Centre Inria at Université Côte d'Azur, France
Group: Algorithms-Biology-Structure
Frederic.Cazals@inria.fr
Web: `http://team.inria.fr/abs`

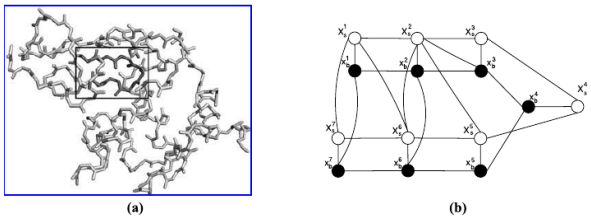

Figure 1: Protein structure in which the spatial proximity between amino acids is encoded using a graphical model. From [1].

PhD thesis proposal

Attention mechanisms for graphical models, with applications to protein structure analysis.

Keywords: deep learning, attention mechanisms, transformers, belief propagation, approximation algorithms, free energy, protein structure analysis.

Context. Emerging from the field of natural language processing, (self-)attention mechanisms have proven essential to understand the coupling between tokens in a sentence [2, 3]. In a different context, graphical models make it possible to express the conditional dependence of random variables encoded in graph nodes via the edges of the graph. On such models, message passing algorithms provide effective ways to compute various quantities of interest, in particular partition functions and free energies [4, 5].

Recently, attention mechanisms have also proven key to encode the coupling between spatial patterns observed between amino acids in a protein structure [6]. The corresponding tool, `AlphaFold2` by Deepmind, is considered a major achievement to predict a plausible structure of a protein from its amino-acid sequence. In related work, message passing algorithms have been used to compute average properties of proteins [1].

Goals. `AlphaFold2` is a key achievement but outputs a single structure. In fact, statistical physics teaches us that observable properties of molecules depend on ensemble of conformations (weighted by Boltzmann's factor). (For a gentle introduction, see also AI, molecular design and the Covid19.)

The goal of this PhD thesis will be to extend attention mechanisms in the context of graphical models, to study ensembles of conformations rather than isolated observations. The generation of conformations and the local check of their coherence will be based on advanced geometric models for protein geometry, which we recently developed [7, 8]. The work envisioned encompasses the design and analysis of algorithms, their coding (C++ and python), as well their experimental evaluation.

Training. Master 2 or equivalent degree in Computer science (algorithms) or machine learning or bioinformatics or biophysics.

References

[1] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. *Journal of Computational Biology*, 15(7):755–766, 2008.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[3] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

[4] Jonathan S Yedidia, William T Freeman, Yair Weiss, et al. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236-239):0018–9448, 2003.

[5] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

[6] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[7] T. O'Donnell and F. Cazals. Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry. *J. Comp. Chem.*, 2023.

[8] T. O'Donnell, V. Agashe, and F. Cazals. Geometric constraints within tripeptides and the existence of tripeptide reconstructions. *J. Comp. Chem.*, 2023.

Executive summary

(En) Alphafold by Deepmind predicts the structure of a protein from its amino acid sequence. This is considered a key achievement boosting the investigations of protein functions by biologists. Alas, Alphafold predicts a unique conformation while protein functions rely on ensembles of conformations.
The goal of this PhD thesis is be to extend attention mechanisms (which are key in Alphafold) in the context of graphical models, to study ensembles of molecular conformations rather than isolated conformations. The generation of conformations and the local check of their coherence will borrow and expand ideas in the realms of machine learning, geometry, and high dimensional sampling.
(Fr) Alphafold développé par Deepmind prédit la structure d'une protéine à partir de sa séquence d'acides aminés. Cette contribution est considérée comme majeure, car elle permet aux biologistes d'accélérer l'étude des fonctions des protéines. Hélas, Alphafold prédit une structure unique, alors que la fonction d'un protéine dépend d'un ensemble de conformations.
L'objet de cette thèse sera d'étendre les mécanismes d'attention (qui jouent un rôle clef dans Alphafold) dans le contexte des modèles graphiques, pour étudier des ensembles de conformations. La génération de conformations et l'étude de leur cohérence utilisera et contribuera a des thèmes tels que l'apprentissage machine, la géométrie, et les algorithmes d'échantillonage en grande dimension.