.
3. **UFO:** An approximation scheme that introduces determinism to empower constraint propagation
4. **AOBB-K*-UFO:** UFO empowered AOBB-K*
5. **Empirical analysis:** Evaluation on 62 real protein benchmarks comparing with previous AOBB-K* and state-of-the-art BBK*.

Problem

Redesign of proteins to form higher affinity complexes



Find new amino acid assignments to residues of interest that optimize affinity between interacting subunits

K* Objective

An approximation of binding affinity between molecules (based on the biological association constant known as K_a)

$$K^*(r) = \frac{Z_{PL}(r)}{Z_P(r)\, Z_L(r)}$$

$$Z_\gamma(r) = \sum_{c \in C_\gamma(r)} \exp\{-E_\gamma(c)/\mathcal{RT}\}$$

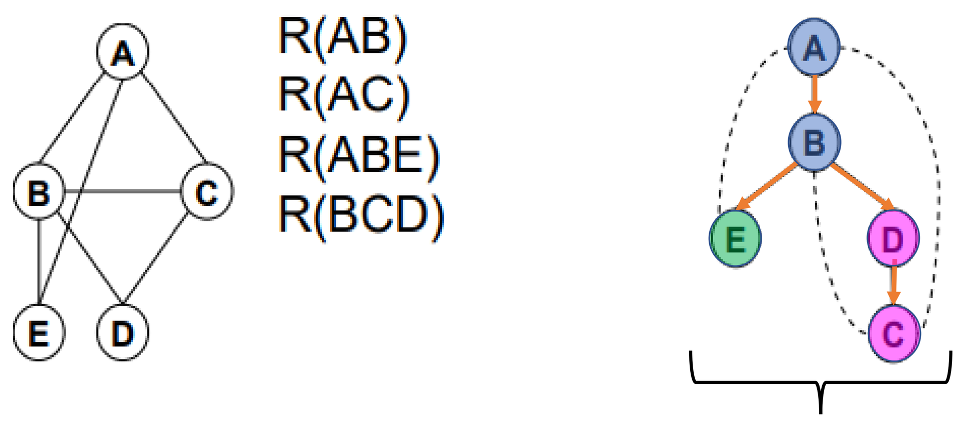
captures the favorability of the subunit(s) in form $\gamma \in \{P, L, PL\}$

r = amino acid assignments to residues
C(r) = possible conformations given r
E(c) = energy given conformation c
$\mathcal{R}$ = universal gas constant
$\mathcal{T}$ = absolute temperature (Kelvin)

AND/OR Search

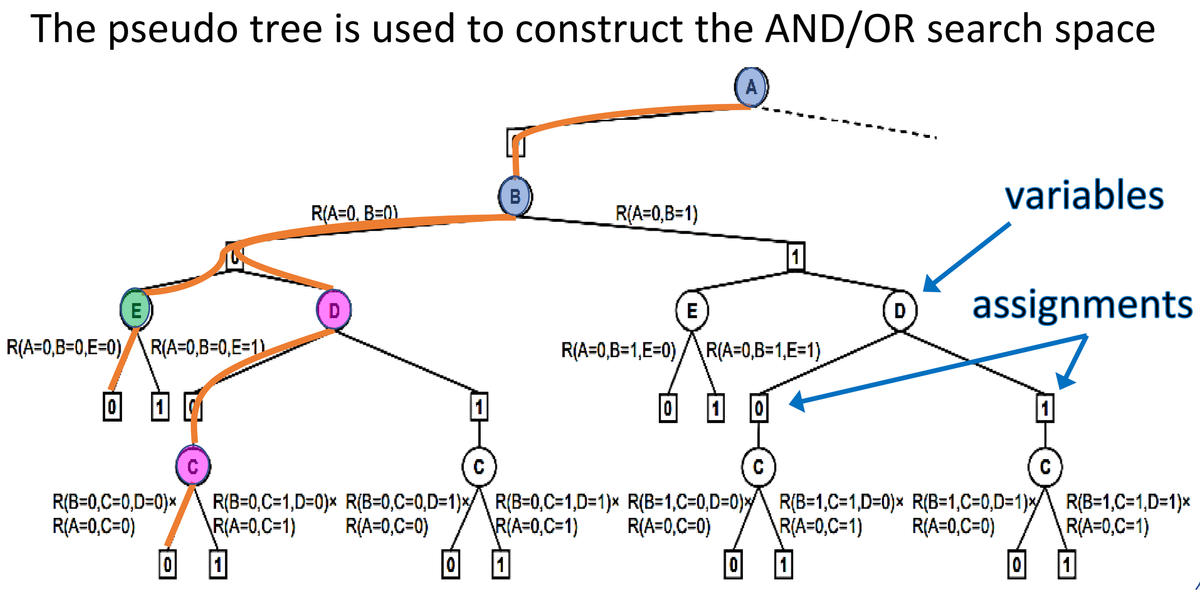
Compact search space taking advantage of conditional independences present in the model

Graphical Model Network | Possible Pseudo Tree



Directed tree (based on a variable ordering) that branches when conditional independences exist given assignments to ancestors.

The pseudo tree is used to construct the AND/OR search space



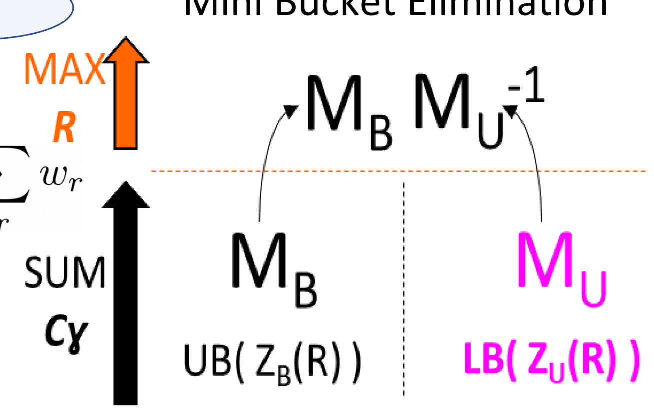
wMBE-K*

K* heuristic to guide search, based on variable elimination message-passing scheme: Mini Bucket Elimination

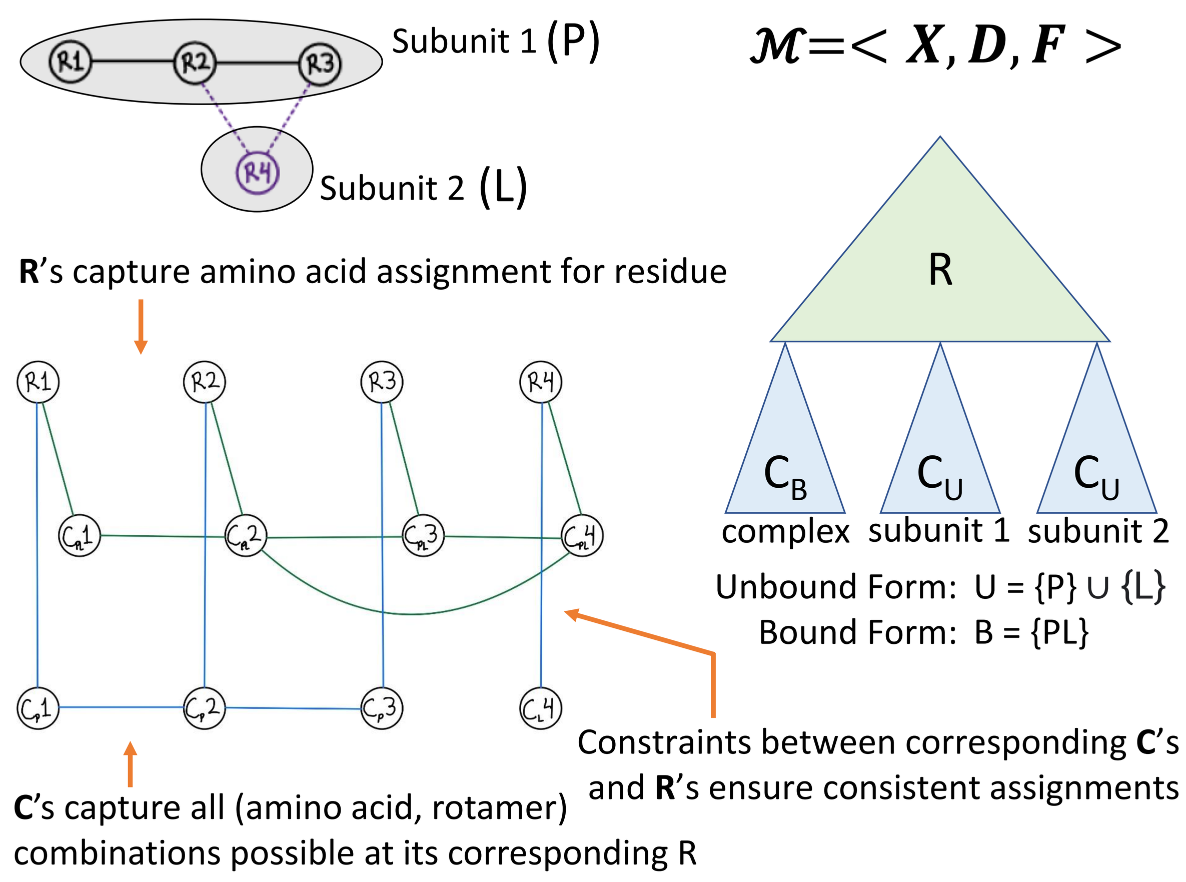
$$K^*(r) = \frac{Z_{complex}(r)}{Z_{subunit\,1}(r)\, Z_{subunit\,2}(r)}$$

$$\sum^{w}[\prod_r(\psi_r)] \le \prod_r[\sum^{w_r}] \quad w = \sum_r w_r$$

$$\sum_x^{w} f(x) \triangleq [\sum_x f(x)^{\frac{1}{w}}]^w$$



Graphical Model Formulation



K*MAP Task

Find amino acid assignments to the residues that maximize K*

$$Z_\gamma(R_1 \ldots R_N) = \sum_{C_1,\ldots,C_N} \prod_{\mathscr{C}_{\gamma(i)} \in \mathscr{C}} \mathscr{C}_{\gamma(i)}(R_i, C_{\gamma(i)}) \cdot \prod_{E^{sb}_{\gamma(i)} \in \boldsymbol{E}^{sb}_\gamma} e^{-\frac{E^{sb}_{\gamma(i)}(C_{\gamma(i)})}{\mathscr{RT}}} \cdot \prod_{E^{pw}_{\gamma(ij)} \in \boldsymbol{E}^{pw}_\gamma} e^{-\frac{E^{pw}_{\gamma(ij)}(C_{\gamma(i)}, C_{\gamma(j)})}{\mathscr{RT}}}$$

$$K^*(R_1, \ldots, R_N) = Z_B(R_1, \ldots, R_N) / Z_U(R_1, \ldots, R_N)$$

$$task{:}\ \ K^*MAP = \max_{R_1,\ldots,R_N} K^*(R_1, \ldots, R_N)$$

AOBB-K*

- ❑ Branch-and-bound algorithm over AND/OR search spaces
  - ▪ AOBB-K* is exact
- ❑ Can use wMBE-K* to guide search
- ❑ Exploits determinism by using constraint propagation
- ❑ Incorporates a global constraint enforcing biologically relevant solutions

Subunit Stability Constraints

Condition to enforce the stability of each subunit to be no less than a given threshold from that of the wild-type stability

$$Z_{subunit\ i}(r) > Z_{subunit\ i}(r^{wt}) * \exp\{-5/\mathcal{RT}\}$$

Stability of naturally occurring version — Constant factor for thresholding

Infusing Determinism: τ-Underflows

Replace unfavorable assignments with hard constraints to exploit the strength of constraint propagation during search

- non-negative function f
- $\tau \in \mathbb{R}^+$

$$f_\tau(x) = \begin{cases} f(x), & f(x) \ge \tau \\ 0, & otherwise \end{cases}$$

wMBE-K*-b

wMBE-K* with sequential modifications to improve estimates at the cost of losing bound guarantees.

1) Lower bounding power-sum replaced with zero-omitted power-sum:

$$f^{\triangleleft w}(y) := f(y)^w \text{ for } f(y) \ne 0 \text{ and } 0 \text{ otherwise}$$

$$\sum\nolimits_X^{\triangleleft w} f := (\sum\nolimits_X f(x)^{\triangleleft \frac{1}{w}})^w \text{ where } \tfrac{0}{0} := 0$$

2) Cost-shifting with only non-zero values:

$$\boldsymbol{F}_\lambda = \boldsymbol{\mathcal{C}} \cup \boldsymbol{E^{sb}} \cup \boldsymbol{E^{pw}} \cup \boldsymbol{\Lambda},$$

$$\lambda^{(i)}_\Omega \in \boldsymbol{\Lambda}, \quad \prod_{\omega \in D_\Omega} \prod_i \lambda^{(i)}_\Omega(\omega) = 1, \quad \lambda^{(i)}_\Omega(\boldsymbol{\Omega}) > 0$$

3) Maximization step prioritizes finite values:

$$max'_X(f) = \begin{cases} max_{x'} f(x'), & f(x') \ne \infty \\ \infty, & otherwise \end{cases}$$

AOBB-K*-b

AOBB-K* with search tuned to find good solutions sooner

Value-ordering of nodes modified so that:

1) The wild-type assignment is prioritized first, ensuring a strong initial lower bound.
2) Nodes with finite wMBE-K*-b upper bounds are explored first, prioritizing consistent solutions

AOBB-K*-DH

AOBB-K* with a dynamic heuristic scheme

Dynamically tightens heuristic during search when:

1) $ub(K^*) > dhThreshold \in (0, \infty)$
2) $depth \le maxDepth \in \mathbb{I}^+$

UFO

input : Graphical model $\mathcal{M}$; SAT solving algorithm, $SAT(.)$; a deflation factor $\delta \in (0, 1]$
output : A proposed threshold τ to use

```
begin
  τ_min = 0;  τ_max = max_{F,X} f(x)
  τ = (τ_max + τ_min) / 2
  while time remains for τ binary search do
    if SAT(M_τ) = False then τ_max = τ
    else τ_min = τ
    τ = (τ_max + τ_min) / 2
  τ = τ_min · δ
  return τ
```

CPD Empirical Analysis

Competing Scheme: BBK*

[Ojewole et al., 2018]

State-of-the-art protein redesign algorithm, part of the software package, OSPREY, developed for over ten years for protein design

- A*-like best first
- utilizes dynamic optimistic greedy heuristic
- approximate scheme w/ tightness parameter

Sample Empirical Results

Comparison of the AOBB-K*-b-[UFO/DH] and BBK* on problems with 4 or 5 mutable residues. For each 1hr experiment, we see: • **i** (i-bound) • **Soln** (best-found K*) • t_{best} (time best solution was first discovered) • t_{final} (completion time • **wt K*** (wild-type K* value)

The algorithms are displayed in a top-down ranking per problem. Ranking is based first on greater K* and then by faster Anytime times. **Large text** highlights values responsible for a higher ranking. and blue color indicates better performance vs. BBK*.

		AOBB-K*-b-[DH/UFO]						BBK*	
M	Problem	Algorithm	i	Soln	t_{best}	t_{final}	wt K*	Soln	Time
4	d7-4-2	AOBB-K*-b-UFO	3	**14.89**	3391.78	3600.00	14.08	14.54	278.08
		AOBB-K*-b-DH	3	**14.49**	3543.27	3600.00	14.08	14.54	278.08
		AOBB-K*-b	3	14.49	3293.62	3600.00	14.08	14.54	278.08
	d13-4-1	AOBB-K*-b-UFO	3	15.03	**12.69**	1974.43	13.25	15.03	46.46
		AOBB-K*-b-DH	3	15.03	**22.05**	79.88	13.25	15.03	46.46
		AOBB-K*-b	4	15.03	165.48	3600.00	13.25	15.03	46.46
	d17-4-1	AOBB-K*-b-UFO	3	10.86	**29.39**	3600.00	10.52	10.80	89.94
		AOBB-K*-b	4	10.86	**657.54**	3600.00	10.52	10.80	89.94
		AOBB-K*-b-DH	3	10.86	660.16	3600.00	10.52	10.80	89.94
	d21-4-1	AOBB-K*-b-UFO	3	11.92	**196.30**	3600.00	9.37	11.72	687.66
		AOBB-K*-b-DH	3	11.92	614.88	3600.00	9.37	11.72	687.66
		AOBB-K*-b	4	11.72	264.92	3600.00	9.37	11.72	687.66
	d43-4-1	AOBB-K*-b-UFO	3	18.19	76.49	484.69	18.04	18.18	119.88
		AOBB-K*-b-DH	3	18.19	**386.49**	3600.00	18.04	18.18	119.88
		AOBB-K*-b	3	18.19	896.67	3600.00	18.04	18.18	119.88
	d47-4-2	AOBB-K*-b-UFO	3	**22.87**	72.53	239.88	22.70	22.83	1339.15
		AOBB-K*-b	3	22.74	**130.95**	3600.00	22.70	22.83	1339.15
		AOBB-K*-b-DH	3	22.74	140.66	3600.00	22.70	22.83	1339.15
5	d7-5-1	AOBB-K*-b-UFO	3	**15.17**	1570.30	3600.00	14.08	14.73	401.09
		AOBB-K*-b-DH	3	14.73	**57.91**	3600.00	14.08	14.73	401.09
		AOBB-K*-b	3	14.73	62.53	3600.00	14.08	14.73	401.09
	d7-5-3	AOBB-K*-b-UFO	3	**14.84**	891.90	3600.00	14.08	15.60	205.56
		AOBB-K*-b	3	14.73	**67.53**	3600.00	14.08	15.60	205.56
		AOBB-K*-b-DH	3	14.73	156.68	3600.00	14.08	15.60	205.56
	d27-5-1	AOBB-K*-b	3	15.55	**274.30**	3600.00	15.48	15.55	1270.65
		AOBB-K*-b-UFO	3	15.55	**276.91**	3600.00	15.48	15.55	1270.65
		AOBB-K*-b-DH	3	15.55	321.02	3600.00	15.48	15.55	1270.65
	d31-5-1	AOBB-K*-b-UFO	3	7.88	**22.35**	128.75	7.63	7.88	130.04
		AOBB-K*-b	3	7.88	**129.43**	3600.00	7.63	7.88	130.04
		AOBB-K*-b-DH	3	7.88	145.63	3600.00	7.63	7.88	130.04
	d47-5-1	AOBB-K*-b-UFO	3	**23.08**	2068.22	3600.00	22.70	23.05	3600.00
		AOBB-K*-b	3	22.74	**222.66**	3600.00	22.70	23.05	3600.00
		AOBB-K*-b-DH	3	22.74	241.88	3600.00	22.70	23.05	3600.00

Summary

- ❑ Scalability to problems with 3, 4, and 5 mutable residues
- ❑ AOBB-K*-b-UFO shows particularly good performance
  - ❑ Competitive run-times
  - ❑ Good solution quality
- ❑ Competitiveness tapers off at 5 mutable residues

Additional Materials

Supplemental | Background | Benchmarks | UFO

Acknowledgements

Supported in part by NSF Grant IIS-2008516