AND/OR Branch-and-Bound for Computational Protein Design Optimizing K*

BOBAK PEZESHKI, RADU MARINESCU, ALEX IHLER, RINA DECHTER

1st Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE), AAAI-22
Computational Protein Design (CPD)

[Re]design proteins to perform desired biological functions.

CPD often manifests as an optimization problem:

*Ex. find the optimal composition that maximizes binding between subunits.*
Special thanks to…

THOMAS SCHIEX
Special thanks to...

JONATHAN JOU
GRAHAM HOLT
BRUCE DONALD
Key Contributions

- Two formulations as a graphical model for optimizing $K^*$.  
- wMBE-$K^*$ heuristic that can bound the optimal $K^*$ and guide search.  
- AOBB-$K^*$, a depth-first branch-and-bound algorithm for maximizing $K^*$ that uses a compact AND/OR search space.  
- An experiments on over 40 protein design problems as an empirical proof-of-concept.
Background
Review of Proteins

Each amino acid position = a residue

Primary Protein Structure
Sequence of a chain of amino acids

Secondary Protein Structure
Local folding of the polypeptide chain into helices or sheets

Tertiary Protein Structure
Three-dimensional folding pattern of a protein due to side chain interactions

Quaternary Protein Structure
Protein consisting of more than one amino acid chain
Review of Proteins

There Are ~20 Naturally Occurring Amino Acids

Chemical Structure
- NAME
  - single letter code
  - three letter code
  - DNA codons

Chart Key:
- ALIPHATIC
- AROMATIC
- ACIDIC
- BASIC
- HYDROXYLIC
- SULFUR-CONTAINING
- AMIDIC
- NON-ESSENTIAL
- ESSENTIAL

Cannot be made by the human body
Review of Proteins

Amino Acid Rotamers: Select conformational isomers of an amino acid

Peter Carlsson, Konrad F. Koehler, and Lennart Nilsson
Proteins are Dynamic Structures

**K* Objective**

\[ K^*(r) = \frac{Z_{PL}(r)}{Z_P(r)Z_L(r)} \]

(Note that \( K^* \) not only considers the “goodness” of the bonded state (PL), but also weighs it relative to the “goodness” of the unbound (dissociate) states.)

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

- amino acid assignments to the residues
- possible rotamer conformations given a.a. sequence \( r \)
- energy given conformation \( c \)
- universal gas constant (for unit conversion between kJ and K)
- absolute temperature (Kelvin)

**Partition Function (Z) Normalizes the Likelihood of the Protein In A Particular Conformational State**

Lilien, Stevens, Anderson, Donald, 2004

(approximates Ka, a biological measure of affinity)
**K*MAP Task**

\[ K^* MAP = \max_R K^*(r) \]

ie. Find the sequence with the greatest \( K^* \sim Ka \)
Marginal MAP (MMAP)

\[ \text{MMAP}(\mathcal{M}, X_{\text{MAP}}) = \max_{X/X_{\text{MAP}}} \sum_{\alpha} \prod_{\alpha} f_{\alpha}(X_{\alpha}) \]

- State-of-the-art search and sampling algorithms

State-of-the-art Marginal MAP (MMAP) algorithms

- Learning Depth-First AND/OR Search [Marinescu, Dechter, Ihler, 2018]
- Stochastic Best-First AND/OR Search [Marinescu, Dechter, Ihler, 2018]
- Recursive Best-First AND/OR Search [Marinescu, Dechter, Ihler, Kishimoto, Botea, 2018]

State-of-the-art sampling algorithms

- Dynamic Importance Sampling [Liu, Dechter, Ihler, 2017]
- Abstraction Sampling [Kask, Pezeshki, Broka, Ihler, Dechter, 2020]

\[ f(x^*) = \max_{\alpha} \prod_{\alpha} f_{\alpha}(x_{\alpha}) \]

\[ Z = \sum_{\alpha} \prod_{\alpha} f_{\alpha}(x_{\alpha}) \]

\[ f(x^*_M) = \max_{x_M} \sum_{x_S} \prod_{\alpha} f_{\alpha}(x_{\alpha}) \]

- NP-hard: exponentially many terms

MMAP < K*MAP

13
Graphical Models for K*MAP Task
Problem Formulation: Simplifications

- **Select Residues**: Model using only a subset of the residues.
- **Discrete Rotamers**: Use discrete side-chain conformations.
- **Fixed Backbone**: Fix the position of the residues in space.
Problem Formulation 1:

R’s capture amino acid assignment for residue

C’s index rotamer of amino acid assigned to corresponding R
Problem Formulation 2:

R’s capture amino acid assignment for residue

C’s capture all (amino acid, rotamer) combinations possible at its corresponding R

Constraints between corresponding C’s and R’s ensure consistent assignments
wMBE-K*

Based on wMBE-MMAP [Marinescu, Dechter, Ihler, 2014]
wMBE Heuristic for $K^*$

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

$M_B M_U^{-1}$

$M_B \quad \text{UB}(Z_B(R))$

$M_U \quad \text{LB}(Z_U(R))$

$$\sum_r w_r \leq \prod_r \left( \sum_x \psi_r(x) \right)$$

$$\sum_x f(x) \triangleq \left( \sum_x f(x) \frac{1}{w} \right) w \quad w = \sum_r w_r$$

MAX R

SUM Cγ
AOBB-K*

Based on AOBB-MMAP [Marinescu, Dechter, Ihler, 2014]
Algorithm 2: AOBB-K*

input : CPD graphical model $\mathcal{M}$; pseudo-tree $T$; $K^*$
upper-bounding heuristic function $h_{K^*}(\cdot)$; $Z_\gamma$
upper-bounding heuristic function $h_{Z\gamma}(\cdot)$; and
subunit stability threshold $S_\gamma$ for each subunit $\gamma$

output : $K^* MAP(\mathcal{M})$

begin

1. $\pi \leftarrow$ root OR node $s$
2. $ub_{K^*}(s) \leftarrow h_{K^*}(s)$
3. $lb_{K^*}(s) \leftarrow -\infty$
4. $g(s) \leftarrow 1$
5. foreach $\gamma \in \varphi$ do
   6. $UB_{Z\gamma}(s) \leftarrow \prod_{m \in ch_{T\gamma}(s)} h_{Z\gamma}(m)$

while $n_X \leftarrow EXPAND(\pi)$ do

9. if $\text{ConstraintPropagation}(\pi) = \text{false}$ then
   10. $\text{PRUNE}(\pi)$
   else if $\exists \gamma \in \varphi \text{ s.t. } UB_{Z\gamma}(n_X) < S_\gamma$ then
       11. $\text{PRUNE}(\pi)$
   else if $X \in R$ then
       12. if $\exists a \in \text{anc}^{OR}(n) \text{ s.t. } ub_{K^*}(a, \pi) < lb_{K^*}(a)$ then
           13. $\text{PRUNE}(\pi)$
       else if $ch_{T}^{unexp}(n) = \emptyset$ then
           14. $\text{BACKTRACK}(\pi)$
   return $ub_{K^*}(s) = lb_{K^*}(s) = K^* MAP(\mathcal{M})$
AOBB-K* High Level Overview

- Exact branch-and-bound algorithm over AND/OR search spaces
- Can use the statically compiled wMBE-K* heuristic
- Exploits determinism by using constraint propagation
- Incorporates a global constraint enforcing biologically relevant solutions

[Ojewole, Jou, Fowler, Donald, 2018]
Empirical Evaluation
### Results vs. State-of-the-art BBK*

| Problem   | iB | X | Dmax | w* | d | UB  | OR  | AND | CPP | UBP | SSP | time | *MAP | BBK* t | BBK* t sln |
|-----------|----|---|------|----|---|-----|-----|-----|-----|-----|-----|------|------|-------|------------|------------|
| 1gwc_00021 | 4  | 12 | 203  | 4  | 6 | 10.29 | 28766 | 134930 | 77823 | 55  | 2  | 16   | 9.79 | 152    | 9.79       |
| 2hnu_00025 | 4  | 14 | 203  | 5  | 7 | 15.08 | 22010 | 105458 | 76657 | 38  | 0  | 7    | 13.18 | 437    | 13.18      |
| 2hnu_00025 | 4  | 16 | 203  | 6  | 8 | 15.04 | 115194 | 297138 | 84882 | 39  | 0  | 16   | 13.65 | 962    | 13.65      |
| 2r9_00018  | 6  | 18 | 205  | 7  | 9 | 16.68 | 20137 | 87306 | 78   | 0  | 15  | 15.79 | 187    | 15.79      |
| 2rfd_00035 | 6  | 16 | 205  | 6  | 8 | 17.70 | 896239 | 4253159 | 3273123 | 40  | 0  | 15   | 16.50 | 182    | 16.50      |
| 2rfe_00030 | 6  | 14 | 203  | 5  | 7 | 11.53 | 20933 | 164126 | 359007 | 87  | 40  | 15   | 10.50 | 182    | 10.50      |
| 2rfe_00043 | 6  | 16 | 203  | 6  | 8 | 18.48 | 15390 | 40297 | 422357 | 34  | 43  | 80   | 18.04 | 50     | 18.04      |
| 2rfe_00044 | 6  | 16 | 203  | 6  | 8 | 18.62 | 37887 | 99927 | 1047107 | 30  | 3  | 86   | 18.10 | 75     | 18.10      |
| 2r10_00008 | 4  | 10 | 203  | 3  | 5 | 11.16 | 2   | 3   | 0   | 40  | 0  | 3    | 11.16 | 9.46   | 11.16      |
| 2xgy_00020 | 4  | 14 | 203  | 5  | 7 | 11.47 | 43643 | 262523 | 743880 | 40  | 0  | 14   | 16.80 | 888    | 16.80      |
| 3cal_00032 | 6  | 16 | 203  | 6  | 8 | 13.38 | 133851 | 1067419 | 531976 | 32  | 6  | 125  | 11.62 | 1429   | 11.62      |
| 3u7y_00009 | 5  | 12 | 203  | 4  | 6 | 4.51  | 2    | 3   | 0   | 40  | 0  | 6    | 4.51  | 191    | 4.51       |
| 4kt6_00023 | 4  | 16 | 203  | 6  | 8 | 14.80 | 38186 | 101546 | 23877 | 16  | 19  | 7    | 12.69 | 136    | 12.69      |
| 4wwi_00019 | 5  | 14 | 203  | 5  | 7 | 15.43 | 8094 | 30874 | 177864 | 40  | 0  | 12   | 14.98 | 37     | 14.98      |
| 1gwc_00021 | 4  | 13 | 203  | 4  | 7 | 12.51 | 33884 | 506021 | 473189 | 32  | 6  | 11.92 | 16.18 | 13.65  | 11.92      |
| 2hnu_00025 | 4  | 17 | 203  | 6  | 8 | 18.48 | 21517 | 596559 | 220825 | 77  | 0  | 12   | 16.18 | 13.65  | 16.18      |
| 2rfe_00012 | 5  | 15 | 205  | 5  | 8 | 14.36 | 3127  | 10003 | 32610 | 57  | 0  | 85   | 13.92 | 15.33  | 13.92      |
| 2rfe_00014 | 5  | 15 | 205  | 5  | 8 | 14.79 | 4087  | 13087 | 39411 | 57  | 0  | 85   | 14.36 | 14.36  | 14.36      |
| 2rfe_00017 | 5  | 15 | 205  | 5  | 8 | 11.46 | 245894 | 1063198 | 6389737 | 227 | 25  | 333  | 10.86 | 10.80  | 10.80      |
| 2rfe_00030 | 4  | 15 | 203  | 5  | 8 | 13.61 | 256957 | 1327425 | 2816050 | 726 | 83  | 726  | 11.12 | 10.90  | 11.12      |
| 2xgy_00020 | 5  | 15 | 203  | 5  | 8 | 11.39 | 398102 | 2383318 | 7422285 | 42  | 0  | 360  | 14.98 | 14.98  | 14.98      |
| 3u7y_00011 | 4  | 13 | 203  | 4  | 7 | 12.29 | 5758  | 16108 | 68579 | 50  | 0  | 99   | 4.51  | 216    | 4.51       |
| 4wwi_00019 | 4  | 15 | 203  | 5  | 8 | 16.05 | 22945 | 87485 | 91677 | 176 | 7  | 180  | 14.99 | 34     | 14.99      |

*Ojewole, Jou, Fowler, Donald, 2018*
Future Work

- Design new, more compact, problem representations
- Explore new heuristic functions and use of a dynamic heuristic
- Extend to search to approximate anytime methods and n-best solutions
- Extend to more complex formulations
Thank You!
wMBE Heuristic for MMAP

- **Mini-bucket elimination** [Dechter & Rish 2001]
  - “i-bound”, limit on the number of variables in a single mini-bucket

- **Weighted Mini-bucket** [Liu & Ihler, 2012]
  - Holder’s inequality

\[
\sum_{r} \prod \psi_r \leq \prod_{r} \left( \sum_{x} \right)
\]

\[
\sum_{x} f(x) \triangleq \left[ \sum_{x} f(x) \frac{1}{w} \right] w \quad w = \sum_{r} w_r
\]

\[
\sum_{E} [\psi(A, E) \psi(C, E)] \leq [\sum_{E} \psi(A, E)] [\sum_{E} \psi(C, E)]
\]
Problem Formulation: Subunit-Stability Constraints

Do not want dissociate subunits to be too unstable

\[ Z_{\text{subunit } i}(r) > Z_{\text{subunit } i}(r^{\text{wt}}) \times \exp\left\{-\frac{5}{RT}\right\} \]

- \( i \) = index of dissociate subunit
- \( r \) = amino acid sequence assignments
- \( D \) = indicating dissociate subunit
- \( r^{\text{wt}} \) = naturally occurring in nature amino acid sequence (wild type)
- \( R \) = universal gas constant (for unit conversion between kJ and K)
- \( T \) = absolute temperature (Kelvin)
Problem Formulation: Pseudo Tree Overview for K*MAP

More information:

Key Takeaway:
Can take advantage of decomposition
GMEC Objective

Lower Energy ➔ More Stable ➔ Structure More Likely To Exist

Def. Global Minimum-Energy Conformation (GMEC):
- conformation that minimizes the energy of the complex

\[ GMEC(r) = \min_{c \in C(r)} E(c) \]

- \( r = \) amino acid assignments to the residues
- \( C(r) = \) possible rotamer conformations given a.a. sequence \( r \)
- \( E(c) = \) energy given conformation \( c \)
GMEC MAP Task

\[ GMEC\ MAP = \min_{R} GMEC(r) \]

\( M = \text{minimum} \)

ie. Find the sequence with the lowest GMEC
- ie. Find sequence that has the most stable conformation
Proteins are Dynamic Structures

A protein’s structural state is dynamic

Proteins continuously transition between various energetically favorable conformation.

Not captured by the GMEC objective.
**K*MAP**

\[
Z_X(r) = \sum_{C_Y} \prod_{E_Y} e^{-\frac{E_Y(i,j)(r_i,C_Y(i),r_j,C_Y(j))}{RT}}
\]

\[
K^*(r) = \frac{Z_{Bound}(r)}{Z_{Dissociate}(r)} = \frac{Z_{PL}(r)}{Z_P(r)Z_L(r)}
\]

\[
K^*MAP = \max_R K^*(r)
\]

\[X \in \{Bound, Dissociate\}\]