

F. Cazals
Research Director, Geometrica group
INRIA Sophia-Antipolis, 2004 route des Lucioles
F-06902 Sophia-Antipolis, FRANCE
Frederic.Cazals@sophia.inria.fr

Algorithmic Problems in Computational Structural Biology

Understanding the lineage between species and the genetic drift of genes and genomes, apprehending the control and feed-back loops governing the behaviour of a cell, a tissue, an organ or a body, and inferring the relationship between the structure of biological (macro)-molecules and their functions are amongst the major challenges of modern biology. This course will focus on the third challenge, which actually subsumes two related problems: the way a protein adopts its 3D structure —folding, and the way two or several molecules assemble to form a complex —docking.

The emphasis will be on the algorithmic and mathematical techniques rather than on the investigation of specific macro-molecular systems. Yet, for all the algorithms covered, insights developed on such systems will be provided. No specific prior knowledge in bio-chemistry is required.

Each topic addressed will be covered during a lecture (2h), plus a paper reading session (2h).

1 Modeling Van der Waals models (2h+2h)

Van der Waals models, where an atom is represented by a ball whose radius depends on the chemical properties of the atom and its covalent environment are essential for a number of problems: defining molecular surfaces, modeling interfaces in protein - protein and protein - drug complexes, modeling the role of solvent, calibrating scoring functions etc. In this course, we shall review the fundamental algorithms to manipulate collections of balls, together with selected applications:

- Delaunay - Voronoi diagrams [BY98]
- The α -complex [Ede92, EM94]
- Applications to molecular surfaces and interfaces [Caz06a, CP⁺06]

2 Partial Shape Matching and structural similarities (2h+2h)

Given two bodies in $3d$, this course will address the Partial Shape Matching (PSM) problem: report the *maximal common parts* shared by these objects. First, in the combinatorial setting, where the objects are given as graphs, we shall see how PSM reduces to the calculation of Maximal Common Induced Subgraphs (MCIS) or Maximal Common Edge Subgraphs (MCES). Second, in the geometric setting, we shall examine two specific techniques based upon geometric hashing and dynamic programming.

- Combinatorial partial shape matching and problems on product graphs [Koc01, CK05a, CK05b]
- Molecular distance measures [Kav06]
- Geometric partial shape matching using hashing and dynamic programming [Che06, MGGP06]

3 Modeling protein flexibility I (2h+2h)

Proteins in vivo vibrate at various frequencies as high frequencies are found apart from chemical bonds, while low frequencies characterize more global deformations¹. Flexibility is actually central in modeling folding and docking. This course will review fundamentals of flexibility from the *kinematic* viewpoint (the study of movements independently from its causes), that is independently from energetic considerations.

- Kinematic chains and configuration spaces [Lat91, LMDL06]

¹See e.g. www.molmovdb.org for an inspection of macro-molecular motions.

4 Modeling protein flexibility II (2h+2h)

The energetic state of a protein and of its environment involves two terms: a *potential energy* term, and an *entropy* term directly related to the vibration properties. These two terms combine into the *free energy* of the system, which is the quantity used to state a system is at a minimum of (free) energy. This course will focus on the coupling between kinematic and energetic properties of molecules, so as to describe protein flexibility at two levels: the side-chain level, and the more global protein level.

- Side chains and rotamer libraries [Dun02]
- Force fields, potential and free energies [FS02]
- Modeling flexibility using normal modes [Hin04]

5 Computational topology I (2h+2h)

Given a function defined on a topological space, distinguishing the signal from the noise associated to the function is a central question arising in all sciences². While this question is classical in signal processing, its investigation underwent significant progresses in a more algorithmic setting, with the development of persistence theory. We shall present some important notions in this realm, together with two applications in molecular modeling.

- Notions of persistence theory [ELZ02, CSEH05]
- Application to the formation of micelles [KZP⁺]

6 Computational topology II (2h+2h)

Interestingly, persistence theory can also be approached from Morse theory, which consists of studying functions on manifolds, so as to encode (global) topological properties of the manifold. While Morse theory has essentially been developed in differential topology, an interesting incarnation of Morse concepts for collections of points or balls in $3d$ is provided by the flow complex. We shall review this construction, together with applications in molecular modeling.

- Morse theory and Morse-Smale diagrams [Mil63]
- The flow complex [GJ03, Caz06b]
- Application to the identification of pockets on macro-molecules [EFL98, CCL03]

References

- [BY98] Jean-Daniel Boissonnat and Mariette Yvinec. *Algorithmic Geometry*. Cambridge University Press, UK, 1998. Translated by Hervé Brönnimann.
- [Caz06a] F. Cazals. La morphologie des contacts entre bio-molécules. *Pour la Science*, Juillet - Août, 2006.
- [Caz06b] F. Cazals. Robust construction of the extended three-dimensional flow complex. Technical Report 5903, INRIA, 2006.
- [CCL03] F. Cazals, F. Chazal, and T. Lewiner. Molecular shape analysis based upon the morse-smale complex and the connolly function. In *ACM Symposium on Computational Geometry*, 2003.
- [Che06] P. Chew. Exact computation of protein structure similarity. In *ACM SoCG*, 2006.
- [CK05a] F. Cazals and C. Karande. An algorithm for reporting maximal c -cliques. *Theoretical Computer Science*, 349(3):484–490, 2005.
- [CK05b] F. Cazals and C. Karande. Reporting maximal cliques: new insights into an old problem. Rapport de recherche 5615, INRIA, 2005.
- [CP⁺06] F. Cazals, F. Proust, , R. Bahadur, and J. Janin. Revisiting the voronoi description of protein-protein interfaces. *Protein Science*, 15(9), 2006.
- [CSEH05] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *ACM Symp. Comp. Geometry*, 2005.
- [Dun02] R.L. Dunbrack. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*, 12(4):431–440, 2002.
- [Ede92] H. Edelsbrunner. Weighted alpha shapes. Technical Report UIUCDCS-R-92-1760, Dept. Comput. Sci., Univ. Illinois, Urbana, IL, 1992.

²In mathematics, see e.g. the lecture Fields lecture of T. Tao at the ICCM, 2006.

- [EFL98] H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.*, 88:83–102, 1998.
- [ELZ02] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.
- [EM94] H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graph.*, 13(1):43–72, January 1994.
- [FS02] D. Frenkel and B. Smit. *Understanding molecular simulation*. Academic Press, 2002.
- [GJ03] J. Giesen and M. John. The flow complex: A data structure for geometric modeling. In *ACM SODA*, 2003.
- [Hin04] K. Hinsin. Normal mode theory and harmonic potential approximations. EMBO Course Notes, 2004. www.pasteur.fr/recherche/unites/Binfs/EMBO2004/coursenotes/hinsen_norma%1_modes.pdf.
- [Kav06] L. Kavradi. Molecular distance measures. In *Course notes*. 2006. <http://cnx.org>.
- [Koc01] I. Koch. Fundamental study: Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Comp. Sc.*, 250(1-2):1–30, 2001.
- [KZP⁺] P. Kasson, A. Zomorodian, S. Park, N. Singhal, L. Guibas, and V. Pande. A new structural metric for membrane fusion: analysis of molecular dynamics data. *Submitted*.
- [Lat91] J.-C. Latombe. *Robot Motion Planning*. Kluwer Academic Publishers, Boston, 1991.
- [LMDL06] G. Liu, J. Milgram, A. Dhanik, and J.C. Latombe. On the inverse kinematics of a fragment of protein backbone. In *10th Symp. on Advances in Robot Kinematics*, Ljubljana, Slovenia, 2006.
- [MGGP06] N. Mitra, L. Guibas, J. Giesen, and M. Pauly. Probabilistic fingerprints for shapes. In *Symp. Geometry Processing*, 2006.
- [Mil63] John W. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963.