A constructive approach for graph concepts with long range dependencies

Fabrizio Costa
Department of Computer Science
Albert-Ludwigs University Freiburg
Freiburg, 79085
costa@informatik.uni-freiburg.de

Stefan Mautner
Department of Computer Science
Albert-Ludwigs University Freiburg
Freiburg, 79085
mautner@cs.uni-freiburg.de

Abstract

Machine learning constructive approaches offer a way to answer interesting 'design' questions on the basis of a collection of examples. In particular, graph constructive methods are of interest in chemo- and bio-informatics domains where the task is to synthesize novel molecules with a desired bioactivity. Most molecules, however, exhibit complex dependencies between their different constituent parts. RNA polymers, for example, self interact, with nucleotides forming bounds that can typically span the entire length of the sequence. Since modeling long range dependencies is a difficult problem, we propose an efficient solution, based on graph coarsening that builds on top of a recent constructive approach and we show encouraging experimental results on a RNA synthesis task.

1 Introduction

Constructive machine learning addresses the problem of automatically 'design' artifacts given a concept expressed as a representative set of examples. The task becomes particularly challenging in a discrete setting when the solution space is exponential and direct enumeration approaches become unfeasible, as it is the case in the domain of graphs with discrete labels. In Costa [2016] the problem of generating elements of a structured domain was framed as the equivalent problem of sampling from a corresponding underlying probability distribution defined over a (learned) class of structures. Specifically they employ a context-sensitive grammar to accurately model complex dependencies between different parts of an instance. The authors acknowledge that an approach based exclusively on a grammar is not sufficient since the number of proposed graphs grows exponentially with the number of production rules in the grammar. To address the issue Costa [2016] proposes to use a Metropolis Hastings (MH) Markov Chain Monte Carlo (MCMC) method, where the problem of sampling is reduced to the easier task of *simulation*. They use the context sensitive graph grammar to inform the MH proposal distribution, but also introduce a probability density estimator to define the MH acceptance procedure. This allows to deal separately with local and global constraints: the locally context-sensitive graph grammar is used for the local constraints and the regularized statistical model is used for the global or long range constraints. The two approaches complement each other: the grammar is a flexible non-parametric approach that can model complex dependencies between vertices that are within a short path distance from each other; the statistical model, instead, can employ the bias derived from the particular functional family (linear) and the type of regularization used (a penalty over the ℓ_2 norm of the parameters) to generalize long range dependencies to similar cases. This approach is therefore adequate when the underlying concept exhibit local dependencies that are more complex than long range ones. Unfortunately in some application domains instances can exhibit complex long range dependencies between their different constituent parts. This is the case for RNA polymers, long sequences of atomic entities (nucleotides) that self interact, establishing pairwise bounds that can typically span the entire length of the sequence.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

A different way to view the issue of long range dependencies is that of the appropriate scale of representation. Certain application domains exhibit natural encodings, i.e. instances are encoded as graphs where nodes represent specific entities, such as nucleotides in the case of RNA sequences. However, it is known that a more effective functional description of RNA polymers can be obtained in terms of structural components such as *stems* (stretches of consecutive paired nucleotides) and *loops* (stretches of consecutive unpaired nucleotides). Under this view, dependencies that are local at the coarser scales correspond to longer range dependencies at the original scale.

A constructive system suitable for these domains needs to be able to adequately model complex long range dependencies and is more effective if it operates at a convenient coarser scale rather than that of the individual units. Here we tackle all these issues extending the approach presented in Costa [2016] with two key ideas: 1) we allow a user/domain defined graph coarsening procedure, and 2) we allow domain specific optimization procedures to ensure that generated instances are always viable.

2 Method

Costa [2016] presented an approach to sample graphs from an empirical probability distribution using a variant of the Metropolis Hastings (MH) algorithm [Metropolis et al., 1953]. The MH approach decomposes the transition probability of an underlying Markov process in two factors, 1) the proposal and 2) the acceptance-rejection probability distributions. The algorithm starts from a seed element and iteratively proposes new instances that are stochastically accepted if they fit appropriate membership requirements. The key element for an efficient MH design is building the proposal distribution in such a way that the generated elements will not be rejected too often. To do so Costa suggests to use a grammar and infer its rules from the available data using grammatical inference techniques upgraded to structured domains (i.e. domains where instances are naturally represented as labeled graphs).

More precisely, a *graph grammar* [Rozenberg and Ehrig, 1999] is a finite set of productions; a production is a triple $S = (M, D, E)$ where M is the "mother" graph, D is the "daughter" graph and E is an embedding mechanism. Given a "host" graph H, a production is applied as follows: 1) identify one occurrence of M as an isomorphic subgraph of H; 2) remove M from H and obtain H^-; 3) replace M by an isomorphic copy of D; and finally 4) use the embedding mechanism E to attach D to H^-. Costa [2016] introduced an efficient graph grammar based on the concept of *distributional semantics* [Harris, 1954, 1968] and on the *substitutability principle* [Clark and Eyraud, 2007]. Two key notions are defined: *internal core graphs* and *interface graphs*. An internal core graph (or *core* for simplicity), denoted $C^v_R(G)$, is a neighborhood graph of G of radius R rooted in v. An interface graph, denoted $I^v_{R,T}(G)$ is the difference graph of two neighborhood graphs of G with the same root in v and radius R and $R + T$. The difference graph is the graph induced by the difference of the two vertex sets. Intuitively an interface corresponds to the "external shell" of a core with *thickness* T. This shell represents the context available for the definition of a production rule. The general concepts of mother and daughter graphs are specialized as follows: given a production S, the mother graph M and the daughter graph D are union graphs of an interface graph and a correspondent internal core graph with the additional constraint that the interface graphs in the daughter and mother graph must be identical (up to isomorphism); finally, the embedding mechanism E is given by the isomorphic bijection between the interface graphs. In words, new elements are produced by *swapping inner cores* in identical contexts. In Fig. 1 (left) a production step is applied to a molecular graph G. The core $C^v_R(G)$ (dark) is determined by vertices at maximal distance R ($R = 0$ in the figure) from a chosen root vertex v. The interface $I^v_{R,T}(G)$ with *thickness* $T = 2$ is in a lighter shade. Given a new core-interface pair (CIP) with matching interface, the substitution can take place to yield the replacement of a carbon with a nitrogen atom. Note that this type of production rules can be inferred efficiently (i.e. in linear complexity) from a representative set of graphs (see [Costa, 2016] for details).

2.1 Method extension

Two main issues arise when dealing with complex structures such as RNA polymers, we call them 1) the *resolution* and 2) the *viability* problem. Replacing only few nodes at each iteration can be inefficient when instances consist of several hundreds of nodes. It is known that for RNAs a more effective description is obtained considering larger structural components such as stems and loops. As for the second issue, it is known that the function of RNA polymers depends on their global structure

(i.e. the set of pairs of interacting nucleotides), which can significantly change when even a single nucleotide is altered. To deal with these issues we propose two enhancements to Costa's approach: 1) a grammar coarsening procedure and 2) a constraint integration procedure.

Grammar coarsening. The idea is to allow users to specify a coarsening procedure in a simple way. We do so via the notion of *edge contraction*: an operation which removes an edge from a graph while simultaneously merging the two vertices that it previously joined. In addition we allow the more general notion of *vertex contraction*, which removes the restriction that the contraction must occur over vertices sharing an incident edge. Both procedures are defined using a node attribute $cid \in \mathbb{N}$ called *contraction-identifier* and contracting all vertices that share the same cid. We propose to use the contraction operation as a flexible way to transform a graph to its coarser version, $G \mapsto G'$. Cores and interfaces can now be defined exploiting the contracted graph. Starting from a CIP on the coarse graph, we define the core as the sub graph induced by the vertices of the original graph that have been contracted to vertices of the core of radius R in the coarse graph, $C^v_R(G', G)$. The new interface graph $I^v_{R,B}(G', G)$ is defined as the Cartesian product of the graph induced by the nodes adjacent to the core nodes in G at maximal distance B (the thickness in the base graph) and the interface graph on the coarse graph. In words, we require that both the interface at base level and at coarse level match for a core swap to take place. In Fig. 1 we depict a RNA polymer graph at nucleotide level (center) and its coarse version (right) where the contraction was informed by the notion of structural components such as stems, hairpin loops and multiloops. Note that a core at coarse level (in dark) made only of one node, corresponds to multiple nodes at base level. Moreover, while the interface at base level requires only the presence of few nucleotides, these have to belong to a context defined at the coarse level that can span a much larger fraction of the instance and can be viewed as a global localization indicator.

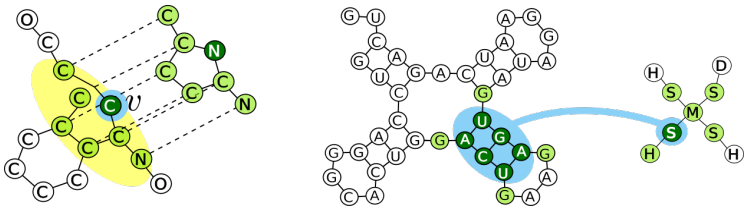

Figure 1: Core and interface subgraphs are marked in dark and light color respectively. Left: Molecular graph with indicated core substitution. Mid: RNA encoding with nucleotide level resolution. Right: RNA encoding with structural element resolution, with symbols: H)airpin, M)ultiloop, S)tem, D)angling end.

Constraint integration. It should be possible to easily make use of specific feasibility constraints when these are available for a given domain. For RNAs the relation between sequence and structure is known to be governed by thermodynamical forces that seek to minimize the amount of free energy of the molecule. The structure for a given sequence can be computed using sophisticated dynamic programming optimization algorithms. To integrate these constraints in the constructive protocol we let the user specify a transformation function which maps candidate instances to feasible instances. In our case the transformation takes a graph representing an RNA structure constructed by the proposed procedure, removes the pairwise bounds and recomputes them as the solution of a given folding algorithm. This transformation ensures that we are computing the acceptance probability always on viable RNA structures.

3 Empirical evaluation and discussion

RNA polymers cover essential biological roles ranging from coding, decoding, regulation and expression of genes. The RFam database [S. Griffiths-Jones, 2003] groups known RNA sequences in functionally and evolutionarily related *families*. To empirically investigate the performance of the proposed constructive approach we selected the SAM-I/IV variant riboswitch (RF01725) and SAM riboswitch (S box leader) (RF00162), which are families that exhibit a rich structure (e.g. that do not consist only of a single hairpin). The general aim is to synthesize functionally equivalent but novel sequences, a task with important biomedical applications.

When working with RNAs, the feasibility constraints are related to the structure determination. To improve accuracy we use a two step approach: given a candidate sequence we 1) identify k nearest neighbors among the available training instances, then 2) we align the $k+1$ sequences using the *MUSCLE* (MUltiple Sequence Comparison by Log- Expectation) program [Edgar, 2004] and compute the consensus structure using the *RNAalifold* program [Stephan H. Bernhart, 2008]. This procedure identifies the most representative structure in the ensemble of possible suboptimal solutions.

In order to evaluate if the constructed sequences are functionally equivalent to the original examples, we use an independent state-of-the-art computational approach as *oracle* [1]. To define the membership of a sequence to a given family, the RFam database uses the covariance model computed by the program *Infernal* (INFERence of RNA ALignment) [E. P. Nawrocki, 2013] induced over a hand curated set of representative sequences. In Fig. 2 we report the average score achieved by the constructed sequences as the training set increases. The horizontal line indicates the family specific threshold above which the Infernal model accepts a sequence as a valid instance of the family. In blue we report the results for the proposed extension and in red the results for the original method. To make it comparable to our method, we used MUSCLE and RNAalifold to train a new Infernal model and report the performance of its generated sequences (dashed). We note that the coarsening strategy consistently improves the quality of the results.

In Costa [2016] *the constructive learning problem for finite samples* is formally specified as the optimization of a parametric generator for two competing objectives: on the one hand one should obtain similar probability density estimators when these are trained on the original data or on the generated data; on the other hand the generated instances should be different from those already known in the training phase [2]. In Fig. 3 we evaluate these properties as the training set increases. We observe that the similarity tends to decrease as more diverse material becomes available and that also the divergence tends to vanish, indicating that the instances generated induce the same probability density as the original ones, albeit being increasingly different.

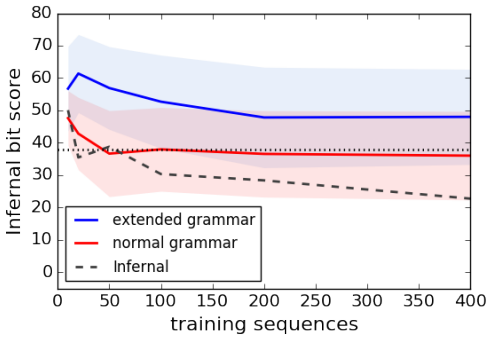

Figure 2: Estimated equivalence by Infernal.

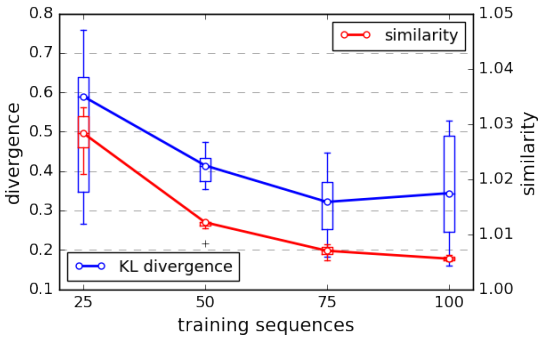

Figure 3: Functional vs structural similarity.

4 Conclusion

We have introduced a generic approach to tackle the problem of long range dependency modeling for a constructive machine learning approach. Preliminary results are encouraging and show that we can propose novel RNA sequences that are functionally equivalent to an original example sample. The coarsening procedure allows the injection of domain knowledge, significantly improving the quality of the results over the original non domain specific approach.

In future work we will investigate how to learn task specific coarsening schemes in a supervised fashion directly from data.

[1] The correct, and expensive, procedure would instead be to synthesize the sequences and test their functionality in a biological experiment.

[2] The notion of similarity between probability density estimators is defined in terms of Kullback-Leibler divergence and the notion of similarity between instances of a structured domain is defined in terms of a set graph kernel, see [Costa, 2016] for details.

References

Alexander Clark and Rémi Eyraud. Polynomial identification in the limit of substitutable context-free languages. *J. Mach. Learn. Res.*, 8:1725–1745, December 2007.

Fabrizio Costa. Learning an efficient constructive sampler for graphs. *Artif. Intell.*, 2016.

S. R. Eddy E. P. Nawrocki. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29: 2933–2935, 2013.

Robert C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792–1797, 2004.

Zellig Sabbettai Harris. Mathematical structures of language. 1968.

Z.S. Harris. Distributional structure. *Word*, 1954.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

Grzegorz Rozenberg and Hartmut Ehrig. *Handbook of graph grammars and computing by graph transformation*, volume 1. World Scientific, 1999.

M. Marshall S. Griffiths-Jones, A. Bateman. Rfam: an rna family database. *Nucleic Acids Res.*, 31: 439–441, 2003.

Ivo L. Hofacker Stephan H. Bernhart. Rnaalifold: improved consensus structure prediction for rna alignments. *BMC Bioinformatics*, 9:474, 2008.