Peter F. Stadler
Graphs in Molecular Biology (Leipzig)

1 Individual Project's contribution to the CRP

1.1 Aims and Objectives

Graphs are ubiquitous models in molecular biology at a variety of different levels of description, modeling nature at different resolutions and implementing relationships between the former types of models. Some of these representations are discrete approximations of geometric objects: molecular structures are seen as graphs whose vertices are either atoms (in models of chemistry) or more complex monomers (in coarse grained models of RNA and protein folding). Graphs and hypergraphs also describe the sequence and phenotype spaces on which evolutionary processes "live", as well as chemical reaction networks or gene regulatory networks describing mass and information flow in a cell. These graph models are intimately linked: the graphs representing of RNA and protein, for instance, are the nodes of the phenotype graph, while molecular graphs form the node set of chemical networks. Regulatory networks, on the other hand, link molecular graphs and the graphs of biopolymer structures. Aspects of (molecular) phylogenetics can also be understood in this setting as the reconstruction of trajectories from a small subset of "observed" vertices. Recent large-scale computer simulations [6] combine several of these discrete models and highlight that the individual models are strongly coupled. From a mathematical point of view, however, the intimate relationships between these layers of description are poorly understood and so far have not been investigated in any detail. **The Leipzig group proposes to explore systematically the connection between the geometric graph models of molecules and polymers on the one hand, and the graphs that represent the relationships among them.**

In the case of biopolymers (both RNA and protein), the processes of structure-formation (folding) and their evolution are naturally described a common graph Γ whose vertices are pairs (x, y) of genetic sequences x and structure graphs y. Only pairs of sequence and structure are admissible that are compatible: the vertices of y are in 1-1 correspondence with the sequence positions, and in the case of RNA, y can only have edges that conform to the rules of base-pairing in nucleic acids. An energy function $e(x, y)$, depending both on the graph structure y and the sequence x (which can be seen as a vertex labeling superimposed on y), determines implicitly which structures are likely to be formed from x. In line with physical reality, the $e(x, y)$ is given as the sum of contributions from individual molecular interactions, i.e., as a sum of contributions evaluating small labeled subgraphs of y. In the case of RNA, for example, the subgraphs are exactly the elements of the minimal cycle basis of y, in lattice models of proteins one typically uses edge weights. Edges in Γ are determined by small changes in either sequence (mutations) or structure, creating a locally product-like structure. The folding problem, thus, is formalized as $y^*(x) = argmin_y e(x, y)$, while evolutionary questions are often approximated as optimization problems on subgraph induced by $(x, y^*(x))$, with a fitness function $f(y^*(x))$. Empirically, both *combinatorial landscapes* (folding: $y \mapsto e(x, y)$ for a fixed sequence x; evolution: $x \mapsto f(argmin_y e(x, y)))$ and the so-called *genotype-phenotype* map $x \mapsto argmin_y e(x, y)$ have been studied in detail by the PI and other [12, 14, 17, 5]. The sequence design problem, i.e., the search for all sequences x that fold into a given structure z, formally $S(z) = \{x | z = \text{argmin}_y e(x, y)\}$, can be seen as just another aspect of the combined sequence-structure-energy landscape.

Here we propose to investigate systematically the consequences of the additive energy functions on the geometric structures of both landscapes, in particular w.r.t. geometric properties such as local optima, saddle points, valleys and basins. From the subgraph of Γ induced by the vertices of the form $(x, y^*(x))$ a phenotype graph Σ derived by interpreting the genotype-phenotype map as a (weak) graph homomorphism. Here we ask when Σ has local product structure since the "coordinate axes" of this product correspond to independently evolvable characteristics [16]. Furthermore, **we wish to understand how these local product structure is related to partitions or other, more complex relationships, on the underlying molecular structure graphs y. In particular, when to geometric**

domains correspond to independently evolvable modules, at least approximately?

The second field in which "intertwined" (hyper)graphs are of central interest is that of metabolism and its regulation. In metabolic networks, vertices are formed by molecular graphs representing small organic molecules. In the context of gene regulation, both small molecules and biopolymers (typically modeled as coarse grained graphs as outlined above) may appear in the same graph. Both chemical reactions and regulatory interactions are naturally modeled as directed hyperedges. Again there is an intimate link between the geometric graph models describing molecules and the network in which they are embedded. Conservation of atom type and bond order and the need to re-arrange chemical bond orders locally implies hyperedges/reactions are restricted to a fairly small subset of feasible transformations, and hence impose constraints on the local structure of chemical networks. Similar locality principles seem to govern regulatory interactions: similar proteins typically have similar functions and interaction partners, giving rise to similar local structures. Again, **we are interested in how these constraints can be understood formally.** Pathways, detours, and cycles are of particular interest in this context, hence the cycle structure of graphs and hypergraphs will be emphasized in this part of the project.

Regulatory network graphs themselves, finally, are subject to evolutionary changes as a consequence of evolutionary changes in the underlying molecules. Indeed, phenotype spaces whose vertices are e.g. Boolean networks have been discussed in the literature [4, 3]. Still, these are ultimately determined by the underlying molecule graphs at the lowest level. **Molecular evolution thus prompts us to develop a theoretical frame for hierarchically organized graphs of graphs of graphs and to understand in detail how lower levels constrain the higher ones as soon as higher-level edges connect entities that similar up to local perturbations.**

1.2 Methodologies

The two topical fields are connected by many methodological similarities and overlaps. Cycle bases underly the standard energy model of RNA secondary structures [10] and have been employed more recently also as a means of describing complex three-dimensional RNA structures [9, 13]. While bases are usually not unique, well-defined unique generating sets such as the relevant cycles [15] have been considered both for biopolymer structures [7] and as a means of characterizing molecular graphs [1]. Here, we will explore alternatives associated i.e., with (strictly) fundamental and robust cycles bases [8, 11] and consider the analogous problems for (directed) hypergraphs. Both abstract characterization of cycle sets and algorithmic approaches will be of interest. Since the (hyper)-edges of the "informational" higher-level graphs are derived from or implicitly determine (partial) homomorphisms of adjacent structural graphs, there is a particular interest in the behaviour of cycle bases and more general cycle sets under these maps. Thus we will in particular explore (generating) cycle sets that are well-behaved under small perturbations.

A second recurrent issue is that of approximate product-like structures, and more generally of locally homeomorphic copies that naturally arise at the higher level when transformation between lower-level graphs commute "most of the time". This is often the case e.g. in chemical reactions affecting two disjoint "functional groups", whence they typically can be exchanged in their order. This suggest a generalization of the fibers of product graphs (which are isomorphic copies of a factor) to images of different homomorphisms from the same factor. Graph bundles are one possible construction of this type. Here one asks for restricted weak homomorphisms that lead to a near product structure over an fiber F and a given base B such that the pre-images of edges of B are locally products. Much more can be done if one does not ask for isomorphic fibers. In general, one may consider partitions of the vertex set such the the subgraphs induced by the classes of a partition play the role of co-factors. Geometric notions, in particular convexity, are an indispensable tool in the context of product structures. In a more general setting, methods from information geometry may also come in handy to determine systems of partitions that are as product-like as possible. This in turn suggest to investigate relations R among graphs that are more general than graph homomorphisms: Interpreting graphs as relations on their vertex sets, one may ask for unique, minimal, maximal, or invertible solutions of $R^+ \circ E \circ R = E'$,

possibly with special rules for the diagonal elements.

The third aspect is the energy, cost, or fitness function defined on the vertex set of higher-level graphs that forms an integral part of the models, turning the graph of graphs into a combinatorial landscape. From this point of view, geometric properties such as local optima, various types of valleys, saddle points, and geodetic paths are naturally characteristics of practical importance, since they determine the behavior of dynamical systems defined on the landscapes, i.e., they govern the time course of evolutionary adaptation and molecular folding kinetics alike. The landscapes are characterized naturally in terms of their projection on the eigenspaces of the Laplace matrix of the underlying graph [12]. Here we specifically plan to explore the connection of geometric properties such as saddle heights and volumes of basins with the spectral analysis of the landscapes, which constrain the number of nodal domains [2], the values of local optima, and the value differences between neighbors.

References

[1] F. Berger, C. Flamm, P. M. Gleiss, J. Leydold, and P. F. Stadler. Counterexamples in chemical ring perception. *J. Chem. Inf. Comput. Sci.*, 44:323–331, 2004.

[2] T. Bıyıkoğlu, J. Leydold, and P. F. Stadler. *Laplacian Eigenvectors of Graphs: Perron-Frobenius and Faber-Krahn Type Theorems*, volume 1915 of *Lecture Notes in Mathematics*. Springer Verlag, Heidelberg, 2007.

[3] G. Boldhaus and K. Klemm. Regulatory networks and connected components of the neutral space: A look at functional islands. *Euro. J. Phys. B*, 2010. in press.

[4] S. Ciliberti, O. C. Martin, and A. Wagner. Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput Biol*, 3:e15, 2007.

[5] C. Flamm, B. M. R. Stadler, and P. F. Stadler. Saddles and barrier in landscapes of generalized search operators. In C. R. Stephens, M. Toussaint, D. Whitley, and P. F. Stadler, editors, *Foundations of Genetic Algortithms IX*, volume 4436 of *Lecture Notes Comp. Sci.*, pages 194–212, Berlin, Heidelberg, 2007. Springer. 9th International Workshop, FOGA 2007, Mexico City, Mexico, January 8-11, 2007.

[6] C. Flamm, A. Ullrich, H. Ekker, M. Mann, D. Högerl, M. Rohrschneider, S. Sauer, G. Scheuermann, K. Klemm, I. L. Hofacker, and P. F. Stadler. Evolution of metabolic networks: A computational framework. *J. Syst. Chem.*, 2010. in press.

[7] P. M. Gleiss, J. Leydold, and P. F. Stadler. Interchangeability of relevant cycles in graphs. *Elec. J. Comb.*, 7:R16 [16pages], 2000.

[8] K. Klemm and P. F. Stadler. Statistics of cycles in large networks. *Phys. Rev. E*, 73:025101, 2006.

[9] S. Lemieux and F. Major. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res.*, 34:2340–2346, 2006.

[10] J. Leydold and P. F. Stadler. Minimal cycle basis of outerplanar graphs. *Elec. J. Comb.*, 5:209–222 [R16: 14 p.], 1998. See `http://www.combinatorics.org/` R16 and Santa Fe Institute Preprint 98-01-011.

[11] J.-P. Ostermeier, M. Helmuth, K. Klemm, J. Leydold, and P. F. Stadler. A note on quasi-robust cycle bases. *Acta Math. Contemp.*, 2:231–240, 2009.

[12] C. M. Reidys and P. F. Stadler. Combinatorial landscapes. *SIAM Review*, 44:3–54, 2002.

[13] K. St-Onge, P. Thibault, S. Hamel, and F. Major. Modeling RNA tertiary structure motifs by graph-grammars. *Nucleic Acids Res.*, 35:1726–1736, 2007.

[14] P. F. Stadler and B. M. R. Stadler. Genotype phenotype maps. *Biological Theory*, 3:268–279, 2006.

[15] P. Vismara. Union of all the minimum cycle bases of a graph. *Electronic J. Comb.*, 4:#R9 (15 pages), 1997.

[16] G. Wagner and P. F. Stadler. Quasi-independence, homology and the unity of type: A topological theory of characters. *J. Theor. Biol.*, 220:505–527, 2003.

[17] M. T. Wolfinger, S. Will, I. L. Hofacker, R. Backofen, and P. F. Stadler. Exploring the lower part of discrete polymer model energy landscapes. *Europhys. Lett.*, 74:726–732, 2006.