GRaphs and Algorithms for 3D proteIn structurE and dyNamics classification

GRADIENT Project

Advisor

Hamamache Kheddouci
U. Claude Bernard, Lyon 1, GOAL, LIRIS
hamamache.kheddouci@univ-lyon1.fr

Shape classification is one of the most important tasks in computer vision as demonstrated by the large body of work dealing with 3D shape analysis [20], [39]. Recent advances in 3D data acquisition as well as the availability of large 3D repositories have been instrumental in the design of new and more efficient algorithms for shape classification. 3D shapes may be represented by graphs and consequently, graph techniques may be strongly useful for their classification. In this project, we address the problem of 3D protein deformable shapes classification. Proteins are macromolecules characterized by deformable and complex shapes which are linked to their function making their classification an important task: namely for drug discovery and disease characterization. Protein shapes can be standardly and robustly generated from their high resolution 3D structures available in the Protein Data Bank (http://www.rcsb.org) or using AlphaFold2 [12]. Their conformational space can be sampled using molecular dynamics simulations [17]. In this project, proteins are assimilated to 3D dynamic deformable objects and their surfaces are represented by graphs, such as triangular tessellations or meshes. Generally, in the graph matching state of the art, only the static aspect is considered. This is due to a limited understanding of dynamic graphs. Since molecular dynamics can be used to efficiently sample the trajectories of molecular 3D objects, they constitute a perfect case of study for dynamic graph matching. The goal of our project is to propose and develop new fast and scalable methods based on dynamic graph matching and machine learning algorithms to address the problem of 3D protein deformable shapes classification. Our research hypotheses and questions can be summarized in the following categories:

1) **What are the most relevant data representations, local and global descriptors and features for 3D protein deformable structures and properties representations?** Our goal is to identify appropriate graph representations that take into account, in addition to protein shape, its physicochemical descriptors such as hydrophobic and electrostatic potential, then extract and compute a set of descriptors and features, allowing to develop powerful and robust graph techniques, including the definition of adequate graph distances and robust models, for 3D protein deformable structures and properties classification. Since the electrostatic potential is a field, hence continuous, special attention will be given to its attribution within the chosen graph representations.

2) **What are the most appropriate graph distances to compare 3D protein deformable shapes?** Our goal is to propose and develop new fast and scalable graph-based approaches, namely based on approximated graph edit distance (GED) techniques, similarity learning and/or hybrid (deterministic and learning) techniques to measure the distance between 3D protein deformable shapes. Besides its NP-completeness, subgraph matching's strict constraints are making it impractical for graph pattern matching in a dynamic context. As a result, and in order to avoid zero answers, relaxed graph pattern matching models have emerged namely graph simulation, strong simulation and strict simulation [18]. These new models offer answers in several cases, so they can approximate the graph edit distance in the dynamic context. In this project, we will study and propose graph simulation methods and their variants, and machine learning models, to compare 3D deformable shapes.

3) **What are the most adequate methods to build robust models to classify 3D protein deformable shapes with the highest performance?** We address this problem by the exploration and the proposal

of new fast and scalable graph methods combined with machine learning approaches, such as graph embedding, graph kernel techniques and GNN models. We will also investigate and propose new methods for features selection and dimensionality reduction. Indeed, some features may be either redundant or highly correlated. Furthermore, for very large datasets, such as the Protein Data Bank, scalability may become an issue. Finally, we will investigate existing and develop new machine learning models and algorithms for graph classification in this particular context taking into account the nature of protein data and their deformations and dynamics. In addition, we will incorporate expert knowledge in the processes of features selection, learning, classification and evaluation.

Research challenges (2 students)

Graph representations for 3D protein structures : The challenge in this task is to propose graph models which capture the structure and the physicochemical properties of the molecules and represent their dynamics

Graph based metrics for dynamic protein graph comparison: In this task, we will explore structural properties of graphs to propose significant metrics which best approximate the Graph Edit Distance - GED. To develop a set of graph techniques for comparing two graphs by using tools like graph decompositions to generate common edges or subgraphs of the two graphs. We will explore and propose new graph decomposition methods. Moreover, we should introduce new graph parameters (dominating sets with constraints) and new particular structures (covering subtrees or cycles) to compute the distance between graphs to compare

3D protein deformable shapes classification. We will explore the most adequate methods to build robust models to classify 3D protein deformable shapes with the highest performance.

Bibliography

[15]* F. Langenfeld et al. Shrec2020 track:Multi-domain protein shape retrieval challenge. Computers & Graphics, 91:189 – 198, 2020.

[16]* G. Levieux et al**.** FD169: UDock, the protein docking entertainment system. Faraday Discuss. 2014.

[17] A. Wang et al. “Large-Scale Biomolecular Conformational Transitions Explored by a Combined Elastic Network Model and Enhanced Sampling Molecular Dynamics”, The Journal of Physical Chemistry Letters, 2020.

[18] W. Fan et al. 2014. Graph simulation: Impossibility and possibility. VLDB Endowment 7, 12 (2014).

[19]* K. Madi et al. “New graph embedding approach for 3d protein shape classification,” IEEE IVPR 2020.

[20]* K. Madi et al. “New graph distance for deformable 3d objects recognition based on triangle-stars decomposition,” Pattern Recognition. 2019.

[21] J. Tangelder et al. “A survey of content based 3d shape retrieval methods,” Multimedia Tools Appl., 2008.

[22] J. Czajkowska et al. “Skeleton graph matching vs. maximum weight cliques aorta registration techniques,” Comp. Med. Imag. and Graph. 2015.

[23] V. Barra et al. “3d shape retrieval using kernels on extended reeb graphs,” Pattern Recognition, 2013.

[24] Y. Kleiman et al, “SHED: shape edit distance for finegrained shape similarity,” ACM Trans. Graph, 2015.

[25] D. Blumenthal et al. Upper Bounding Graph Edit Distance Based on Rings and Machine Learning. IJPRAI 2021.

[26] D. Conte et al. “Thirty years of graph matching in pattern recognition,” IJPRAI, vol. 18, no. 3, 2004.

[27]* K. Madi, et al. “New graph distance based on stable marriage formulation for deformable 3d objects recognition”, IEEE AICCSA 2019.

[28] B. Horst et al. “Towards the unification of structural and statistical pattern recognition,” PRL, 2012.

[29] P. Foggia et al. “Graph matching and learning in pattern recognition in the last 10 years,” IJPRAI, 2014.

[30] F. Escolano et al. “Information-geometric graph indexing from bags of partial node coverages”. 2011.

[31] S. Jouili and S. Tabbone, “Graph embedding using constant shift embedding,” pp. 83–92, 2010.

[32] X. Bai et al. “Learning invariant structure for object identification by using graph methods,” Comput. Vis. Image Underst., 2011.

[33] W. Lee et al. “Selecting structural base classifiers for graph-based multiple classifier systems,” 2010.

[34] E. Borzeshi et al. “Discriminative prototype selection methods for graph embedding” Pattern Recognit, 2013.

[35] Z. Wu et al., “A comprehensive survey on graph neural networks,” IEEE TNNLS, 2021.

[36] H. Dai et al., “Learning steady-states of iterative algorithms over graphs,” ICML 2018.

[37] Defferrard et al: Convolutional neural networks on graphs with fast localized spectral filtering. NeurIPS 2016.
[38] F. Monti et al. “Geometric deep learning on graphs and manifolds using mixture model cnns” CVPR 2017.
[39]* E. Paquet et al. “Deformable Protein Shape Classification based on Deep Learning, and the Fractional Fokker–Planck and Kahler–Dirac Equations”. IEEE TPAMI 2022.
[40] F. Manessi et al. Dynamic graph convolutional networks. Pattern Recognition. Vol 97, 2020.
[41] J. Zhang et al. “Gaan: Gated attention networks for learning on large and spatiotemporal graphs” UAI, 2018.
[42]* S. Berlemont et al. “Class-Balanced Siamese Neural Networks”. Neurocomputing, 2017
[43]* L. Zheng et al. Pairwise Identity Verification via Linear Concentrative Metric Learning. Trans. Cyber. 2018.
[44]* P. Compagnon et al. “Learning Personalized ADL Recognition Models from Few Raw Data”. AI in Med, 2020.
[45]* T. Jaffrelot-Inizan et al. High-Resolution Mining of SARS-CoV-2 Main Protease Conformational Space: Supercomputer-Driven Unsupervised Adaptive Sampling. *Chem. Sci.* 2021.
[46] S. Jouiliet al. “Graph embedding using constant shift embedding,” pp. 83–92, 2010.
[47] Ma, S. et al. Capturing topology in graph pattern matching. VLDB Endowment 5(4), 310–321 (2011).
[48]* S. Bouhenni et al. A Survey on Distributed Graph Pattern Matching in Massive Graphs. Comp. Surv. 2021.
[49] G. Ma et al. Deep graph similarity learning: a survey, In Data Mining and Knowledge Discovery, 2020.
[50] B. Wu et al. DGCNN: disordered graph convolutional neural network based on the gaussian mixture model. Neurocomputing, 2018
[51] Y. Bai et al. SimGNN: a neural network approach to fast graph similarity computation. ACM WSDM, 2019
[52] S. Liu et al: Simple unsupervised representation for graphs, with applications to molecules. NeurIPS. 2019
[53] Y. Li et al. Graph matching networks for learning the similarity of graph structured objects. ICML, 2019
[54] Du SS et al. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. NeurIPS, 2019
[55] B. Chen et al. “An algorithm for low-rank matrix factorization