Positive multistate protein design

Jelena Vucinic, David Simoncini, Manon Ruffini, Sophie Barbe, Thomas Schiex

Supplementary information

POsitive Multistate Protein *d*esign

Jelena Vucinic[1,2], David Simoncini[1,3], Manon Ruffini [1,2], Sophie Barbe [1] and Thomas Schiex [2]**

[1]LISBP, Université de Toulouse, CNRS, INRA, INSA, Toulouse, France.
[2] MIAT, Université de Toulouse, INRA, Auzeville-Tolosane, France.
[3] IRIT UMR 5505-CNRS, Université de Toulouse, 31042 Cedex 9, France.

1 Proof of theorem 1

This Theorem is in two parts. The first part says that positive MSD is NP-complete, the second part says that negative MSD is $\mathsf{NP}^{\mathsf{NP}}$-complete. We start with the first part, proving that positive MSD is only NP-complete:

Proof. The problem is in NP because it is possible to verify a positive instance given a short certificate defined by a sequence $\mathbf{a} \in \prod_i S_i$ and a set of conformation sequences $\mathbf{c}_j = \arg\min_{\mathbf{c} \in \prod_i R^j_{i,a[i]}} E_j(\mathbf{a}, \mathbf{c})$ for sequence $\mathbf{a}$ on each of the backbone $B_j \in \mathbf{B}^+$. It suffices to compute the joint fitness of all states and check if it is lower than the threshold k. It is complete for NP since SSD is just the case where $|\mathbf{B}^+| = 1$ and is NP-complete (Pierce and Winfree, 2002). □

The second requires to show that the general $\oplus$-MSD problem is $\mathsf{NP}^{\mathsf{NP}}$-complete.

Proof. We must prove that:

- it belongs to the class $\mathsf{NP}^{\mathsf{NP}}$;
- any problem in $\mathsf{NP}^{\mathsf{NP}}$ reduces to $\oplus$-MSD in polynomial time.

  Let us introduce the following $\mathsf{NP}^{\mathsf{NP}}$-complete problem, called $\exists\forall$3DNF :

  Given two sets of propositional variables $\mathbf{p} = (p_1, \ldots, p_n)$ *and* $\mathbf{q} = (q_1, \ldots, q_m)$*, and a boolean formula* $H(\mathbf{p}, \mathbf{q})$ *over these variables, in disjunctive normal form (DNF), with each cube conjunction of three literals, is there a valuation* $\nu_{\mathbf{p}}$ *of* $\mathbf{p}$*, such that for any valuation* $\nu_{\mathbf{q}}$ *of the variables of* $\mathbf{q}$*,* $\nu_{\mathbf{q}}\nu_{\mathbf{p}}(H(\mathbf{p}, \mathbf{q}))$ *is true?*

Assuming we had a NP-oracle, that could solve any instance of SSD, it would be possible to verify a positive instance of $\oplus$-MSD defined by a sequence $\mathbf{a} \in \prod_i S_i$, by calling the oracle to compute the minimum conformation energy $E_j(\mathbf{a}) = \min_{\mathbf{c} \in \prod_i C^j_{i,\mathbf{a}[i]}} E_j(\mathbf{a}, \mathbf{c})$ on each backbone $B_j \in \mathbf{B}^+ \cup \mathbf{B}^-$ and combining these energies to check that

$$\left(\bigoplus_{B_j \in \mathbf{B}^+} \min_{\mathbf{c} \in \prod_i C^j_{i,\mathbf{a}[i]}} E_j(\mathbf{a}, \mathbf{c}) \right) - \left(\bigoplus_{B_j \in \mathbf{B}^-} \min_{\mathbf{c} \in \prod_i C^j_{i,\mathbf{a}[i]}} E_j(\mathbf{a}, \mathbf{c}) \right) \leq k$$

Let us reduce $\exists\forall$3DNF to $\oplus$-MSD . Let $\mathbf{p} = (p_1, \ldots, p_n)$, $\mathbf{q} = (q_1, \ldots, q_m)$, $H(\mathbf{p}, \mathbf{q})$ be a $\exists\forall$3DNF instance, where $H(\mathbf{p}, \mathbf{q}) = C_1 \vee \cdots \vee C_k$, and for each $i \in [|1, k|]$, l^1_i, l^2_i and l^3_i are the literals of C_i. Let us construct an instance of $\oplus$-MSD with $n + m + k$ variables, represented as a CFN:

- For each variable $p \in \mathbf{p}$, we introduce the variable V_p with domain $\{T, F\}$, representing the valuation of p;
- For each variable $q \in \mathbf{q}$, we introduce the variable V_q with domain $\{T, F\}$, representing the valuation of q;
- For each cube C_i of $H(\mathbf{p}, \mathbf{q})$, we introduce the variable V_{C_i} with domain $\{l^1_i, l^2_i, l^3_i\}$.

For each cube C_i and each variable $x \in \mathbf{p} \cup \mathbf{q}$ that appears in C_i, the binary cost between V_{C_i} and V_x is defined as follows:

$$E_{x,C_i} : (v, l_i) \in D^{V_x} \times D^{V_{C_i}} = \begin{cases} 1 & \text{if } v = T \text{ and } l_i = x \\ 1 & \text{if } v = F \text{ and } l_i = \neg x \\ 0 & \text{otherwise} \end{cases}$$

If the literal l_i is satisfied by the valuation v of its variable x, the corresponding binary cost is 1.

The total energy is the sum of the binary terms: $E = \sum_{x,C_i} E_{x,C_i}$. Given an assignment of all variables, the energy is zero if and only if all binary terms are zero. This means that for each cube C_i, the variable V_{C_i} is assigned a literal l_i that is not satisfied by the valuation $\nu_{\mathbf{p}}\nu_{\mathbf{q}}$ defined by the assignment. Finally, we consider the $\oplus$-MSD instance with a single negative backbone:

Does there exist $\mathbf{a} \in \prod_i D^{V_{p_i}}$ *such that:*

$$-\left(\min_{\mathbf{c}\in\prod_i D^{V_{C_i}}\times\prod_j D^{V_{q_j}}} E(\mathbf{a},\mathbf{c})\right) \leq -1$$

If the $\exists\forall$3DNF instance is positive, there exists a valuation $\nu_{\mathbf{p}}$ such that $\nu_{\mathbf{p}}H(\mathbf{p},\mathbf{q})$ is a tautology. So, if $\mathbf{a}$ is the assignment of the variables $V_p, p \in \mathbf{p}$ that corresponds to ν_p, then for any assignment of $V_q, q \in \mathbf{q}$, there always exists a cube C_i, which all literals are satisfied, hence, the binary cost is greater than 1. This is equivalent to:

$$\min_{\mathbf{c}\in\prod_i D^{V_{C_i}}\times\prod_j D^{V_{q_j}}} E(\mathbf{a},\mathbf{c}) \geq 1$$

So the $\oplus$-MSD instance is positive.

Conversely, if the $\oplus$-MSD instance is positive, there exists $\mathbf{a} \in \prod_i D^{V_{p_i}}$, corresponding to a valuation $\nu_{\mathbf{p}}$, such that the energy of any assignment of the remaining variables is greater than 1, meaning that $\nu_{\mathbf{p}}H(\mathbf{p},\mathbf{q})$ is a tautology.

Note that the $\oplus$-MSD instance consists of $n+m+k$ variables, each with domain size less than 3, and $k \times (n+m)$ binary energies, that can be described in a 3×2-sized matrix, where each coefficient is straightforward to compute. Therefore, the reduction is valid and polynomial.

□

2 Description of protein benchmark systems

Table $S1$: Description of protein systems: For each instance: system name, reference PDB id, crystallographic resolution or number of conformations for NMR structures, number of amino acid residues(N), SCOP stuctural classification(Class).

System name	PDB ID	Number of conformations	Number of residues
Saccharomyces cerevisiae J-domain	5vso	20	75
Human SNF5/INI1 domain	5l7b	20	75
Trypanosoma brucei Pex14 N-terminal domain	5mmc	20	70
Immunoglobulin binding domain of streptococcal protein G	1gb1	60	56
Cytotoxin-I from the venom of cobra N. oxiana	5t8a	20	61
E2 lipoyl domain from Thermoplasma acidophilum	2l5t	33	77
Spider toxin U4-hexatoxin-Hi1a	2n6r	20	76
Phl PII from timothy grass pollen	1bmw	38	96
Antibacterial factor-2	5ix5	20	68
Peptide toxin SsTx from Scolopendra subspinipes mutilans	5x0s	20	53
Rhabdopeptide NRPS Docking Domain Kj12A-NDD	6ews	20	63
Platelet integrin-binding C4 domain of von Willebrand factor	6fwn	20	85
Human ubiquitin at 298K	6qf8	20	79
Ubiquitin (Q41N variant)	6jlt	20	76
Sushi 1 domain of GABAbR1a	6hkc	20	75

System name	PDB ID	R(Å)	Number of residues
Hydrophobic protein from Soybean	1hyp	1.8	80
Alpha-amylase inhibitor hoe-467A	1hoe	2	74
E.coli Cold-shock protein A	1mjc	2	69
B1 immunoglobulin-binding domain of streptococcal protein G	1pga	2.07	56
PAS Factor from Vibrio vulnificus	2b8i	1.8	77
Apo-GolB	4y2k	1.7	65
Toxin isolated from the Malayan Krait	1f94	0.97	63
Allergen phl p2	1who	1.9	96
Alpha-spectrin src homology 3 domain	1tud	1.77	62
Headpiece Domain of Chicken Villin	1yu5	1.4	67
Ribosomal protein L30 from thermus thermophilus	1bxy	1.9	60
C-TERMINAL DOMAIN OF THE RIBOSOMAL PROTEIN L7/L12	1ctf	1.7	74
C-Myb DNA-Binding Domain	1guu	1.6	52
Domain 3 of human alpha polyC binding protein	1wvn	2.1	82
Type III Antifreeze Protein RD1 from an Antarctic Eel Pout	1ucs	0.62	64

3 Solving positive min -MSD with iCFN

Recently, a guaranteed CFN-based algorithm for both positive and negative min-MSD was introduced as the iCFN method (Karimi and Shen, 2018). The authors did not use a reduction of the problem to CFN but proposed and implemented a new algorithm that exploits some of the underlying machinery of CFN algorithms (arc consistencies (Cooper *et al.*, 2010)). The authors showed that their method outperforms the guaranteed COMETS software (Hallen and Donald, 2016). We therefore decided to compare POMP^d against iCFN only.

The iCFN website (https://shen-lab.github.io/software/iCFN/) gives access to both the software in binary format and to multistate design energy matrices. We wrote a first python script to translate iCFN-formatted problems into the `cfn.gz` CFN format that can be directly read by the CFN solver toulbar2. iCFN uses double resolution floating point energies and the `cfn.gz` format relies on a fixed point representation of energies. We used a "6 digits after the decimal point" representation. We wrote a second python/PyRosetta script to generate energy matrices in iCFN-format directly from PyRosetta. These scripts make it possible to either apply POMP^d to the positive min-MSD instances available on the iCFN website or to apply the iCFN algorithm on our benchmark set (for the min-MSD problem only as iCFN is not able to tackle Σ-MSD).

The iCFN command line used on the positive min-MSD problems was `iCFN -just_pos -ecutDEE=2 -ecutDEE_across=2 -ecutDEE_seq=10 -ecut_stability=5 -max_conf_seq=1 -max_dis_seq=9999` ⟨`files`⟩ which asks for one solution of the min-MSD problem, with no limitation on the number of mutations in the produced design sequence. Except for the effect of the various pruning thresholds used by iCFN that reduce computing time, this precisely matches the min-MSD problem we solve using CFN reductions.

The iCFN multistate designs use a specific rotamer library that includes 2 extra protonated states for glutamate (Glu) and aspartate (Asp) as well as 3 protonated states for histidine (His). Because the 'cpd' branch of toulbar2 relies on the one letter code of amino acids, it is currently unable to process the corresponding energy matrices. We therefore used the 'master' branch of toulbar2 to solve these problems. The command line used in this case is simply `-m -hbfs:` which deactivates the default Hybrid Best First Search algorithm (Allouche *et al.*, 2015) for a simple Depth-First Search and activates the median cost variable ordering heuristic (Allouche *et al.*, 2014). All computations were done on a laptop equipped with 16GB of RAM and a Intel(R) Core(TM) i7-7600U CPU at 2.80GHz.

4 Clustering distance thresholds

Table $S2$: User-defined clustering distance thresholds (d) for each protein structure.

NMR structures		X-ray structures	
PBD ID	d (Å)	PBD ID	d (Å)
5vso	2.0	1hyp	0.5
5l7b	0.4	1hoe	0.4
5mmc	2.0	1mjc	0.6
1gb1	0.3	1pga	0.4
5t8a	0.2	2b8i	0.4
2l5t	1.0	4y2k	0.4
2n6r	0.3	1f94	0.4
1bmw	1.2	1who	0.4
5ix5	0.6	1tud	0.5
5x0s	1.5	1yu5	0.15
6ews	0.5	1bxy	0.5
6fwn	0.8	1ctf	0.3
6qf8	1.2	1guu	0.3
6jlt	0.5	1wvn	0.5
6hkc	1.0	1ucs	0.3

5 Search space sizes for different design problems

Table $S3$: Multistate design problems: for each problem we give the average search space of four SSD problems, search space for the min-MSD problem, defined as the sum of all SSD search space sizes, the raw Σ-MSD search space size, defined by the product of the size of all variable domains and the search space size reduced by the SS constraints that impose that all states use the same sequence.

PBD ID	average $\overline{SSD}$ search space	min-MSD search space	Σ-MSD search space	Σ-MSD reduced search space
		NMR structures		
5vso	$1.3\ 10^{181}$	$5.4\ 10^{181}$	$8.5\ 10^{723}$	$1.6\ 10^{431}$
5l7b	$2.6\ 10^{170}$	$1.0\ 10^{171}$	$6.4\ 10^{680}$	$3.1\ 10^{411}$
5mmc	$5.6\ 10^{158}$	$2.3\ 10^{159}$	$4.1\ 10^{634}$	$8.3\ 10^{380}$
1gb1	$2.5\ 10^{137}$	$1.0\ 10^{138}$	$5.9\ 10^{547}$	$1.6\ 10^{329}$
5t8a	$9.4\ 10^{133}$	$3.8\ 10^{134}$	$5.9\ 10^{535}$	$4.8\ 10^{297}$
2l5t	$2.2\ 10^{185}$	$9.0\ 10^{185}$	$7.2\ 10^{738}$	$2.1\ 10^{438}$
2n6r	$3.4\ 10^{168}$	$1.3\ 10^{169}$	$4.0\ 10^{673}$	$9.3\ 10^{376}$
1bmw	$5.2\ 10^{229}$	$2.1\ 10^{230}$	$1.1\ 10^{914}$	$1.4\ 10^{547}$
5ix5	$4.4\ 10^{148}$	$1.7\ 10^{149}$	$2.0\ 10^{593}$	$7.8\ 10^{327}$
5x0s	$1.2\ 10^{119}$	$4.9\ 10^{119}$	$1.4\ 10^{470}$	$2.0\ 10^{263}$
6ews	$8.1\ 10^{155}$	$3.2\ 10^{156}$	$4.3\ 10^{622}$	$5.4\ 10^{376}$
6fwn	$2.8\ 10^{188}$	$1.1\ 10^{189}$	$2.4\ 10^{751}$	$4.3\ 10^{419}$
6qf8	$2.3\ 10^{188}$	$9.1\ 10^{188}$	$4.0\ 10^{750}$	$9.3\ 10^{453}$
6jlt	$6.9\ 10^{188}$	$2.8\ 10^{189}$	$1.5\ 10^{755}$	$3.5\ 10^{458}$
6hkc	$4.5\ 10^{174}$	$1.8\ 10^{175}$	$6.6\ 10^{694}$	$1.2\ 10^{402}$
		Xray structures		
1hyp	$2.1\ 10^{166}$	$8.2\ 10^{166}$	$3.2\ 10^{664}$	$4.8\ 10^{375}$
1hoe	$2.2\ 10^{171}$	$8.9\ 10^{171}$	$5.8\ 10^{684}$	$8.6\ 10^{395}$
1mjc	$4.3\ 10^{165}$	$1.7\ 10^{166}$	$10.0\ 10^{661}$	$4.9\ 10^{392}$
1pga	$4.7\ 10^{137}$	$1.9\ 10^{138}$	$1.5\ 10^{550}$	$4.0\ 10^{331}$
2b8i	$7.7\ 10^{189}$	$3.1\ 10^{190}$	$5.1\ 10^{758}$	$1.5\ 10^{458}$
4y2k	$2.5\ 10^{161}$	$9.8\ 10^{161}$	$1.7\ 10^{644}$	$3.4\ 10^{390}$
1f94	$2.9\ 10^{134}$	$1.2\ 10^{135}$	$1.4\ 10^{537}$	$1.8\ 10^{291}$
1who	$2.6\ 10^{227}$	$1.0\ 10^{228}$	$6.1\ 10^{907}$	$7.8\ 10^{540}$
1tud	$3.1\ 10^{146}$	$1.3\ 10^{147}$	$5.0\ 10^{585}$	$3.3\ 10^{351}$
1yu5	$9.4\ 10^{165}$	$3.8\ 10^{166}$	$2.9\ 10^{663}$	$9.0\ 10^{401}$
1bxy	$5.8\ 10^{147}$	$2.3\ 10^{148}$	$7.6\ 10^{590}$	$5.0\ 10^{356}$
1ctf	$5.2\ 10^{164}$	$2.1\ 10^{165}$	$1.9\ 10^{658}$	$9.38\ 10^{388}$
1guu	$1.2\ 10^{123}$	$4.7\ 10^{123}$	$1.4\ 10^{492}$	$1.2\ 10^{293}$
1wvn	$8.0\ 10^{179}$	$3.2\ 10^{180}$	$4.5\ 10^{718}$	$6.8\ 10^{429}$
1ucs	$5.9\ 10^{153}$	$2.4\ 10^{154}$	$7.9\ 10^{614}$	$1.3\ 10^{365}$

Table $S4$: iCFN multistate design problems: for each problem we give the position of the redesigned residue, the number of flexible residues around the redesigned residue and the search space for the min-MSD problem, defined as the sum of all SSD search space sizes, the raw Σ-MSD search space size, defined by the product of the size of all variable' domains and the actual search space size, reduced by the SS constraints that impose that all states use the same sequence.

redesigned position	# of flexible residues	min-MSD search size	Σ-MSD search size	Σ-MSD reduced search size
26	18	$7.6\ 10^{30}$	$1.6\ 10^{323}$	$7.7\ 10^{308}$
28	18	$3.1\ 10^{34}$	$6.3\ 10^{362}$	$3.1\ 10^{348}$
98	19	$4.9\ 10^{31}$	$7.7\ 10^{334}$	$3.7\ 10^{320}$
100	29	$1.4\ 10^{42}$	$5.2\ 10^{447}$	$2.5\ 10^{433}$

6 Energy difference between SSD optimal sequences and Σ-MSD sequence

Table $S5$: Difference in energy for each protein in the benchmark between the average of all SSD optimal sequences and the energy of the optimal Σ-MSD sequence (kcal).

NMR PDB	Σ-MSD-$\overline{SSD}$	X-ray PDB	Σ-MSD-$\overline{SSD}$
5vso	16.0	1hyp	12.7
5l7b	10.4	1hoe	14.8
5mmc	14.3	1mjc	8.9
1gb1	11.6	1pga	12.8
5t8a	5.6	2b8i	13.3
2l5t	25.4	4y2k	5.5
2n6r	11.7	1f94	9.2
1bmw	44.9	1who	17.5
5ix5	21.7	1tud	4.2
5x0s	30.3	1yu5	6.3
Mean	19.2	Mean	10.5

7 CPU-time for SSD and Σ-MSD as a function of protein size

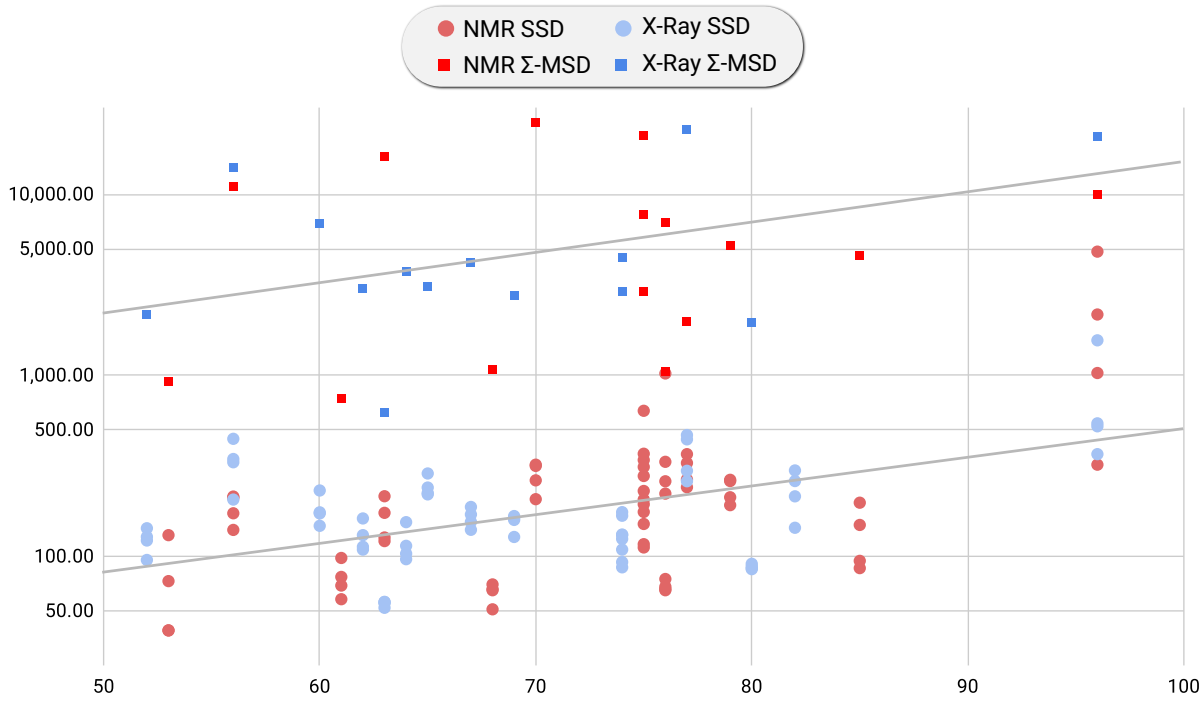

Figure *S*1: The CPU-time (Y logscale axis) is represented for SSD and Σ-MSD for both NMR and X-ray structures as a function of the protein size (X-axis). The general trend is exponential as expected with closely related slopes but a constant shift in computational cost by a factor of 1.5 orders of magnitude.q

8 3D representation of local optima networks

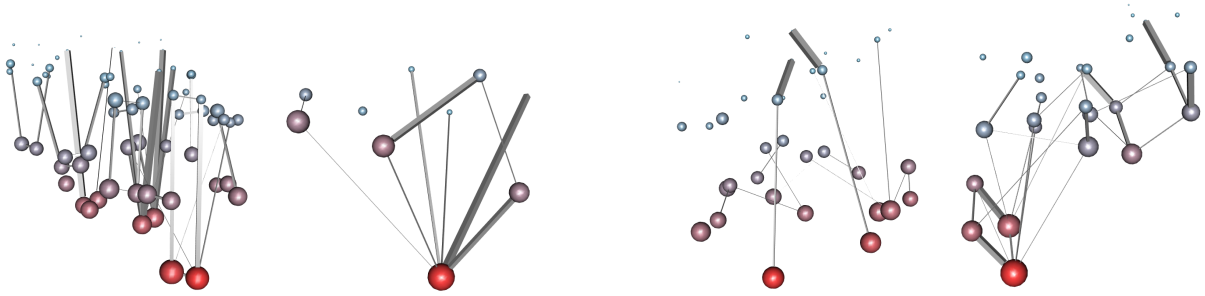

Figure *S*2: 3D view of local optima networks. From left to right: 1bmw with min-MSD and Σ -MSD, 1who with min-MSD and Σ -MSD.

References

Allouche, D. *et al.* (2014). Computational protein design as an optimization problem. *Artificial Intelligence*, **212**, 59–79.

Allouche, D. *et al.* (2015). Anytime hybrid best-first search with tree decomposition for weighted csp. In *International Conference on Principles and Practice of Constraint Programming*, pages 12–29. Springer.

Cooper, M. C. *et al.* (2010). Soft arc consistency revisited. *Artificial Intelligence*, **174**(7-8), 449–478.

Hallen, M. A. and Donald, B. R. (2016). Comets (constrained optimization of multistate energies by tree search): A provable and efficient protein design algorithm to optimize binding affinity and specificity with respect to sequence. *Journal of Computational Biology*, **23**(5), 311–321.

Karimi, M. and Shen, Y. (2018). iCFN: an efficient exact algorithm for multistate protein design. *Bioinformatics*, **34**(17), i811–i820.

Pierce, N. A. and Winfree, E. (2002). Protein design is np-hard. *Protein engineering*, **15**(10), 779–782.