Algorithms and learning for protein science*

Frederic.Cazals@inria.fr

March 1, 2025

Rationale. Proteins underlie all biological functions, yet, understanding their mechanisms at the atomic scale remains a fundamental open problem. The difficulties are inherent to complex dynamics in very high dimensional spaces. Indeed, with circa 5000 atoms and xyz coordinates per atom, a polypeptide chain of median size lives in a configuration space of dimension 15,000. While `AlphaFold` has been a game changer by providing plausible structures of selected (well folded) regions of proteins, it by no means provide insights on dynamics.

In this context, the goal of this class is twofold. First, to cast the main problems related to protein dynamics into a rigorous mathematical / algorithmic framework. Second, to present some of the major ongoing developments, which feature a stimulating interplay between theoretical biophysics, geometry, topology, and machine learning. To get acquainted with real data, one course will be devoted to a computer practical providing background on standard molecular manipulations, and illustrating selected methods studied in class.

Pre-requisites. Training in algorithms / machine learning. Interest for biophysics / biology / medicine.

MVA: positioning.

- Themes: track Santé, machine learning, theory.
- Topics: proteins, molecular conformations, thermodynamics, dynamics, high dimensional spaces, kinematics, sampling, (free) energies.

Course overview. The courses consists of 6 lectures (*cours magistral*) of 3 hours each; plus one lecture mixing theory and demos.

The class is taught in English.

*Original vesion: May 2024. Updated along with the lectures

1 Understanding protein functions at the atomic level

The goal of this lecture is to get acquainted with the structure, dynamics, and functions of proteins. Doing so requires bridging the gap between notions from biology, chemistry, (statistical) physics, and kinematics.

- Proteins: structure and determination; [1], [2]
- Molecular coordinates and potential energy; [3]
- Thermodynamics, PMF and free energy; [4], [5]
- Kinetics, Markov State Models; [6], [7]
- Dynamics and time scales; [8]
- Binding affinity [9]
- Two molecular mechanisms: antibody-antigen interactions, membrane transport; [10], [11], [12]
- Open (mathematical, algorithmic) challenges in protein science

2 `k-means` and mixtures, with applications to torsion angles in flat torii

This lecture focuses on the important topic of torsion angles in proteins (from the backbone and side chains), and their joint probability densities. Studying these requires a proper understanding of mixtures in general (whence the link with Kmeans), and mixture fitting procedures.

- Torsion angles as soft coordinates.
- Kmeans clustering and its initialization [13, 14, 15, 16]
- Gaussian mixture models [17]
- Rotamers and multivariate densities on flat torii; [18]
- Minimum message length and applications to mixture of von Mises [19, 20]

3 Molecular kinematics, inverse problems, loop sampling

This lecture focuses on inverse problems for kinematic chains – aka loop closure problems, and the application to protein loop sampling procedures.

- Direct versus inverse problems: molecular dynamics vs inverse kinematics
- Modeling proteins using internal coordinates; [21]
- Kinematics and loop closure problems; [22], [23]
- Inverse problems and protein loop sampling; [24, 25]

4 Structural alignments and analysis on static structures

This lecture will review the mathematics and algorithms of must-know basic operations on static proteins structures. Implementations available in the Structural Bioinformatics Library [26].

- Molecular surfaces, volumes, and interfaces – using Voronoi diagrams and α-shapes [27]
- Procruste probblems [28], [29]
- Molecular distances [30], [31], [32]
- Iterative structural alignments [33]
- Identification of rigid domains using spectral clustering [34]
- Application: unveiling the mechanism of an antibiotic efflux pump [35]

5 Motions and energies: normal modes, (free) energies, ICA, tICA

This lecture will review the mathematics and algorithms of must-know basic operations on ensembles of molecular conformations. Implementations available in the Structural Bioinformatics Library [26].

- Normal modes [36, 37, 38]
- Thermodynamics and free energies: simplified models [39]
- Time-lagged independent component analysis (tICA) [40]
- Computing in statistical physics: the Wang-Landau algorithm; [41], [42]
- Connexion with volume calculation and polytopes [43], [44], [45]

6 Spatial partitions in high dimensional spaces, nearest neighbors and their significance

Tree like structures are key for nearest neighbor searches in high dimensional spaces, and to perform feature selection. This lecture will review reference methods for these two tasks, which have applications in structural bioinformatics and beyond.

- Random projection trees; [46]
- Applications to nearest neighbor finding, regression, dimension estimation; [47]
- Concentration of distances and significance of nearest neighbors

7 From Darwin to AlphaFold: MSA, DCA, attention mechanisms, and structure prediction

Protein sequences play a crucial role in providing evolution related pieces of information. This lecture will review models to capture couplings in multiple sequence alignments, and the application of these in `AlphaFold`. The quality of `AlphaFold` predictions will also be studied.

- Protein sequences and multiple sequence alignments; [48]
- Direct coupling analysis and variants; [49], [50]
- Coupling analysis with factored transformers; [51, 52], [53]
- `AlphaFold`; [54]
- `AlphaFold-DB`: assessment of reconstructions; [55]

8 General references

- Bioinformatics: [48]
- Biophysics and theoretical biophysics: [2], [4], [56]
- Algorithms, machine learning: [5], [43], [57]

9 Validation mode

Projects for students working in tandem (15 points) + individual quizz and/or short exercises (5 points).

A project will consists of reproducing / expanding results recently published. Students are asked to return a report, plus a git repo / notebook / code archive. Students will be given one month to complete the project.

Catch-up: oral exam on the lectures–typically a quizz with one or two questions per lecture.

10 Links with other classes

Overall, this class would complement lectures of the *Track Santé*, providing a molecular view of various topics studied at a macroscopic level in the courses *Méthodes mathématiques pour les neurosciences* (E. TANRE, R. VELTZ), and *Medical image analysis based on generative, geometric and biophysical models* (H. DELINGETTE, X. PENNEC).

More specifically, several topics are linked to other classes, in particular: lecture 2 (`AlphaFold`) is connected the course *Deep learning* (V.LEPETIT, M. VAKALOPOULOU); lecture 3 (loop sampling) is partly connected to the course *Computational statistics* (S. Alassonniere), as it covers the design of probabilistic mixtures in flat torii (using the Minimum Message Length framework); lecture 4 (High-dimensional sampling) is connected to lectures dealing with sampling / MCMC in general, with a specific view as we deal with molecular geometry models.

References

[1] Carl Ivar Branden and John Tooze. *Introduction to protein structure*. Garland Science, 2012.

[2] John Kuriyan, Boyana Konforti, and David Wemmer. *The molecules of life: Physical and chemical principles*. Garland Science, 2012.

[3] D. J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

[4] K. Dill and S. Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 2010.

[5] T. Lelièvre, G. Stoltz, and M. Rousset. *Free energy computations: A mathematical perspective*. World Scientific, 2010.

[6] J.C. Schön and M. Jansen. Prediction, determination and validation of phase diagrams via the global study of energy landscapes. *Int. J. of Materials Research*, 100(2):135, 2009.

[7] V. Pande, K. Beauchamp, and G.R. Bowman. Everything you wanted to know about markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[8] S.A. Adcock and A.J. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–1615, 2006.

[9] H-X. Zhou and M. Gilson. Theory of free energy and entropy in noncovalent binding. *Chemical reviews*, 109(9):4092–4107, 2009.

[10] Ruchao Peng, Lian-Ao Wu, Qingling Wang, Jianxun Qi, and George Fu Gao. Cell entry by SARS-CoV-2. *Trends in biochemical sciences*, 46(10):848–860, 2021.

[11] Satoshi Murakami. Multidrug efflux transporter, AcrB–the pumping mechanism. *Current opinion in structural biology*, 18(4):459–465, 2008.

[12] A. Schmidt, H. Xu, A. Khan, T. O'Donnell, S. Khurana, L. King, J. Manischewitz, H. Golding, P. Suphaphiphat, A. Carfi, E. Settembre, P. Dormitzer, T. Kepler, R. Zhang, A. Moody, B. Haynes, H-X. Liao, D. Shaw, and S. Harrison. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *PNAS*, 110(1):264–269, 2013.

[13] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SODA*, page 1035. Society for Industrial and Applied Mathematics, 2007.

[14] Silvio Lattanzi and Christian Sohler. A better k-means++ algorithm via local search. In *International Conference on Machine Learning*, pages 3662–3671. PMLR, 2019.

[15] Lorenzo Beretta, Vincent Cohen-Addad, Silvio Lattanzi, and Nikos Parotsidis. Multi-swap k-means++. *Advances in Neural Information Processing Systems*, 36:26069–26091, 2023.

[16] Guillaume Carrière and Frédéric Cazals. Improved seeding strategies for k-means and gaussian mixture fitting with expectation-maximization. *Submitted*, 2024.

[17] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[18] P. Amarasinghe, L. Allison, P. Stuckey, M. Garcia de la Banda, A. Lesk, and A. Konagurthu. Getting '$\phi\psi\chi$al'with proteins: minimum message length inference of joint distributions of backbone and sidechain dihedral angles. *Bioinformatics*, 39(Supplement_1):i357–i367, 2023.

[19] Chris S Wallace and Peter R Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 49(3):240–252, 1987.

[20] Christopher S Wallace. *Statistical and inductive inference by minimum message length*. Springer Science & Business Media, 2005.

[21] T. O'Donnell and F. Cazals. Modeling the dynamics of proteins: techniques from geometry and kinematics. *In preparation*, 2024.

[22] E. Coutsias, C. Seok, M. Jacobson, and K. Dill. A kinematic view of loop closure. *Journal of computational chemistry*, 25(4):510–528, 2004.

[23] Kimberly Noonan, David O'Brien, and Jack Snoeyink. Probik: Protein backbone motion by inverse kinematics. *The International Journal of Robotics Research*, 24(11):971–982, 2005.

[24] T. O'Donnell, V. Agashe, and F. Cazals. Geometric constraints within tripeptides and the existence of tripeptide reconstructions. *J. Comp. Chem.*, 44(13):1236–1249, 2023.

[25] T. O'Donnell and F. Cazals. Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry. *J. Comp. Chem.*, 44(11):1094–1104, 2023.

[26] F. Cazals and T. Dreyfus. The Structural Bioinformatics Library: modeling in biomolecular science and beyond. *Bioinformatics*, 7(33):1–8, 2017.

[27] F. Cazals, H. Kanhere, and S. Loriot. Computing the volume of union of balls: a certified algorithm. *ACM Transactions on Mathematical Software*, 38(1):1–20, 2011.

[28] G. Golub and C.F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[29] Jim Lawrence, Javier Bernal, and Christoph Witzgall. A purely algebraic justification of the kabsch-umeyama algorithm. *Journal of research of the National Institute of Standards and Technology*, 124:1, 2019.

[30] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

[31] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. tm-score–tm-align.

[32] F. Cazals and R. Tetley. Characterizing molecular flexibility by combining lRMSD measures. *Proteins: structure, function, and bioinformatics*, 87(5):380–389, 2019.

[33] D. Ritchie, A. Ghoorah, L. Mavridis, and V. Venkatraman. Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, 28(24):3274–3281, 2012.

[34] F. Cazals, J. Herrmann, and E. Sarti. Simpler protein domain identification using spectral clustering. *Proteins: structure, function, and bioinformatics*, NA(NA), 2025.

[35] M. Simsir. *Modélisation structurale des pompes à efflux de la famille des RND: de la résistance aux antibiotiques à la résistance à la chimiothérapie*. Phd thesis, Université Côte d'Azur, Nice, 2020.

[36] T. Ichiye and M. Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Genetics*, 11(3):205–217, 1991.

[37] Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[38] Ivet Bahar, Timothy R Lezon, Ahmet Bakan, and Indira H Shrivastava. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical reviews*, 110(3):1463–1497, 2010.

[39] G. Meng, N. Arkus, M.P. Brenner, and V.N. Manoharan. The free-energy landscape of clusters of attractive hard spheres. *Science*, 327(5965):560–563, 2010.

[40] Steffen Schultze and Helmut Grubmüller. Time-lagged independent component analysis of random walks and protein dynamics. *Journal of Chemical Theory and Computation*, 17(9):5766–5776, 2021.

[41] G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre, and G. Stoltz. Convergence of the Wang-Landau algorithm. *Mathematics of Computation*, 84(295):2297–2327, 2015.

[42] A. Chevallier and F. Cazals. Wang-Landau algorithm: an adapted random walk to boost convergence. *J. of Computational Physics*, 410(1):1–19, 2020.

[43] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of data science*. Cambridge, 2020.

[44] B. Cousins and S. Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016.

[45] A. Chevallier, F. Cazals, and P. Fearnhead. Efficient computation of the the volume of a polytope in high-dimensions using Piecewise Deterministic Markov Processes. In *AISTATS*, 2022.

[46] S. Dasgupta and K. Sinha. Randomized partition trees for exact nearest neighbor search. *JMLR: Workshop and Conference Proceedings*, 30:1–21, 2013.

[47] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 537–546. ACM, 2008.

[48] J. Pevsner. *Bioinformatics and functional genomics*. John Wiley & Sons, 2015.

[49] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. Marks, C. Sander, R. Zecchina, J. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*, 108(49):E1293–E1301, 2011.

[50] Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt. Efficient generative modeling of protein sequences using simple autoregressive models. *Nature communications*, 12(1):5800, 2021.

[51] Nicholas Bhattacharya, Neil Thomas, Roshan Rao, Justas Dauparas, Peter K Koo, David Baker, Yun S Song, and Sergey Ovchinnikov. Single layers of attention suffice to predict protein contacts. *Biorxiv*, pages 2020–12, 2020.

[52] Francesco Caredda and Andrea Pagnani. Direct coupling analysis and the attention mechanism. *BMC bioinformatics*, 26(1):41, 2025.

[53] Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. Mapping of attention mechanisms to a generalized potts model. *Physical Review Research*, 6(2):023057, 2024.

[54] 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.

[55] F. Cazals and E. Sarti. Alphafold predictions on whole genomes at a glance. *Submitted*, 2025.

[56] D.M. Zuckerman. *Statistical Physics of Biomolecules: An Introduction*. CRC Press, 2010.

[57] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of statistical learning : data mining, inference and prediction*. Springer, 2001. htf-esldm-01.