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



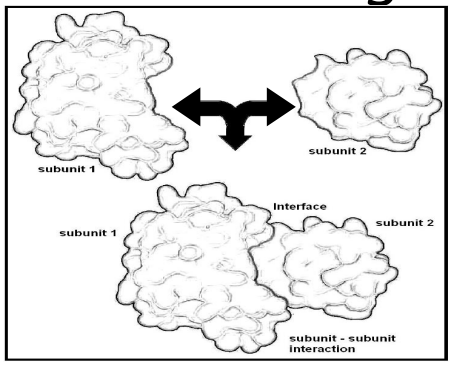

BOBAK PEZESHKI, RADU MARINESCU, ALEX IHLER, RINA DECHTER

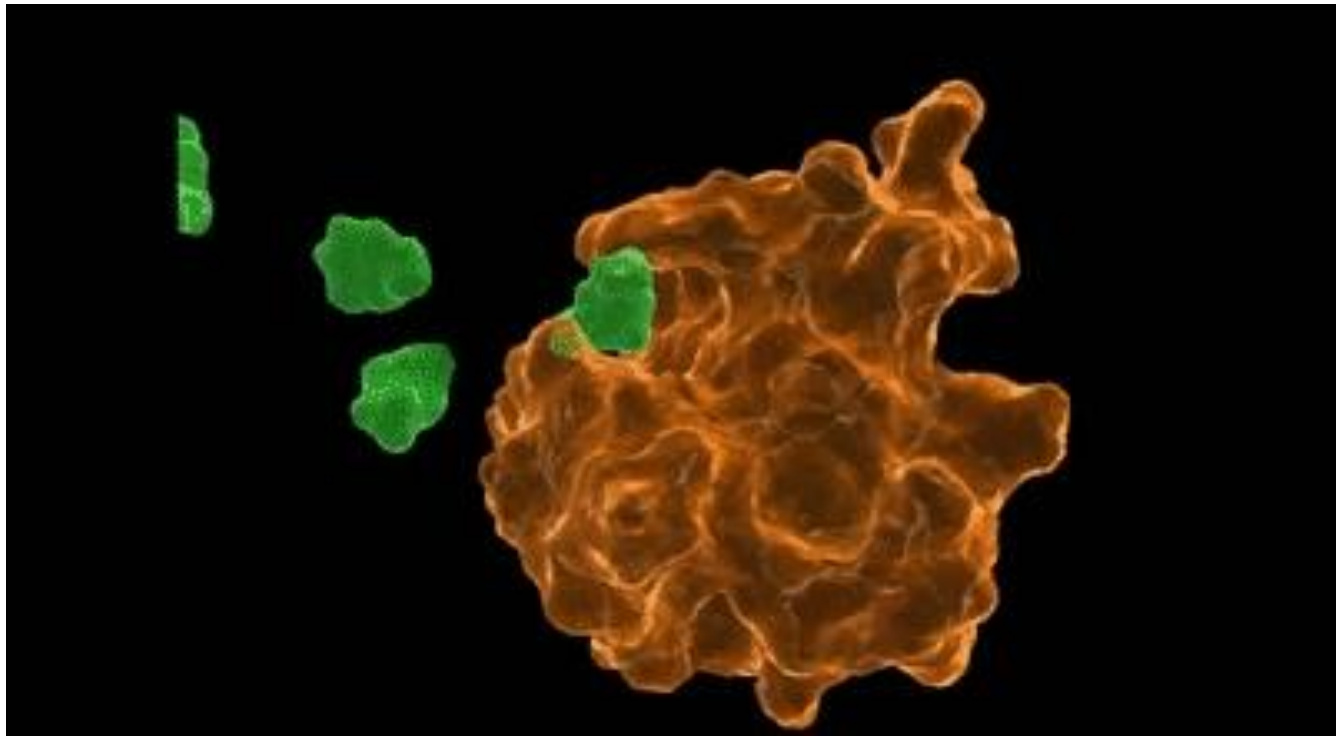
Computational Protein Design (CPD)

[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.



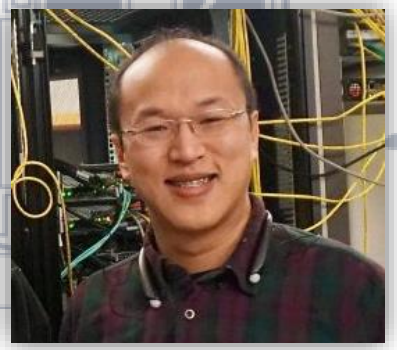


https://www.mrdubuque.com/uploads/2/4/5/0/24509062/u4sp7h_orig.gif

Special thanks to…

THOMAS SCHIEX

Special thanks to…



JONATHAN JOU

GRAHAM HOLT

BRUCE DONALD

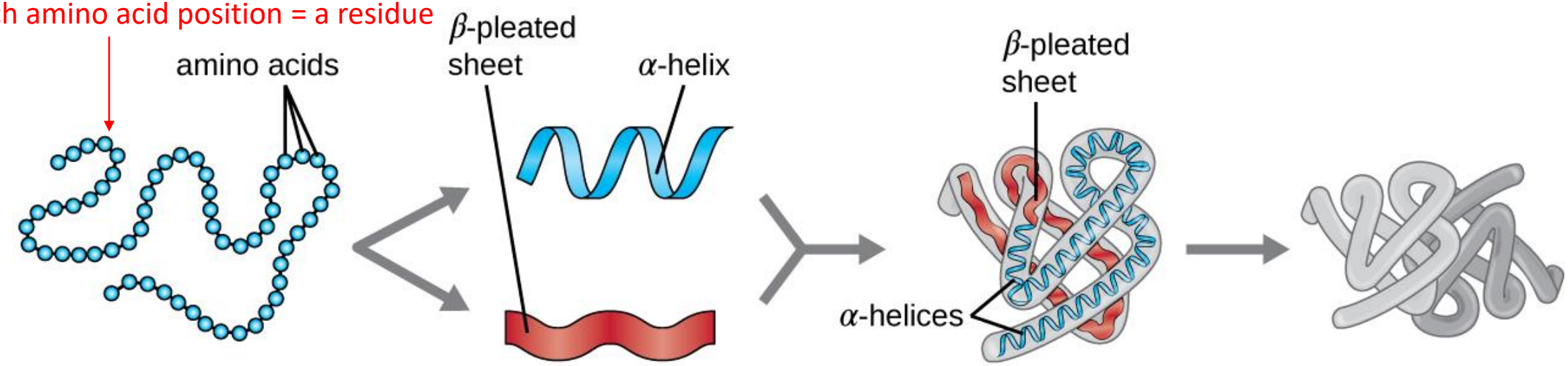
Key Contributions

- **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

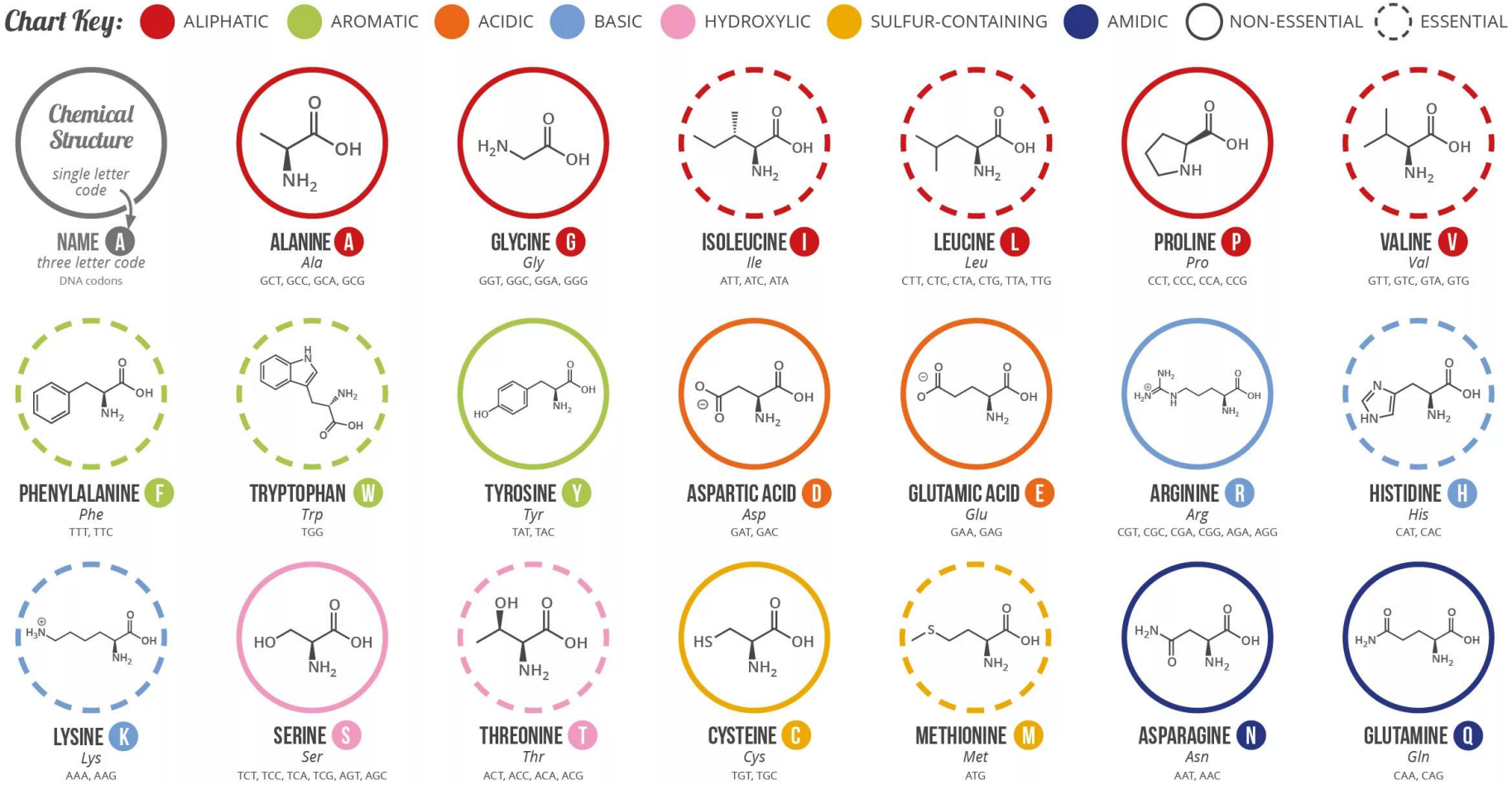
Tertiary Protein Structure
three-dimensional folding pattern of a protein due to side chain interactions

Quaternary Protein Structure
protein consisting of more than one amino acid chain

Review of Proteins

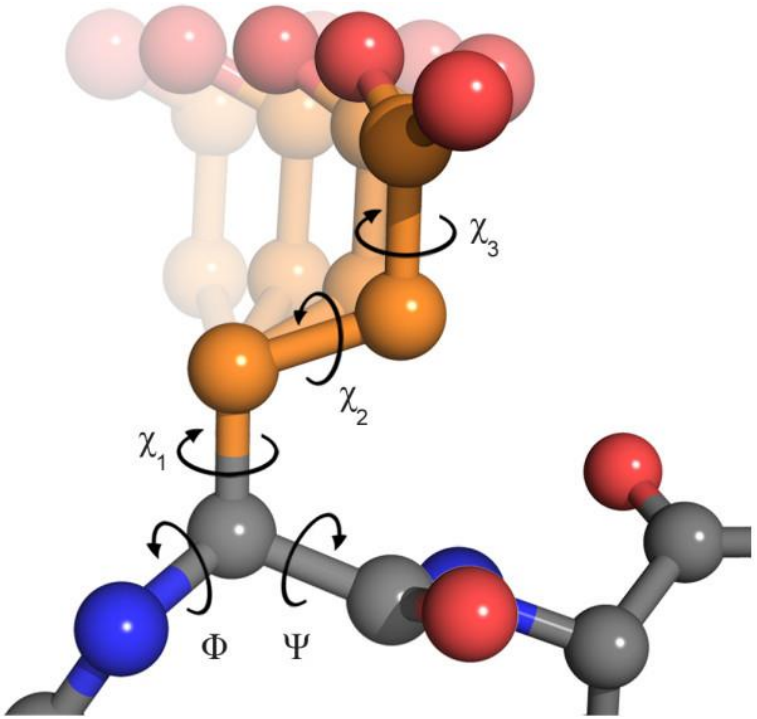
Cannot be made by the human body

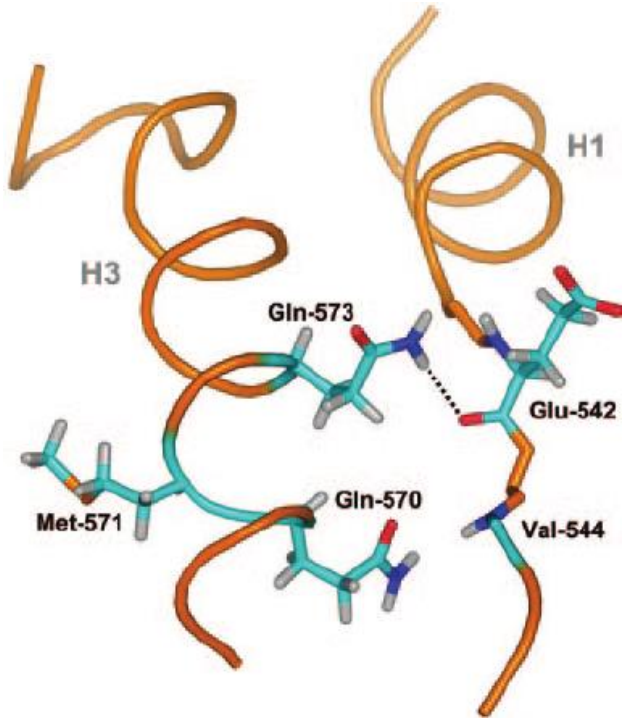
There Are ~20 Naturally Occurring Amino Acids



Review of Proteins

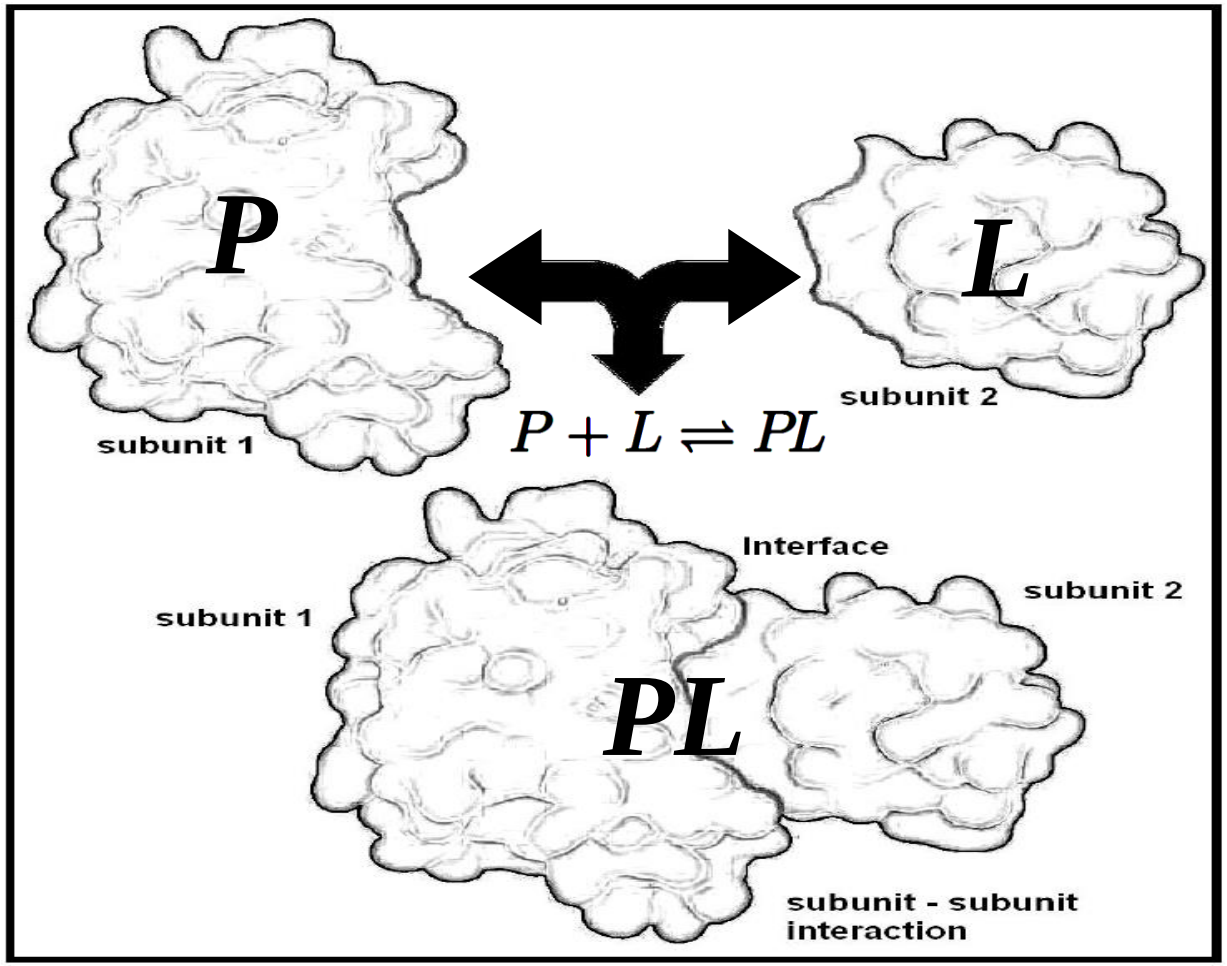
Amino Acid Rotamers: Select conformational isomers of an amino acid





Peter Carlsson, Konrad F. Koehler, and Lennart Nilsson
Molecular Endocrinology 19(8):1960–1977. https://doi.org/10.1210/me.2004-0203

Proteins are Dynamic Structures



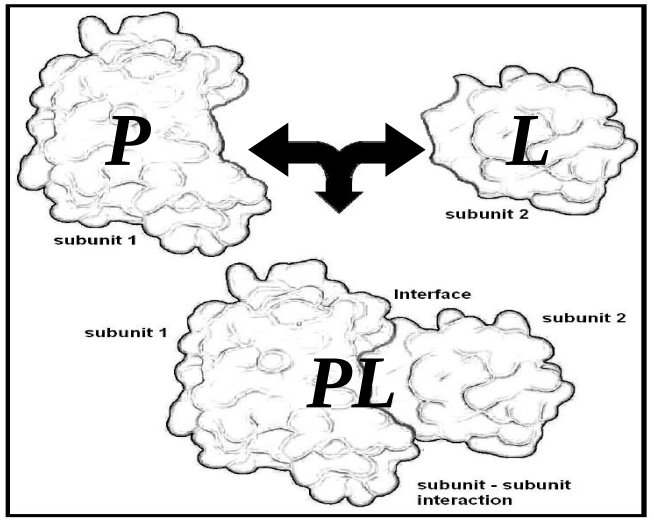
Sowmya, Gopichandran & Vaishnavi, A. & Jigisha, A. & Kangueane, Pandjassarame. (2011). Protein-protein complexes.

K* Objective

(approximates Ka, a biological measure of affinity)

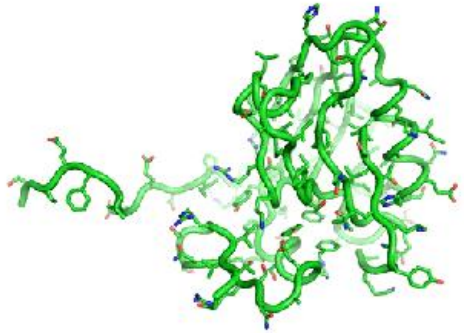
$$K^*(r) = \frac{Z_{PL}(r)}{Z_P(r)\,Z_L(r)}$$

[Lilien, Stevens, Anderson, Donald, 2004]



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

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

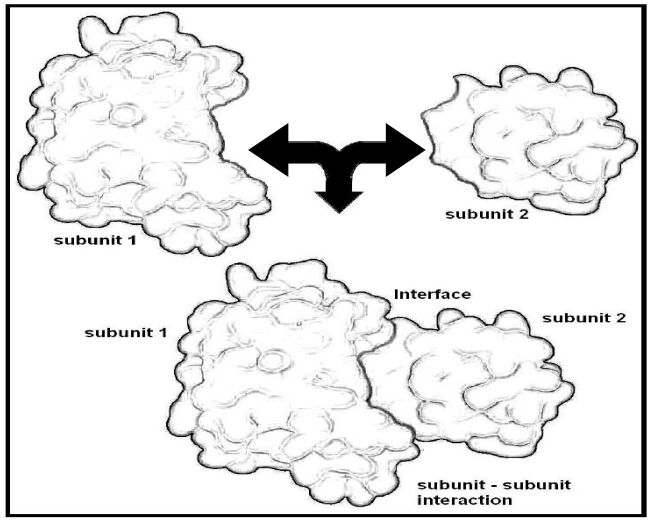


mino acid assignments to the residues
= possible rotamer conformations given a.a. sequence r
= energy given conformation c
universal gas constant (for unit conversion between kJ and K)
absolute temperature (Kelvin)

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

K*MAP Task

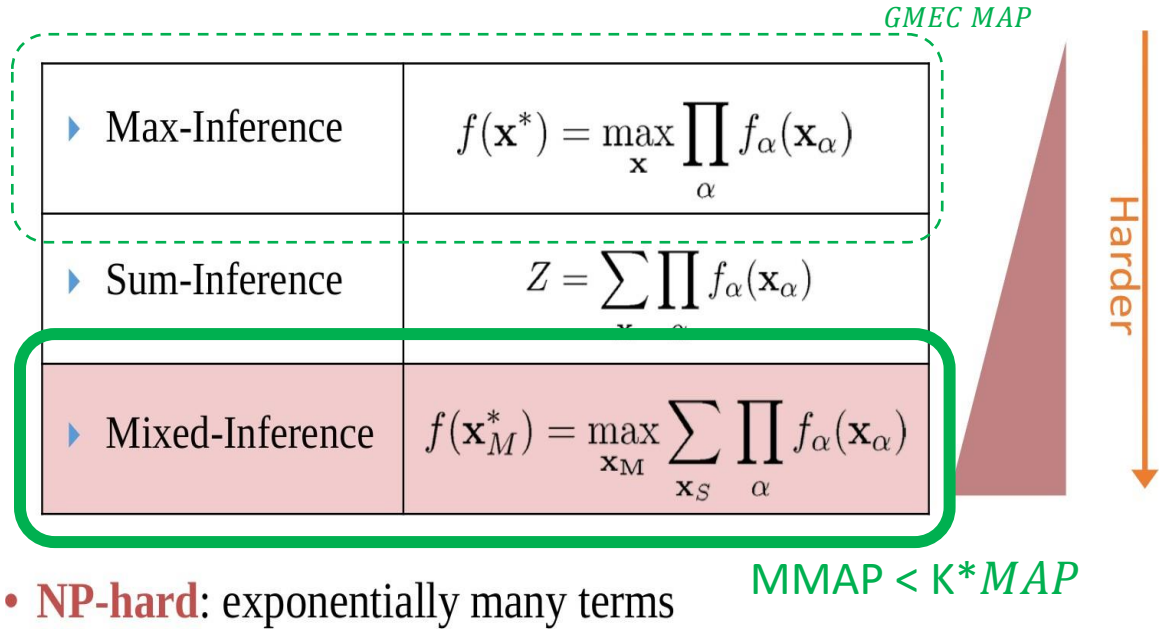
$$K^*MAP = \max_R K^*(r)$$



ie. Find the sequence with the greatest K* ~ Ka

Marginal MAP (MMAP)

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



- **NP-hard**: exponentially many terms

- 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]

[Marinescu, Lee, Dechter, Ihler, 2018]

State-of-the-art sampling algorithms

Dynamic Importance Sampling [Liu, Dechter, Ihler, 2017]

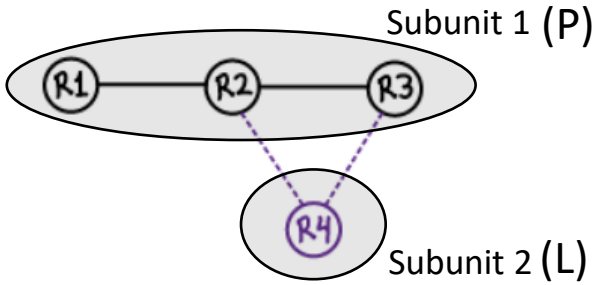
Abstraction Sampling [Kask, Pezeshki, Broka, Ihler, Dechter, 2020]

Graphical Models for K*MAP Task

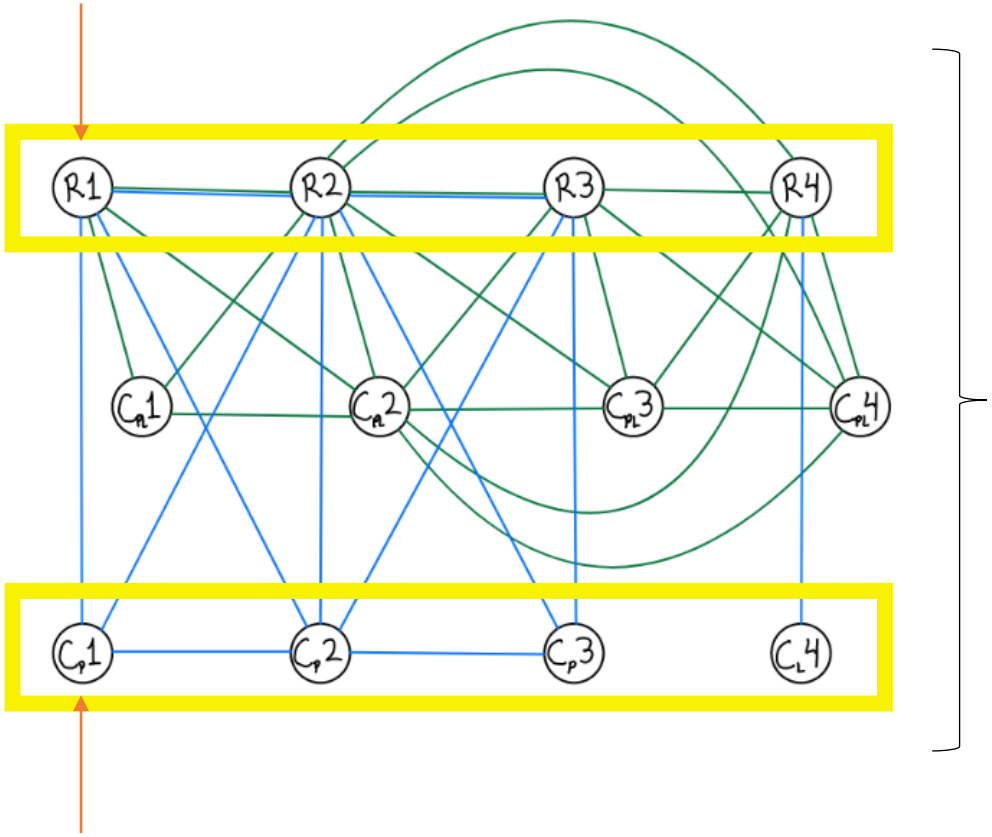
Problem Formulation: Simplificaitons

- **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.

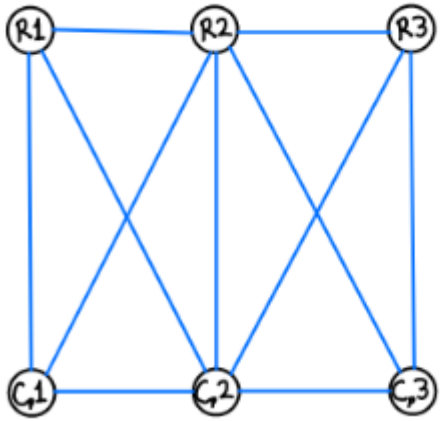
Problem Formulation 1:



R's capture amino acid assignment for residue

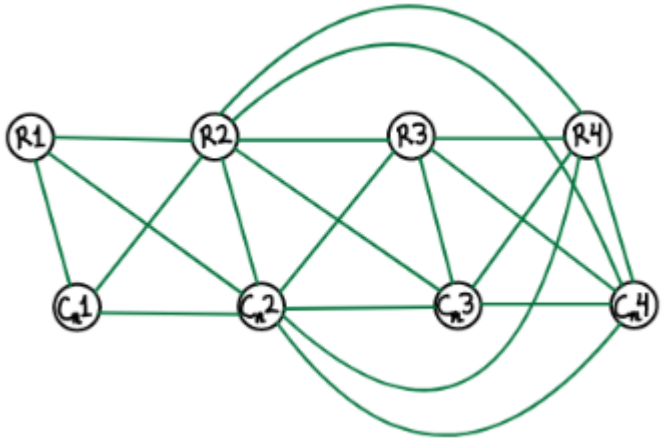


P



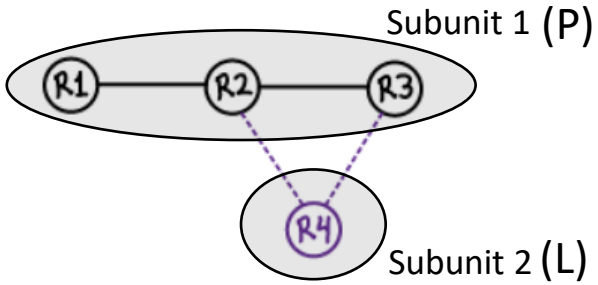
L

PL

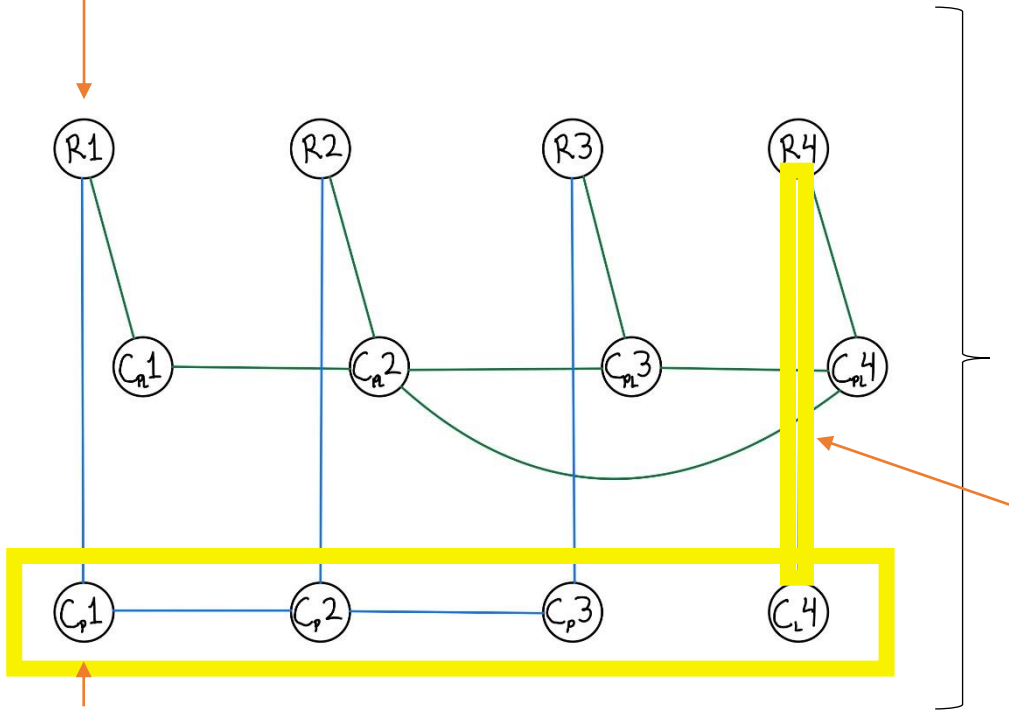


C's index rotamer of amino acid assigned to corresponding R

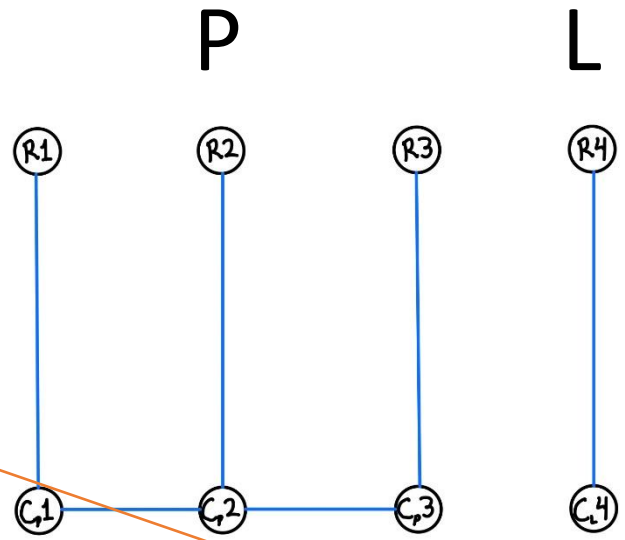
Problem Formulation 2:

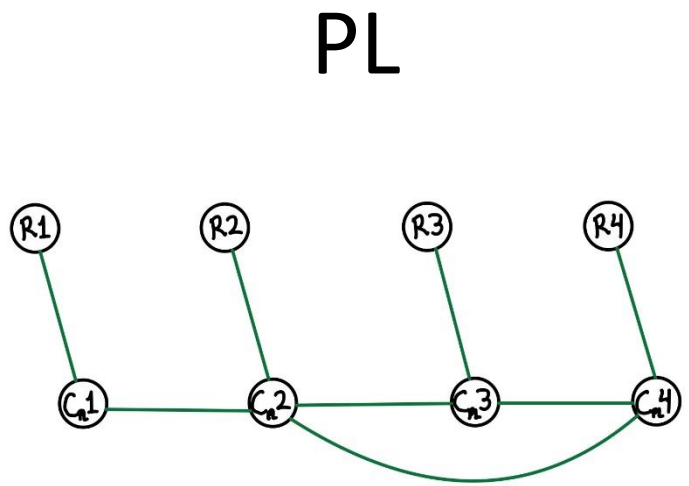


R's capture amino acid assignment for residue



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





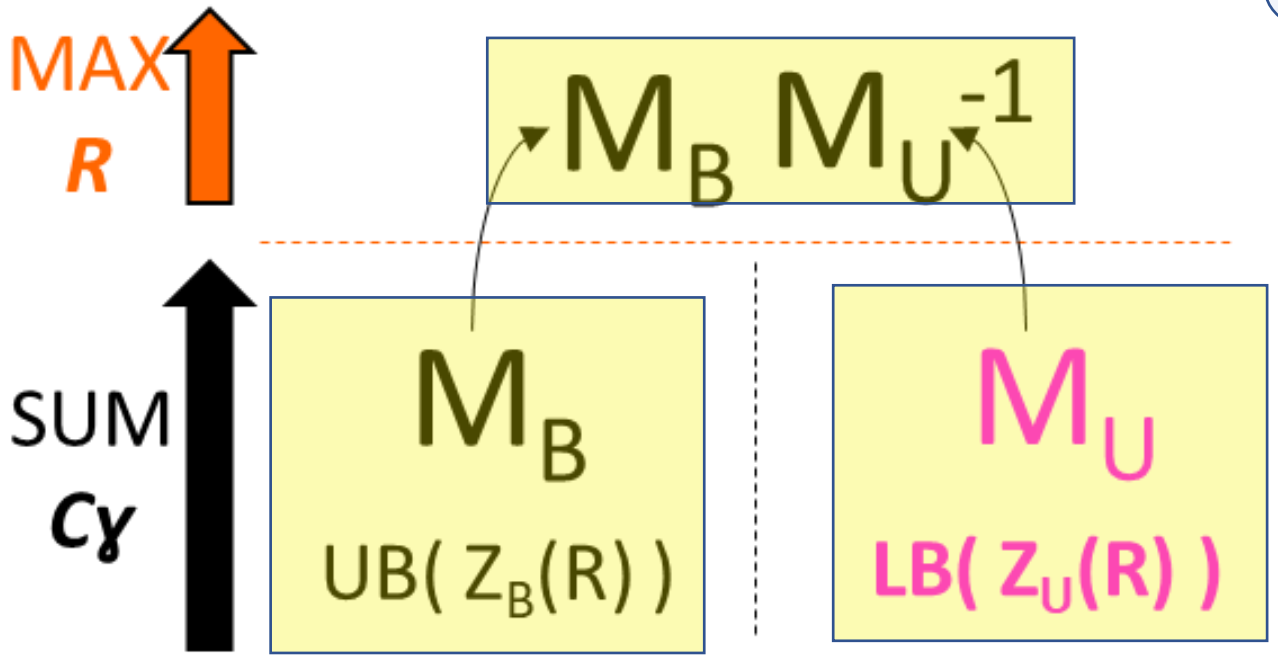
Constraints between corresponding **C**'s and **R**'s ensure consistent assignments

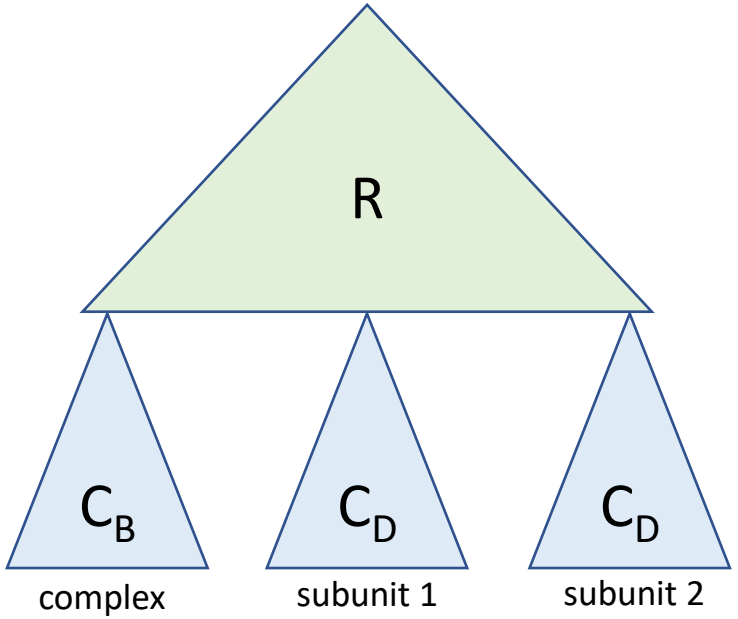
wMBE-K*

Based on wMBE-MMAP [Marinescu, Dechter, Ihler, 2014]

wMBE Heuristic for K*

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





$$\sum^{w}[\prod_{r}(\psi_r)] \leq \prod_{r}[\sum_{x}^{w_r}]$$

$$\sum_{x}^{w} f(x) \triangleq [\sum_{x} f(x)^{\frac{1}{w}}]^{w} \qquad w = \sum_{r} w_r$$

AOBB-K*

Based on AOBB-MMAP [Marinescu, Dechter, Ihler, 2014]

AOBB-K* Algorithm

Algorithm 2: AOBB-K*

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

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

1 **begin**
2 Encode deterministic relations in $\mathcal{M}$ into CNF
3 $\pi \leftarrow$ root OR node s
4 $ub_{K^*}(s) \leftarrow h_{K^*}(s)$
5 $lb_{K^*}(s) \leftarrow -inf$
6 $g(s) \leftarrow 1$
7 **foreach** $\gamma \in \varphi$ **do**
8 $UB_{Z_\gamma}(s) \leftarrow \prod_{m \in ch_{T_\gamma}(s)} h_{Z_\gamma}(m)$
9 **while** $n_X \leftarrow EXPAND(\pi)$ **do**
10 **if** $ConstraintPropagation(\pi) = false$ **then**
11 $PRUNE(\pi)$
12 **else if** $\exists \gamma \in \varphi$ *s.t.* $UB_{Z_\gamma}(n_X) < S_\gamma$ **then**
13 $PRUNE(\pi)$
14 **else if** $X \in \boldsymbol{R}$ **then**
15 **if** $\exists a \in anc^{OR}(n)$ *s.t.* $ub_{K^*}(a, \pi) < lb_{K^*}(a)$ **then**
16 $PRUNE(\pi)$
17 **else if** $ch_T^{unexp}(n) = \emptyset$ **then**
18 $BACKTRACK(\pi)$
19 **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*

[Ojewole, Jou, Fowler, Donald, 2018]

Problem	iB	|X|	Dmax	w*	d	UB	OR	AND	CPP	UBP	SSP	E	time	*MAP	BBK* t	BBK* sln
1gwc_00021	4	12	203	4	6	10.29	28766	134930	77823	55	2	5	**16**	9.79	152	9.79
2hnu_00026	4	14	203	5	7	15.08	22010	105458	76657	38	0	4	**7**	13.18	437	13.18
2hnv_00025	4	16	203	6	8	15.04	115194	297138	84882	39	0	3	**16**	13.65	962	13.65
2rf9_00018	6	18	205	7	9	16.68	20137	85033	87306	78	0	4	**15**	15.79	187	15.79
2rfd_00035	6	16	205	6	8	17.70	896239	4253159	3273123	40	0	4	**381**	**17**	VS.	16.77
2rfe_00030	4	14	203	5	7	11.53	20393	164126	359007	87	40	7	**19**	10.50	182	10.50
2rfe_00043	6	16	203	6	8	18.48	15390	40297	422357	34	43	4	80	18.04	50	18.04
2rfe_00044	6	16	203	6	8	18.62	37887	99927	1047107	30	3	6	86	18.19	75	18.19
2rl0_00008	4	10	203	3	5	**11.16**	2	3	0	40	0	3	**3**	**11.16**	VS.	9.46
2xgy_00020	4	14	203	5	7	11.47	43643	262523	743860	40	0	2	**14**	10.60	887	10.60
3cal_00032	6	16	203	6	8	13.38	133851	1067419	531976	32	6	4	**125**	11.62	1429	11.62
3u7y_00009	5	12	203	4	6	**4.51**	2	3	0	40	0	3	**6**	4.51	191	4.51
4kt6_00023	4	16	203	6	8	14.80	38186	101546	23877	16	19	4	**7**	12.69	136	12.69
4wwi_00019	5	14	203	5	7	15.43	8094	30774	17888	40	0	2	**7**	14.99	26	14.99
1gwc_00021	4	13	203	4	7	12.51	33881	590621	473189	388	6	5	203	**11.92**	VS.	11.72
2hnv_00025	4	17	203	6	9	18.38	215171	550559	220825	77	0	4	**153**	**16.18**	VS.	13.65
2rfe_00012	5	15	205	5	8	14.36	3127	10003	32610	57	0	3	85	13.93	12	13.93
2rfe_00014	5	15	205	5	8	14.79	4087	13087	39411	57	0	3	85	14.36	45	14.36
2rfe_00017	5	15	203	5	8	11.46	245894	1063198	6389737	227	25	3	333	**10.86**	VS.	10.80
2rfe_00030	4	15	203	5	8	13.61	256957	1327425	2816050	726	83	7	**274**	**11.12**	VS.	10.97
2xgy_00020	5	15	203	5	8	11.39	398102	2383318	7422285	42	0	0	**360**	10.90	1388	10.90
3u7y_00009	4	13	203	4	7	4.95	36760	228568	564654	204	7	3	**99**	4.51	216	4.51
3u7y_00011	4	13	203	4	7	12.29	5758	16108	68579	50	0	5	86	11.85	27	11.85
4wwi_00019	5	15	203	5	8	16.05	22945	87485	91677	176	75	5	180	14.99	34	14.99

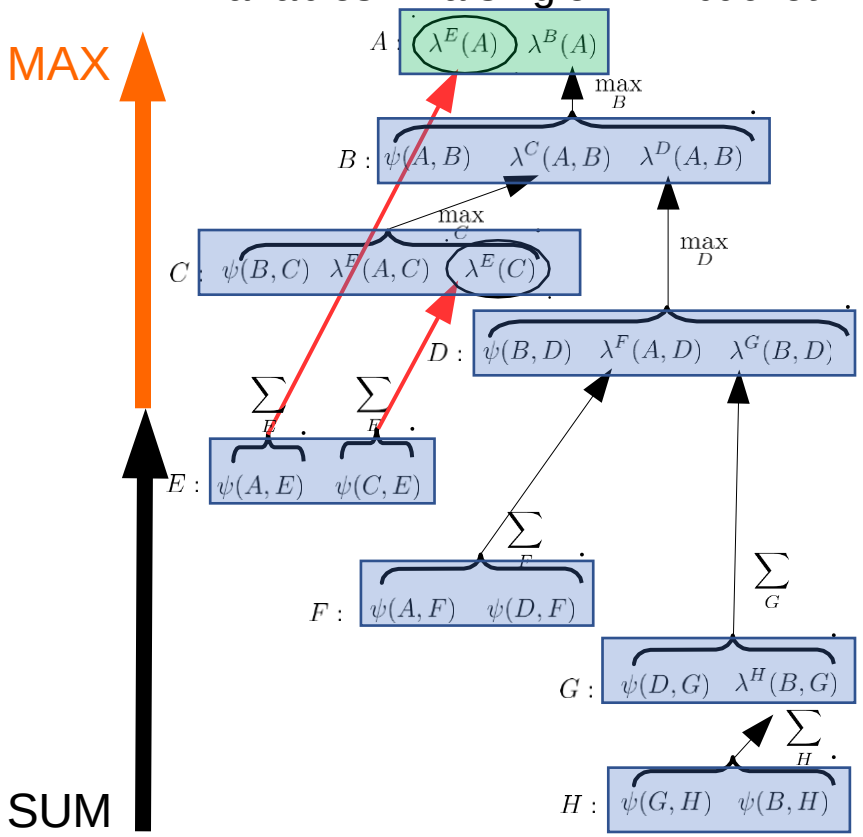
Future Work

- Design new, more compact, problem representations
- Explore new heuristic functions and use of a dynamic heuristc
- 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

$$\sum_x^w [\prod_r (\psi_r)] \le \prod_r^{w_r} [\sum_x]$$

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

$$\sum_E [\psi(A,E)\psi(C,E)] \le [\sum_E \psi(A,E)][\sum_E \psi(C,E)]$$

Problem Formulation: Subunit-Stability Constraints

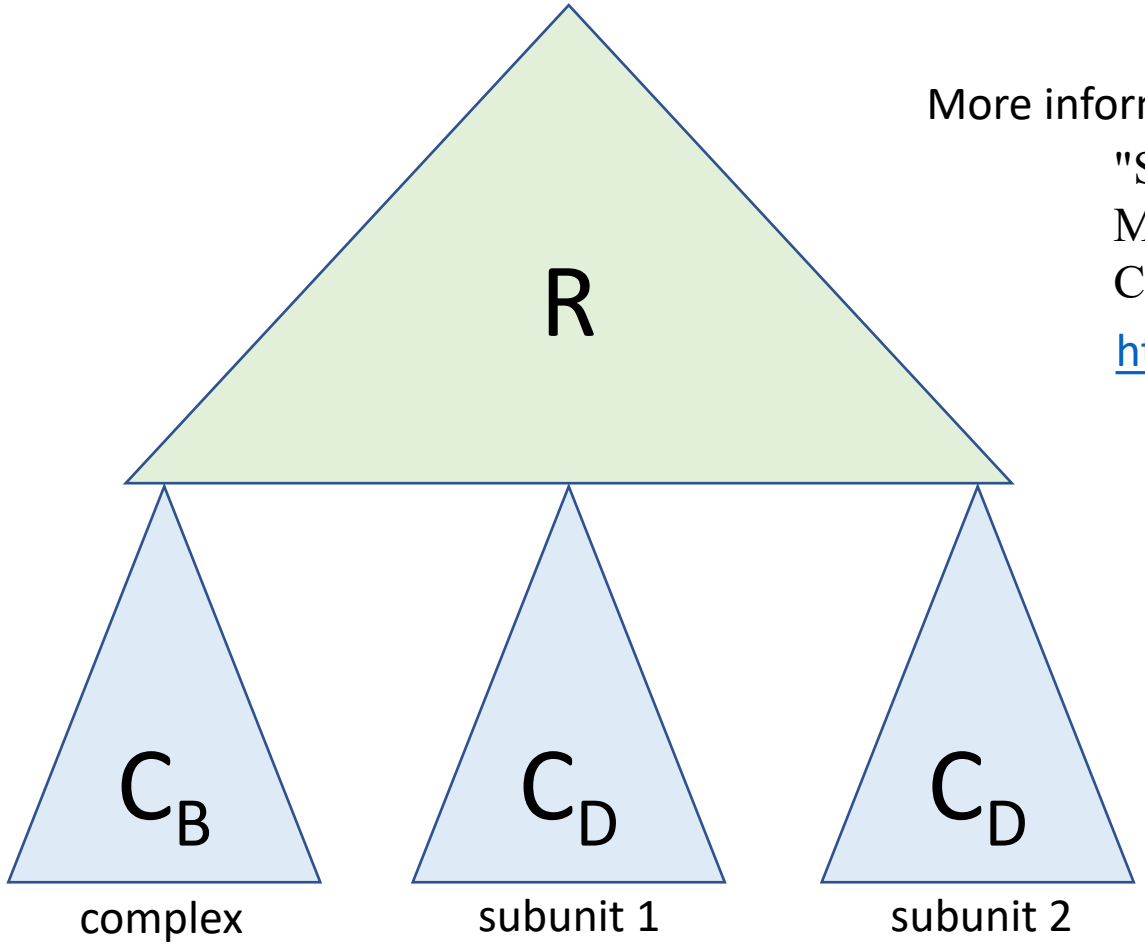
$$K^*(r) = \frac{Z_{complex}(r)}{Z_{subunit\ 1}(r)\, Z_{subunit\ 2}(r)}$$

Do not want dissociate subunits to be too unstable

$$Z_{subunit\ i}(r) > \underbrace{Z_{subunit\ i}(r^{wt})}_{\text{Likelihood of naturally occurring version}} * \underbrace{\exp\{-5/\mathcal{R}\boldsymbol{T}\}}_{\text{Constant factor to threshold with}}$$

i = index of dissociate subunit
r = amino acid sequence assignments
D = indicating dissociate subunit
r^{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:

"Search Algorithms for Solving Queries on Graphical Models and the Importance of Pseudo-trees in their Complexity" *UCI ICS Technical Report, June 2017.*

https://www.ics.uci.edu/~dechter/publications/r243.pdf

Key Takeaway:
Can take advantage of decomposition

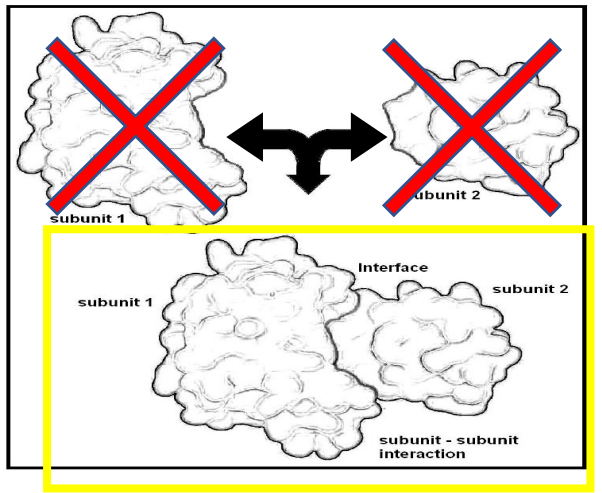
GMEC Objective

Lower Energy ➔ More Stable ➔ Structure More Likely To Exist

Def. Global Minimum-Energy Conformation (GMEC):

- conformation that minimizes the energy of the complex

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

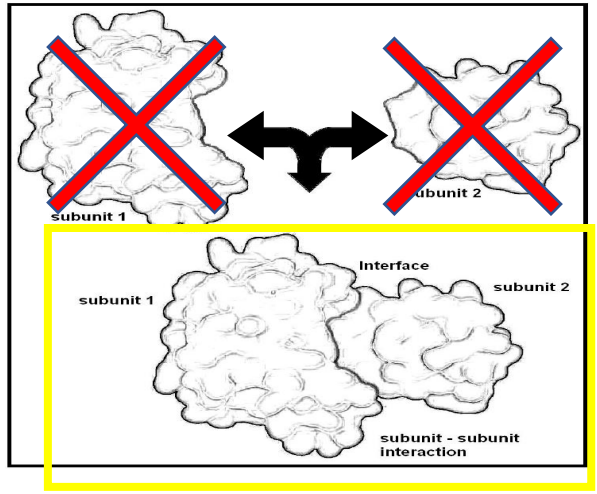


r = amino acid assignments to the residues
C(r) = possible rotamer conformations given a.a. sequence r
E(c) = energy given conformation c

GMEC MAP Task

M = minimum

$$GMEC\ MAP = \min_{R} GMEC(r)$$



ie. Find the sequence with the lowest GMEC

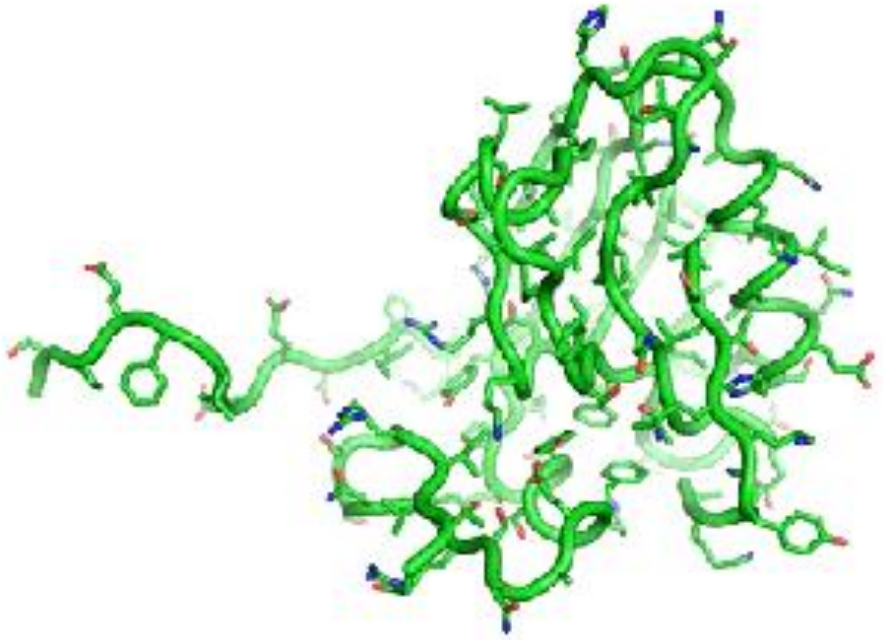
- ie. Find sequence that has the most stable conformation

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

$X \in \{Bound, Dissociate\}$

$$Z_X(\boldsymbol{r}) = \sum_{\boldsymbol{C_\gamma}} \prod_{\boldsymbol{E_\gamma}} e^{-\frac{E_{\gamma(i,j)}\left(r_i, C_{\gamma(i)}, r_j, C_{\gamma(j)}\right)}{\mathcal{R}T}}$$

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

$$K^*MAP = \max_{\boldsymbol{R}} K^*(\boldsymbol{r})$$