Coupling CP with Deep Learning for Molecular Design and SARS-CoV2 Variants Exploration

Thomas Schiex
Universite Fédérale de Toulouse, ANITI, INRAE, UR 875, 31326 Toulouse, France

Abstract

The use of discrete optimization, including Constraint Programming, for designing objects that we completely understand is quite usual. In this talk, I'll show how designing specific biomolecules (proteins) raises new challenges, requiring solving problems that combine precise design targets, approximate laws, and design rules that can be deep-learned from data.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning; Theory of computation → Constraint and logic programming; Computing methodologies → Learning in probabilistic graphical models

Keywords and phrases graphical models, deep learning, constraint programming, cost function networks, random Markov fields, decision-focused learning, protein design

Digital Object Identifier 10.4230/LIPIcs.CP.2023.4

Category Invited Talk

1 Introduction

Proteins are biomolecules that support most mechanisms in living organisms, from viruses to human beings. They already have major commercial applications in green chemistry (as enzymes) but also in the health domain (e.g., the anti-CoViD Regeneron™ antibodies are proteins). Most commercially used proteins are either natural proteins or engineered versions of natural proteins. To go beyond the repertoire of natural proteins, it is important to be able to reliably and efficiently design new proteins, with new capacities [9]. Proteins are defined by their amino acid sequence, a discrete object defined over an alphabet of 20 characters. Once the sequence of a protein is fixed, it can be encoded into a suitable microbe, enabling the cheap manufacturing of these complex microscopic assemblies.

Optimization is often used to design objects such as schedules, assignments, time-tables or packing, which we completely understand. Instead, proteins are tiny physical objects that live in the realm of quantum physics. Their behavior is hard to formally, precisely and concisely capture. Designing new proteins therefore requires to combine knowledge, expressed as approximate laws of physics, with targeted design constraints and criteria, in the context of large sets of data of past successful designs (natural proteins) that also embody the many hidden laws which a successfully expressed protein must satisfy.

2 Designing proteins and SARS-CoV2 variants with CP

In this talk, we will see how Cost Function Networks (CFNs), a weighted variant of Constraint Networks/CP) can help us design new proteins [1, 6]. Alone, CFNs can already capture both logical information (constraints) and numerical information, enabling the simultaneous representation of approximate laws of physics and design targets. By solving suitable instances of Weighted Constraint Satisfaction Problems, one can already produce protein sequences that can be tested *in silico* (with e.g., AlphaFold2 [7]), characterized experimentally, and lead to successful designs [8].

© Thomas Schiex;
licensed under Creative Commons License CC-BY 4.0

By leveraging the exhaustive enumeration capabilities of exact discrete solvers, it becomes possible to tackle previously unsolvable questions. To infect us, the SARS-Cov2 virus relies on its own spike protein, designed by evolution to be stable and efficiently bind to the human ACE2 receptor. Using a protein structure produced in the early months of 2020, we exhaustively enumerated SARS-CoV2 variants that would, in theory, bind to ACE2 and kept those that remained sufficiently stable. After a drastic selection among tenths of millions of predicted variants, 59 sequences were tested experimentally for affinity, infectivity, and resistance to antibodies, resulting in a list of non-yet-existing infectious therapeutic-antibodies-resistant variants that could be used to design vaccines proactively [3].

3 Learning how to play the Protein Design and Sudoku games

Because the laws of physics and modeling assumptions used in such approaches lead to approximate results, it becomes crucial to exploit the massive amount of data that has been produced by experimentalists in terms of natural protein structures and sequences. This raises the exciting question of learning CFNs describing the "quality" of sequences for a given protein structure to eventually learn how to design proteins. This problem is reminiscent of learning "how to reason" or "how to play Sudoku" which has been addressed by various recent decision-focused learning architectures. By leveraging a usual probabilistic interpretation of CFNs, we recently proposed a simple scalable learning architecture [4] that combines Deep Learning with an exact CFN solver (toulbar2 [2]) to learn how to design proteins (or how to play Sudoku) which outperforms existing architectures in terms of training time, data-efficiency and accuracy. Because solving the WCSP is NP-hard, powerful polynomial time relaxations then become handy [5].

References

1 David Allouche, Isabelle André, Sophie Barbe, Jessica Davies, Simon de Givry, George Katsirelos, Barry O'Sullivan, Steve Prestwich, Thomas Schiex, and Seydou Traoré. Computational protein design as an optimization problem. *Artificial Intelligence*, 212:59–79, 2014.

2 David Allouche, Simon De Givry, George Katsirelos, Thomas Schiex, and Matthias Zytnicki. Anytime hybrid best-first search with tree decomposition for weighted CSP. In *Principles and Practice of Constraint Programming: 21st International Conference, CP 2015, Cork, Ireland, August 31–September 4, 2015, Proceedings 21*, pages 12–29. Springer, 2015.

3 Mireia Solà Colom, Jelena Vucinic, Jared Adolf-Bryfogle, James W Bowman, Sébastien Verel, Isabelle Moczygemba, Thomas Schiex, David Simoncini, and Christopher D Bahl. Deep evolutionary forecasting identifies highly-mutated SARS-CoV-2 variants via functional sequence-landscape enumeration. *Research Square*, pages rs–3, 2022.

4 M. Defresne, S. Barbe, and T. Schiex. Scalable coupling of deep learning with logical reasoning. In *Proc. of the 32nd IJCAI*, Macau, A.S.R., China, 2023.

5 Valentin Durante, George Katsirelos, and Thomas Schiex. Efficient low rank convex bounds for pairwise discrete graphical models. In *International Conference on Machine Learning*, pages 5726–5741. PMLR, 2022.

6 Mark A Hallen and Bruce R Donald. Protein design by provable algorithms. *Communications of the ACM*, 62(10):76–84, 2019.

7 John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

8 Hiroki Noguchi, Christine Addy, David Simoncini, Staf Wouters, Bram Mylemans, Luc Van Meervelt, Thomas Schiex, Kam YJ Zhang, Jeremy RH Tame, and Arnout RD Voet. Computational design of symmetrical eight-bladed β-propeller proteins. *IUCrJ*, 6(1):46–55, 2019.

9 Ilan Samish, editor. *Computational Protein Design*. Humana New York, NY, 2017. doi:10.1007/978-1-4939-6637-0.