Approximate Counting with Deterministic Guarantees for Binding Affinity Computation

Clément Viricel[1,2], David Simoncini[1], David Allouche[1], Simon de Givry[1], Sophie Barbe[2] and Thomas Schiex[1]

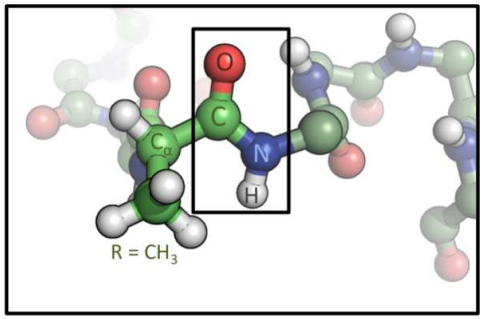
[1]*Unité de Mathématiques et Informatiques Appliquées UR 875, INRA, F-31320 Castanet Tolosan, France,*

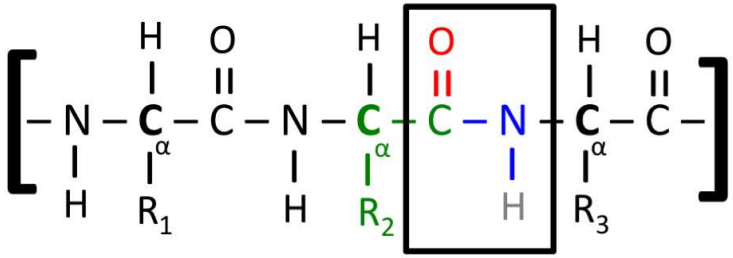
[2]*Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés, INSA, UMR INRA 792/CNRS 5504, F-31400 Toulouse, France*

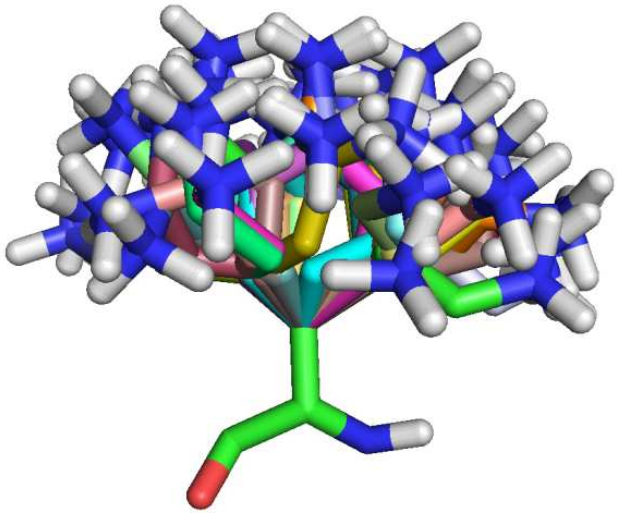
What is a protein ?

Protein: amino acids (AA) sequence

Protein: Backbone + side-chains







Side-chain have different conformations

From sequence to 3D structure

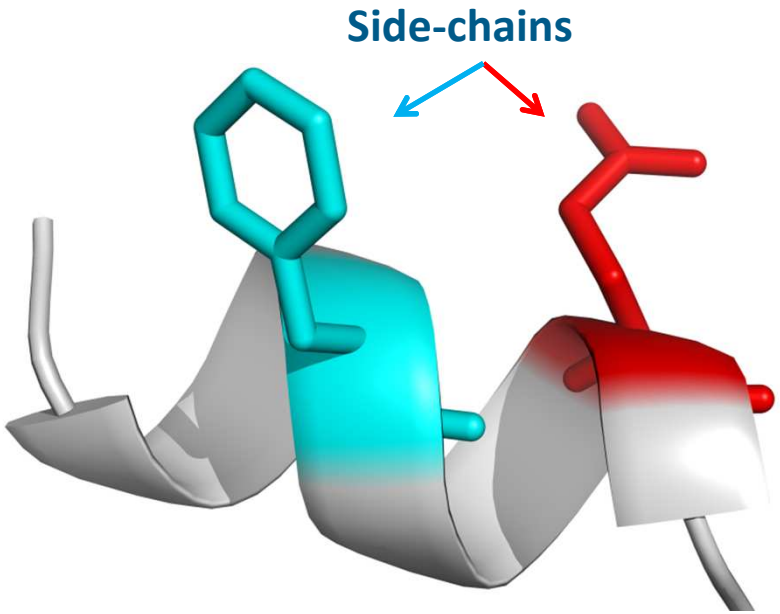
Amino Acid Sequence

---Trp Asp Pro Glu Phe Ala Ser Glu Gln Ser---

Folding

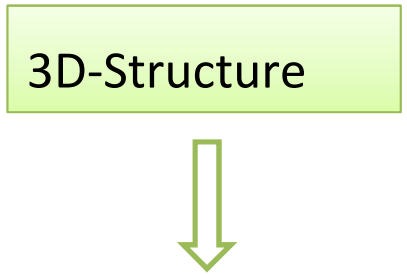
Defines

3 Dimensional Structure

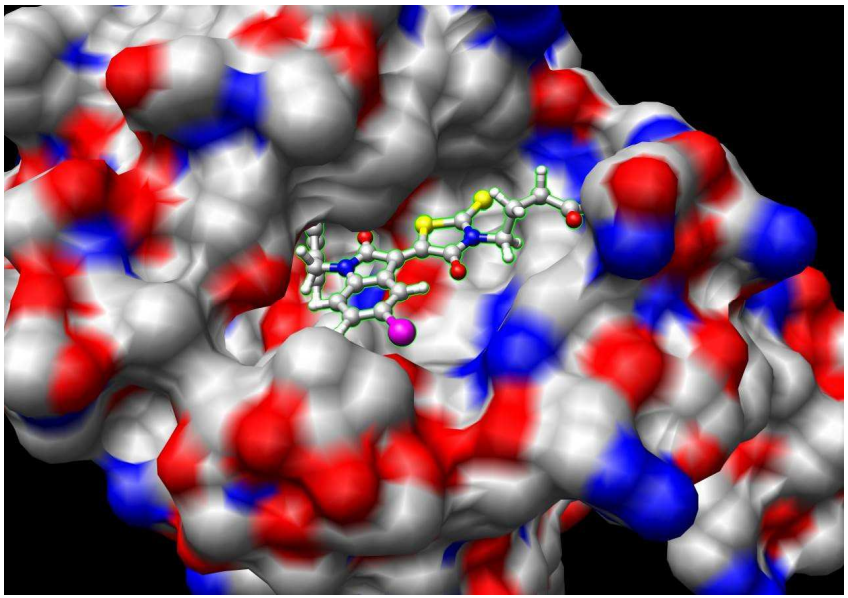


From 3D Structure to Function

3D-Structure



Function



- Protein Design Objective
  - Sequence ⟶ structure ⟶ function so new function requires new sequence
  - Identify sequences that adopt 3D structure with suitable function (enhances activity, **control recognition of partners**)

Protein Design

Issue : Combinatorial Explosion

For a n amino acids protein, 20 natural amino acid types
$\Rightarrow 20^n$ sequences
For a 50 amino acids protein : $20^{50} \approx 10^{65}$ sequences.
For 1 µg/prot $\Rightarrow \sim 10^{21}$ times the Earth's mass.

The Computational Protein Design (CPD)

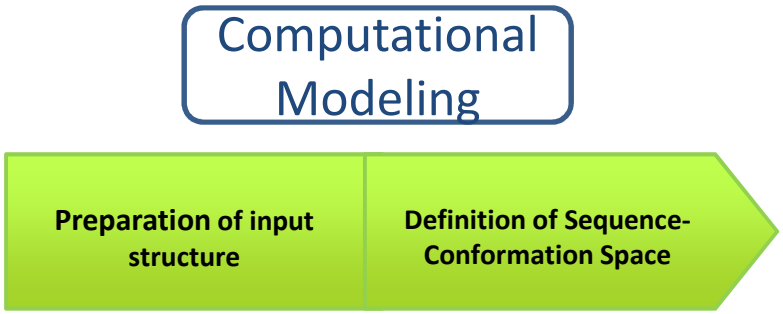
Goal:

- Increase the odd of finding hits
- Reduce cost and time development

How:

- Mathematical model of proteins
- Criteria and algorithms for finding suitable sequences.

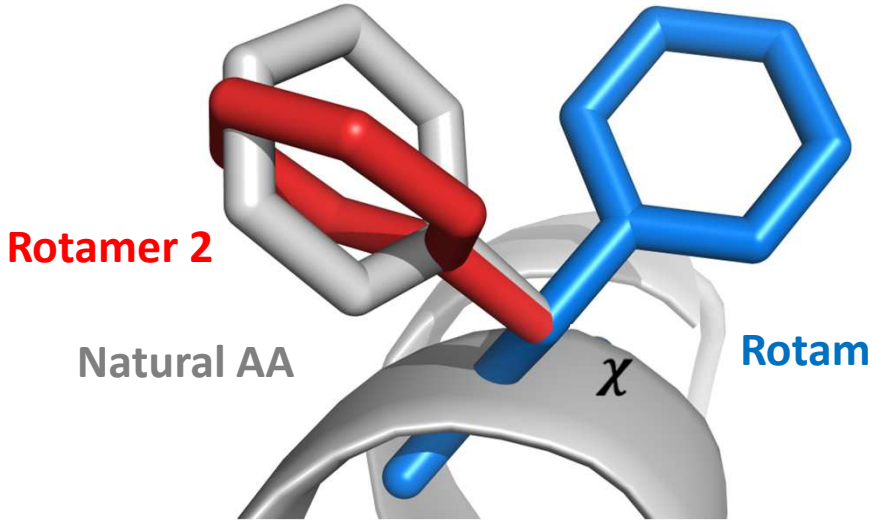
Computational Modeling

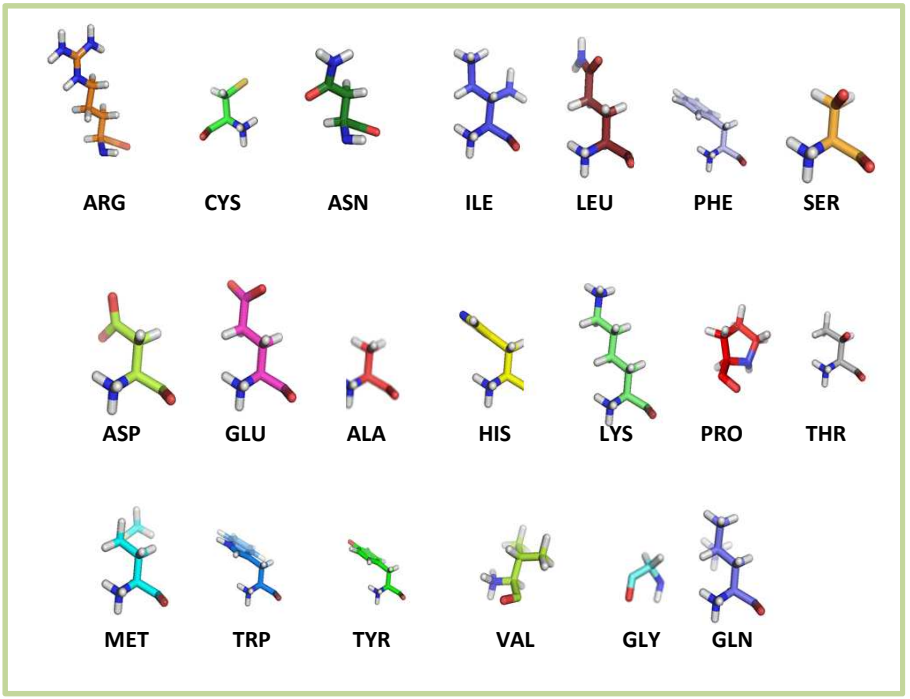


Modeling Protein Flexibility

Usual modeling assumptions:

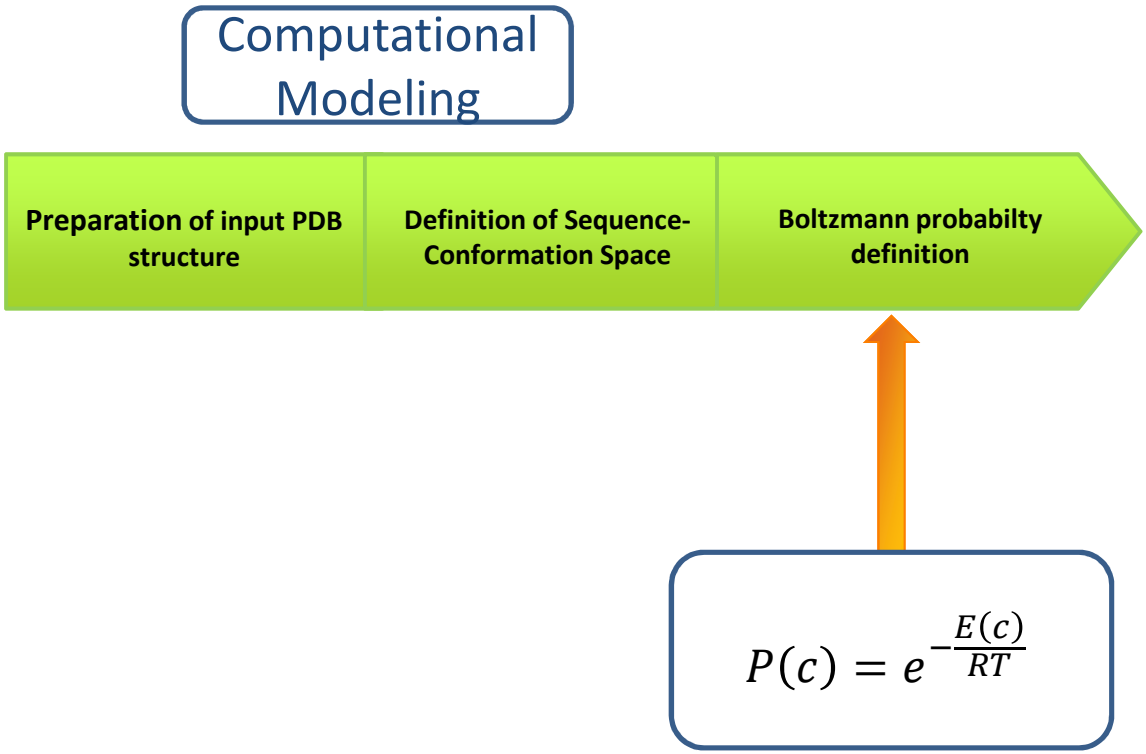
- Rigid backbone
- Discrete side-chain orientations (rotamers : most frequent conformations)





Search space = Sequence space x conformation space

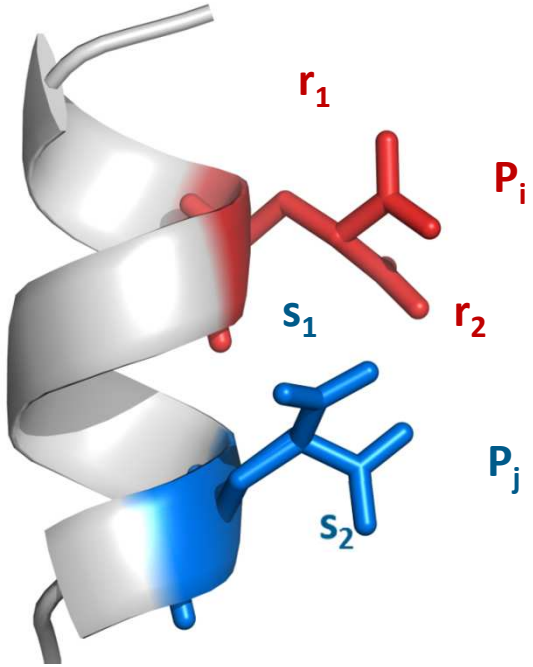
Binding Affinity Computation



All-Atom Energy Function

$$E_P = E_t + \sum_{i=0}^{n} E_{Unary}(i_r) + \sum_{i,j=0}^{n} E_{Binary}(i_r, j_s)$$

Interactions			
	Backbone Backbone	Backbone Rotamer	Rotamer Rotamer



Energy matrix

	P_i	r_1	r_2
P_j		0.1	-3.2
s_1	1.7	-3.5	3.4
s_2	-4.2	0	-2.1

Computing Z

- Limited Guarantees
  - Monte Carlo(sampling), mean field , message passing (TRW).
- Exact
  - Cachet, #SAT (SAT solver, caching)
- (δ, ε)-guarantees:
  - WISH(+optimisation), MIS: XOR hashing based
  - Gumbel perturbations (+optimization)
- ε-guarantees
  - OSPREY-K*

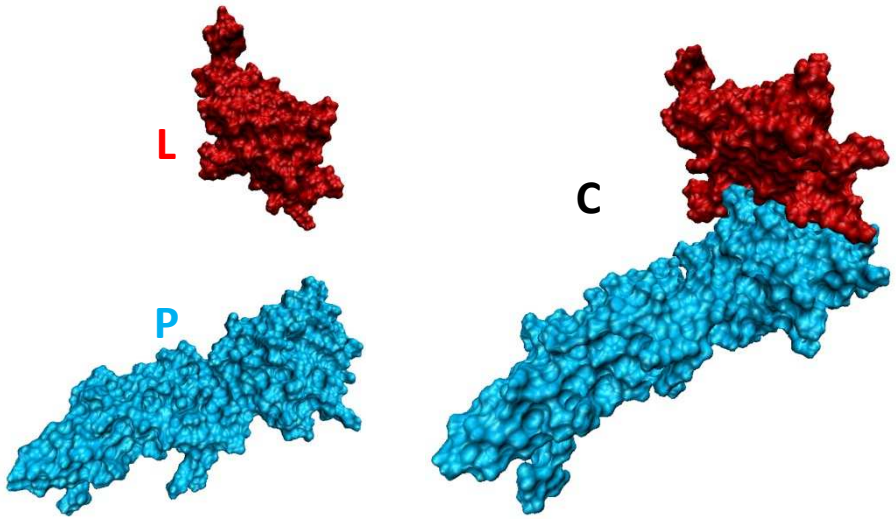
Binding Affinity Constant

The binding constant K_A represents the affinity between two proteins (for each sequence)

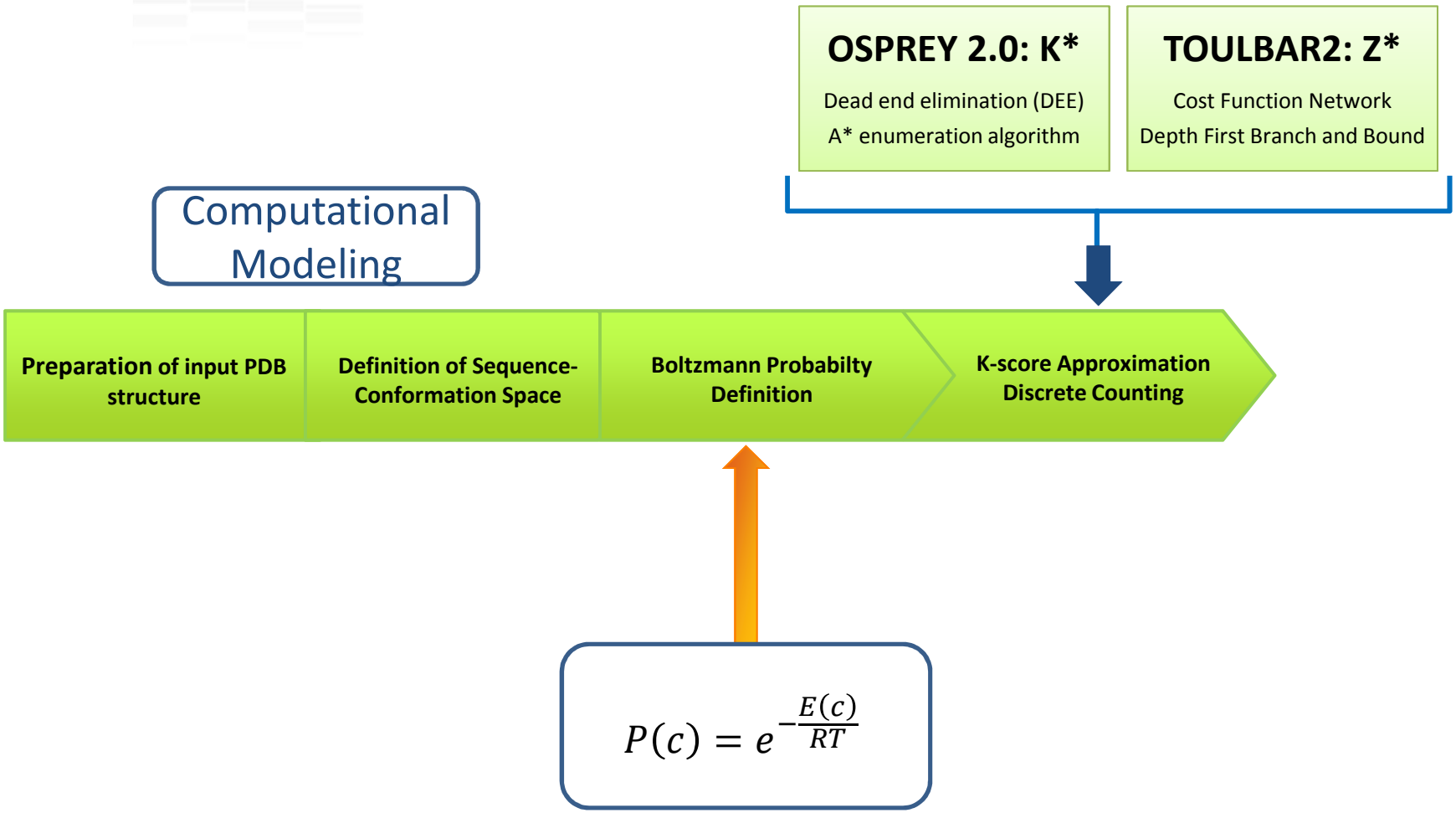
$$K_A \propto \frac{Z_C}{Z_P Z_L} \approx \frac{\sum_{c \in C} e^{-E_c/RT}}{\sum_{p \in P} e^{-E_p/RT} . \sum_{l \in L} e^{-E_l/RT}}$$

Z computation

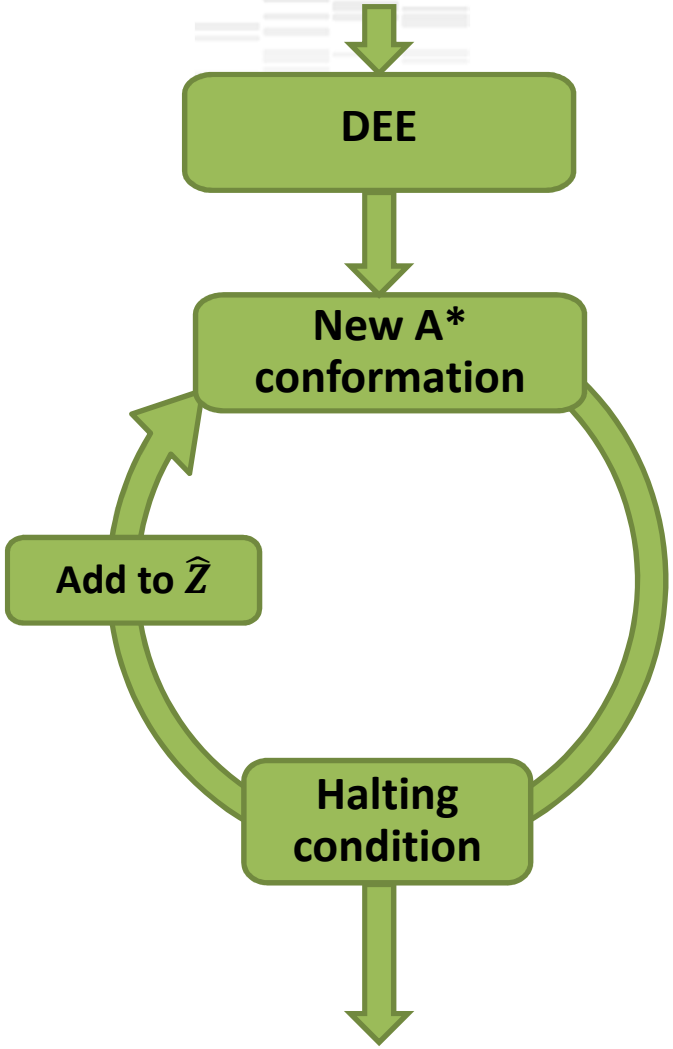
#P-complete



Discrete Counting



K* algorithm (OSPREY)



DEE removes strongly dominated rotamers

A* enumeration produces conformations in decreasing order of probability mass

A Novel Ensemble-Based Scoring and Search Algorithm for Protein Redesign, and its Application to Modify the Substrate Specificity of the Gramicidin Synthetase A Phenylalanine Adenylation Enzyme. [Ryan H. Lilien] RECOMB'04

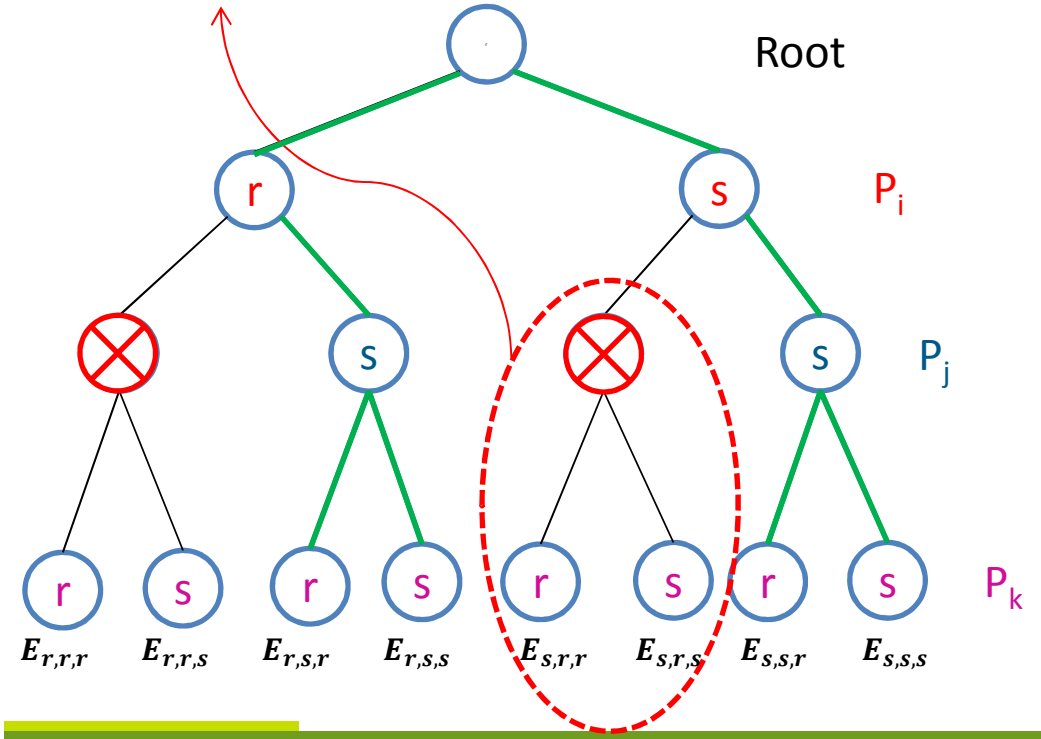
Tree Exploration

Enforce invariant: $U < \varepsilon\hat{Z}$

U : Amount of pruned probability mass $\hat{Z}$: Current Z approximation

$$\hat{Z} < Z < \hat{Z} + U \iff \hat{Z} < Z < (1+\epsilon)\hat{Z}$$

Upper U_b bound on $\sum e^{-\frac{E}{RT}}$



$Initially: U \leftarrow 0$

$if\ U_b + U < \varepsilon\hat{Z}$

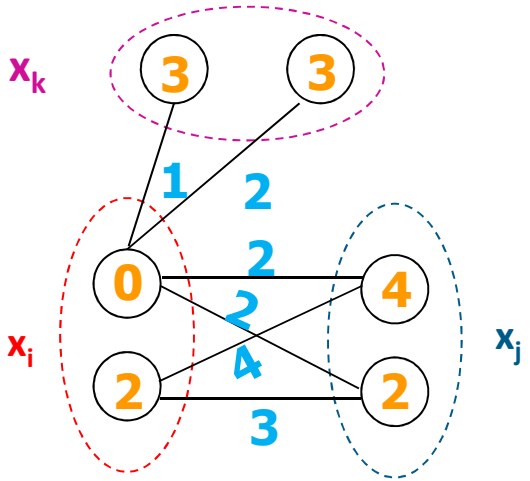
$Prune\ and\ U \leftarrow U + U_b$

$else\ Branch$

$At\ a\ leaf: \hat{Z} \leftarrow \hat{Z} + e^{-\frac{E}{RT}}$

Cost Function Networks (Toulbar2)

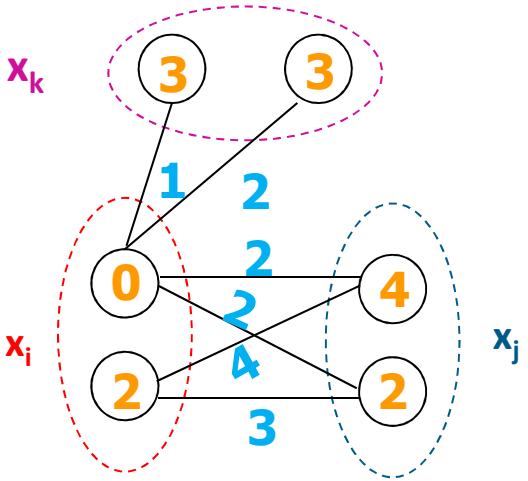
- $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ set of variables
- $\boldsymbol{D_i}$ set of domains over $\boldsymbol{x_i}$, $|\boldsymbol{D_i}| \leq \boldsymbol{d}$
- W set of non-negative cost functions w_S each with a scope S
- Goal: Minimize $\sum w_S \longrightarrow$ NP-hard
- $w_\emptyset$: constant function → Lower bound



$$w_\emptyset = 0$$

Local Consistency

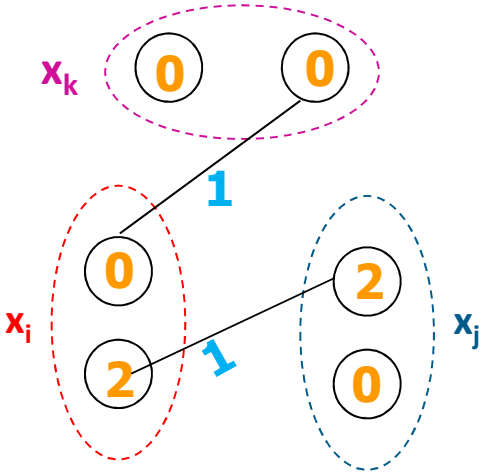
Local consistencies transform the problem into an equivalent one, increasing the upper bound $w_\emptyset$.



$w_\emptyset = 0$

LCs

EDAC (directional consistency)
VAC (virtual consistency)



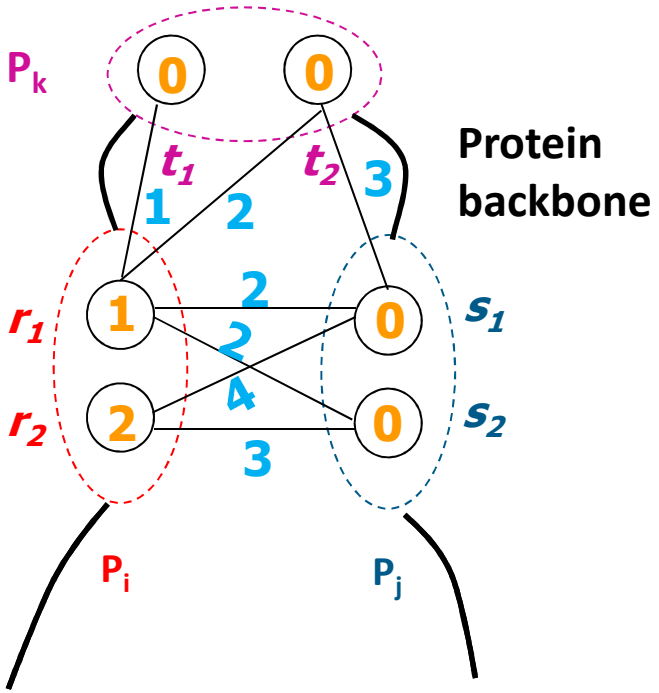
$w_\emptyset = 8$

Computational Protein Design as Cost Function Network

CPD as CFN

- n AA positions, $X = \{P_1, P_2, .. P_n\}$
- D_i set of rotamers of position P_i
- W pairwise energy functions

$$W = \{E(i), \ldots, E(i,j)\}$$



[Allouche et al. CP2012]
[Allouche et al. AI 2014]
[Traoré et al. Bioinformatics 2013]

Upper Bound on The Partition Function

Z_0^* algorithm:

$$U_b = N \times exp\left(\frac{-c_\emptyset}{RT}\right)$$

Takes in account the number of leaves N below the current node

Z_1^* algorithm:

$$U_b = exp\left(\frac{-c_\emptyset}{RT}\right) \prod_{i \in X} \sum_{a \in d_i} exp\left(\frac{-E_i(a)}{RT}\right)$$

Takes in account unary costs

Z_2^* algorithm:

$$U_b = Z_{STP}$$

Takes in account unary costs + binary costs on a spannning tree

Comparison $Z^*_{0,1,2}$ and K^*

$\varepsilon = 10^{-3}$	Z_0^*		Z_1^* *vs* Z_0^*		$(Z_1^* + \mathbf{VAC})$ *vs* Z_1^*		Z_2^* *vs* Z_1^*		K^*	
PDB ID (#Seq.)	Nodes	Times	Nodes	Times	Nodes	Times	Nodes	Times	Nodes	Times
1ACB (6)	129	0.2 sec	$\approx 0\%$	$\approx 0\%$	$\approx 0\%$	$\approx 0\%$	$\approx -2\%$	$\approx 0\%$	$\propto 10^5$	4,859 *min*
1AMU (1584)	8.45×10^4	$\frac{1}{2}$ *min*	$\approx -23\%$	$\approx -10\%$	$\approx +13\%$	$\approx -21\%$	$\approx -3\%$	$\approx +13\%$	6.45×10^6	1,278 *min*
3SGB (173)	2.2×10^6	30 *min*	$\approx 0\%$	$\approx 0\%$	$\approx 0\%$	$\approx -5\%$	$\approx -10\%$	$\approx +35\%$	∞	∞
1TP5 (1121)	3.19×10^6	31 *min*	$\approx -51\%$	$\approx -47\%$	$\approx 0\%$	$\approx -75\%$	$\approx -36\%$	$\approx +11\%$	∞	∞
1B74 (1809)	5.64×10^6	85 *min*	$\approx -41\%$	$\approx -35\%$	$\approx +1\%$	$\approx -70\%$	$\approx -9\%$	$\approx 17\%$	∞	∞
2Q2A (4716)	39.9×10^6	590 *min*	$\approx -56\%$	$\approx -45\%$	$\approx -1\%$	$\approx -72\%$	$\approx -5\%$	$\approx +4\%$	∞	∞

Limit time out: 250 h
64 GB RAM & 1 proc

Perspective

I
- Model and compute an approximation of K for more systems.

II
- Sequence-space elimination

III
- Tighter, more efficient $\boldsymbol{U_b}$

IV
- Compare rankings of affinities with experimental data (biophysics)

V
- Compare Z^* to other program (Cachet, WISH etc…)

Acknowledgements

- Biometrics & Artificial Intelligence Unit
  - David Allouche
  - George Katsirelos
  - Simon de Givry
  - Thomas Schiex
- Catalysis & Enzyme Molecular Engineering Team
  - Sophie Barbe

Branch&bound + Local Consistency

- Preserve the energies
- Increase the lower bound $C_\varnothing$