How to measure the topological quality of protein parse trees?

Mateusz Pyzik, François Coste, Witold Dyrka

Proceedings of Machine Learning Research 93:118–138, 2018 International Conference on Grammatical Inference

How to measure the topological quality of protein parse trees?

Mateusz Pyzik MATEUSZ.PYZIK@PWR.EDU.PL
Politechnika Wrocławska, Wydział Podstawowych Problemów Techniki, Katedra Inż. Biomedycznej

François Coste FRANCOIS.COSTE@INRIA.FR
Univ Rennes, Inria, CNRS, IRISA, Rennes, France

Witold Dyrka WITOLD.DYRKA@PWR.EDU.PL
Politechnika Wrocławska, Wydział Podstawowych Problemów Techniki, Katedra Inż. Biomedycznej, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

Editors: Olgierd Unold, Witold Dyrka, and Wojciech Wieczorek

Abstract

Human readability and, consequently, interpretability is often considered a key advantage of grammatical descriptors. Beyond the natural language, this is also true in analyzing biological sequences of RNA, typically modeled by grammars of at least context-free level of expressiveness. However, in protein sequence analysis, the explanatory power of grammatical descriptors beyond regular has never been thoroughly assessed. Since the biological meaning of a protein molecule is directly related to its spatial structure, it is justified to expect that the parse tree of a protein sequence reflects the spatial structure of the protein. In this piece of research, we propose and assess quantitative measures for comparing topology of the parse tree of a context-free grammar with topology of the protein structure succinctly represented by a contact map. Our results are potentially interesting beyond its bioinformatic context wherever a reference matrix of dependencies between sequence constituents is available.

Keywords: context-free grammar, parse tree, contact map, molecular language

1. Introduction

Context-free (CF) and context-sensitive (CS) grammars are often regarded as more appropriate to model proteins than regular level models such as finite state automata and Hidden Markov Models (HMM). In theory, the claim is well-founded in the fact that many biologically relevant interactions between residues of protein sequences have a character of nested or crossed dependencies. In practice, there is hardly any evidence that grammars of higher expressiveness have an edge over old good profile HMMs (Eddy, 1998; Soeding, 2005) in classical applications including recognition and classification of protein sequences (Eddy, 2011; Remmert et al., 2012; Finn et al., 2015). This is in contrast to RNA modeling, where CFGs power some of the most successful tools (Sakakibara et al., 1993; Eddy and Durbin, 1994; Knudsen and Hein, 1999; Sükösd et al., 2012).

A few explanations of this phenomenon are plausible. On the biology side, one difficulty is that interactions in proteins are often less specific and more *collective* in comparison to RNA. On the modeling side, a difficulty is the larger alphabet which combined with

high complexity of CF and CS grammars imposes considerable trade-offs consisting on information reduction or learning sub-optimal solutions. Indeed, some studies hinted that CF level of expressiveness brought an added value in protein modeling when CF and regular grammars were implemented in the same framework (Dyrka, 2007; Dyrka et al., 2013).

Explanatory potential of grammatical modeling has not been fully used in studying proteins. A notable example is Protomata (Coste and Kerbellec, 2006), which was applied to create a database of motifs for phycobilin lyases (Bretaudeau et al., 2013). Several works attempted to build on explanatory power of grammars beyond regular. For example, derivations generated by mildly CSGs were proposed for prediction of secondary structures for protein sequences (Mamitsuka and Abe, 1994; Abe and Mamitsuka, 1997) and their mutual conformation in transmembrane bundles (Waldispuehl and Steyaert, 2005; Waldispuehl et al., 2006, 2008). Most likely parse trees derived using probabilistic CFGs were suggested to approximate the spatial structure of binding sites (Dyrka and Nebel, 2009), and to reveal expected amino acid properties at particular positions in helix-helix pairs (Dyrka et al., 2013). Trees generated with annotated stochastic CFGs were used to investigate relations between conserved regions in sequences (Sciacca et al., 2011). However, there have been no systematic study of explanatory power provided by various grammatical models.

The first step to this goal is defining objective criteria of such evaluation. Intuitively, a decent explanatory grammar should generate a topology, encoded in the parse tree, consistent with the topology of the protein, or its secondary and/or tertiary structure. In this piece of research we build on this intuition and propose a set of measures to compare the topology of the parse tree of a grammar with the topology of the protein structure represented by a contact map. The advantage of the parse tree as a grammatical descriptor is that it is a direct outcome of parsing; moreover, it can be obtained for the non-probabilistic and probabilistic grammars (typically the maximum-likelihood tree). The advantage of the contact map as the protein structure representation is its apparent simplicity but also availability of good quality predictions (Weigt et al., 2009; Wang et al., 2017), which can be potentially used in case the actual protein structure has not been solved.

2. Definitions

Context-free grammar. A *context-free grammar* is a quadruple $G = \langle \Sigma, V, v_0, R \rangle$, where Σ is a finite alphabet of terminal symbols (representing for instance amino acid species), V is a finite alphabet of non-terminal symbols (also called variables) disjoint from Σ, $v_0 \in V$ is a special start symbol and R is a finite set of rewriting rules such that $R \subseteq (V \to (\Sigma \cup V)^*)$, where $X \to Y = \{x \to y : x \in X \wedge y \in Y\}$ and $XY = \{xy : x \in X \wedge y \in Y\}$.

Chomsky Form with Contacts Consider a context-free grammar $G = \langle \Sigma, V, v_0, R \rangle$ satisfying $V = V_l \uplus V_s$ and $R = R_l \cup R_b \cup R_c$, where $R_l \subseteq (V_l \to \Sigma)$, $R_b \subseteq (V_s \to VV)$ and $R_c \subseteq (V_s \to V_l V_s V_l)$. The subsets R_l, R_b and R_c shall be referred to as *lexical*, *branching* and *contact* rules. Set $R_s = R_b \cup R_c$ is later referred to as *structural* rules. Grammars which satisfy these conditions are hereby defined to be in *Chomsky Form with Contacts* (CFC). When a CFC grammar satisfies $R_c = \emptyset$, it happens to be in Chomsky Normal Form (CNF).

Derivations and parse trees A rule $r = (A \to \alpha)$ derives ω_2 from ω_1, written $\omega_1 \overset{r}{\Rightarrow} \omega_2$, if and only if $\omega_1 = wA\beta$ and $\omega_2 = w\alpha\beta$ for some $w \in \Sigma^*, \beta \in (N \cup \Sigma)^*$. Note that only the

first non-terminal can be rewritten here, enabling to define a canonical (left-most) derivation by a string of rules: we say that a *derivation* $d = r_1 \dots r_l \in R^l$, $l \geq 0$ derives ω_l from ω_0, written $\omega_0 \overset{d}{\Rightarrow} \omega_l$, if and only if $\omega_0 \overset{r_1}{\Rightarrow} \omega_1 \dots \overset{r_l}{\Rightarrow} \omega_l$ for some $\omega_1, \dots, \omega_l \in (N \cup \Sigma)^*$.

A derivation d such that $v_0 \overset{d}{\Rightarrow} x$, $x \in \Sigma^*$ is called a *complete derivation* for G. We say then that d yields x, denoted $\text{yield}(d) = x$. A complete derivation d structures the sequence x by the associated *parse tree* $y(d)$, which is a labeled ordered tree whose root is labeled by v_0 and extended iteratively following the derivation rules: at each step of the derivation, by applying rule $r_i = (v \to \alpha_1 \dots \alpha_k)$ with $\alpha_1, \dots, \alpha_k \in (\Sigma \cup V)$, the parse tree is extended by adding edges from the existing left-most node labeled v to new ordered nodes labeled α_1 to α_k. As derivation is complete, the result is a (complete) parse tree whose nodes are labeled by variables and leaves by terminals.

Unlabeled derivation tree If we are interested only in the shape of the parse tree (called also the topology of the parse tree), we can consider the *unlabeled parse tree* or *skeleton* u obtained by removing the variables that label the nodes of the parse tree y, we write $u = u(y)$ (Sakakibara, 1992). Each (un)labeled tree has an associated $n \times n$ symmetric matrix $\mathsf{p} = (p_{i,j})$, where $p_{i,j}$ is the length of the path from i-th to j-th leaf.

Contact constraints Let $x = x_1 \dots x_n$ be a protein sequence. Most protein sequences fold into complex spatial structures. Let $\mathsf{d} = (d_{i,j})$ be a matrix of spatial distances between amino acids at positions i and j. The distance is typically calculated between C_α or C_β atoms from the main backbone (Wozniak and Kotulska, 2014). Two amino acids at positions i and j are said to be in contact if $d_{i,j}$ is below a given threshold τ (usually assumed to be 8Å (Wozniak and Kotulska, 2014)). A (complete) contact map for a protein of length n is a binary symmetric $n \times n$ matrix $\mathsf{m} = (m_{i,j})$ such that $m_{i,j} = 1 \iff d_{i,j} \leq \tau$. In practice, sometimes only a subset of the contact is determined. A partial contact map for a protein of length n is a binary symmetric matrix m such that $m_{i,j} = 1 \implies d_{i,j} \leq \tau$.

The *complement* of the contact matrix $\overline{\mathsf{m}} = (\overline{m}_{i,j})$ is defined such that $\overline{m}_{i,j} = 1 - m_{i,j}$.

Usually contacts between residues very close in the sequence do not carry much information. For a given separation in sequence s, let $\mathsf{m}^{(s)} = \left(m^{(s)}_{i,j}\right)$ be a *truncation* of contact matrix m such that $m^{(s)}_{i,j} = [|i-j| \geq s] \cdot m_{i,j}$, where $[\cdot]$ is the Iverson bracket.

Given a threshold δ, parse tree t is said to be *consistent* with a contact map m if and only if $m_{i,j} = 1 \implies p_{i,j} \leq \delta$.

A contact map is *non-crossing* when no contact has one end inside and another outside of interval induced by any other contact. In other words, pairs can stand side by side or be nested one in another.

$$(m_{i,j} = 1 \wedge m_{k,l} = 1 \wedge i \leq k \wedge i < j \wedge k < l) \implies (i < k \wedge (j < k \vee l < j))$$

For example, contacts $(1,6)$, $(2,5)$ form a non-crossing contact map, while contacts $(1,6)$ and $(2,7)$ are crossing. Interestingly, fixing threshold $\delta = 4$, contact map m is non-crossing if and only if there exists a grammar G in CFC and a parse tree w.r.t. G which is consistent with map m.

3. Measures

The most straightforward approach to assess descriptive performance is to use the unlabeled derivation tree as a predictor of spatial contacts between positions in sequence, parameterized by the cutoff δ on path length between the leaves. Pairs of terminal symbols linked by paths shorter than δ are considered as *true (false) positives* if actually (not) in contact, while pairs of terminal symbols linked by longer paths are considered as *false (true) negatives* if actually (not) in contact.

Basic measures The essential measures of prediction performance are the *recall* (called also *sensitivity*), which is the fraction of correctly predicted objects which actually belongs to the class of interest, and the *precision* (called also *positive predictive value*), which is the fraction of correctly predicted objects among all predicted to belong to the class of interest. Indeed, they were used, for example, to assess small grammar designs for RNA secondary structure prediction (Dowell and Eddy, 2004). The harmonic mean of recall and precision is usually denoted as the F_1 measure, also known as *Czekanowski binary index* (Czekanowski, 1909), *Dice index* (Dice, 1945) and *Sørensen index* (Sørensen, 1948). In the setting of the contact prediction, the measure can be seen as the overlap between a subset of residues close in a protein structure and a subset of residues close in a parse tree. All three measures are parameterized by the fixed cutoff δ.

Aggregated measures While choice of δ can be obvious for some grammars such as a classical probabilistic CFG for RNA modeling (Knudsen and Hein, 1999), it is not evident in general. A possible workaround consists in calculating an aggregated measure, such as the *area under precision-recall curve* (AUPRC) defined by a sequence of recall-precision points over varying thresholds δ. In practice AUPRC is often approximated by the *average precision* calculated over the thresholds (AP). Another popular aggregated measure of performance of a classifier is the *area under the receiver operating characteristic* (AUROC) calculated for true positive rates achieved over the thresholds of false positive rate (Fawcett, 2006). The threshold-independent measures can be used to evaluate the overall topology (global features of the shape) of the parse tree.

Mean path measures In addition, we propose two novel measures that relate the mean path length in the parse tree between residues in contact $\mathfrak{p}\left(\mathsf{m}^{(s)}\right)$ to the mean path length between residues *not* in contact $\mathfrak{p}\left(\overline{\mathsf{m}}^{(s)}\right)$, where m is the reference contact map, $\mathfrak{p}(\mathsf{a}) = \langle \mathsf{p}, \mathsf{a} \rangle / \|\mathsf{a}\|_1$, $\langle \cdot, \cdot \rangle$ stands for the scalar product and $\|\cdot\|_1$ denotes the ℓ_1 norm. Then, the overall fitness of the parse tree w.r.t. a given contact map m can be calculated as:

$$S_1 = \frac{\mathfrak{p}\left(\overline{\mathsf{m}}^{(s)}\right) - \mathfrak{p}\left(\mathsf{m}^{(s)}\right)}{\max\left\{\mathfrak{p}\left(\mathsf{m}^{(s)}\right), \mathfrak{p}\left(\overline{\mathsf{m}}^{(s)}\right)\right\}}, \quad \text{and} \quad R_1 = \frac{\mathfrak{p}\left(\overline{\mathsf{m}}^{(s)}\right)}{\mathfrak{p}\left(\mathsf{m}^{(s)}\right)},$$

which are respectively the normalized difference, and the ratio of average path lengths between residues in contact and those *not* in contact. S_1 gives scores from -1 (*bad*) to 1 (*good*), while R_1 ranges from 0 (*bad*) to ∞ (*good*). Both measures indicate how much paths to residues in contact are shorter than paths to residues *not* in contact.

Measures defined above are general enough to cope with weighted contact maps where each pair has a score or probability – input matrix of contacts m can be real-valued. Definition of distances p can be changed as well – variants of measure S_1 for weighted grammars

can be defined similarly, for example, by setting distance $p_{i,j}$ to be the sum of negated logarithms of edge weights between i and j.

Local measures w.r.t. residue Local variants of all aforementioned measures can be defined. Intuition behind them is to focus only on residues that are in contact with k-th residue. This effect can be obtained by using only respective row of the contact map $\mathsf{m}_{k,\cdot}$ when calculating the value of a measure for amino acid at position k.

Implementation All measures and simulations were implemented in python 3.5 (van Rossum and de Boer, 1991) with some help of scikit-learn (Pedregosa et al., 2011). Charts were plotted with matplotlib (Hunter, 2007) and parse trees were plotted with dot (Gansner and North, 2000).

4. Data

Simulated data Properties of the proposed measures were first evaluated using a large set of simulated data. It was assumed that parse trees were generated using grammars in the CFC in the form $v_0 \to v_0v_0|v_0v_1|v_1v_0|v_1v_1|v_1v_0v_1$ and $v_1 \to \bullet$, where $\bullet$ denotes a terminal symbol (irrelevant for the shape of the tree). Therefore, the minimum separation between residues in contact s was set to 3 (two residues in between), and the minimal achievable distance between terminals in the tree was 4. Having in mind the goal of assessing topological quality of protein grammars and basing on data from our previous research (Dyrka and Nebel, 2009; Daskalov et al., 2015), we assumed the sequence length of 30 residues with 6 or 10 non-crossing contacts.

To build the set, first, one hundred non-crossing maps were generated, and parse trees that matched them perfectly were constructed. Note that ambiguities in generating the unlabeled parse tree with regard to the non-crossing contact map can occur for stretches of non-pairing residues. In such cases, it was arbitrarily chosen to prefer $v_0 \to v_0v_1$ over $v_0 \to v_1v_0$ rules. Thus, within each non-pairing stretch of residues, the rightmost leaves were closer to the root of the parse tree. Then contact maps were randomly modified such that from one up to 2/3 of contacts were changed either by deletion, addition, or replacement (combination of deletion and addition). Additions were performed either in non-crossing or in unconstrained manner. The process was repeated 10 times for each setup (initial number of contacts, type and number of modifications). The modified maps were then treated as a reference when computing measures against the parse tree reflecting the original map. This reversal of the natural scenario of modifying parse trees was applied for the sake of simpler generation process.

Generation of non-crossing contact maps was achieved by subdividing the initial interval into smaller ones. Each newly chosen contact pair partitions space into left-hand, right-hand and inner part. To pick a contact, it suffices to choose an interval among available and then its endpoints from the interval that was picked. The procedure yields uniform distribution of non-crossing contact maps.

This simple procedure cannot be used for extending an existing non-crossing contact map, since in this case new contacts are not necessarily added in the hierarchical nested order. Thus, we apply a constrain-and-generate approach with a blacklist of choices that failed to produce valid result (desired number of contacts) and a stack of choices that

can be undone if contradiction appears. Backtracking guaranteed completeness and the blacklist ensured termination of the procedure. A set of allowed endpoints was maintained for each position in the sequence. The sets were adjusted at every random choice of a pair in contact, effectively pruning the search space. This method was sufficiently efficient, as we were mostly interested in maps which were not large and not dense enough to make such the search procedure prohibitively time-consuming.

RNA Qualitative analysis of the proposed measured was conducted using an example RNA sequence taken from Fig. 2 in (Dowell and Eddy, 2004). The sequence folds into a three-stem structure stabilized by interactions exhibiting only nesting and branching dependencies. A set of parse trees compatible with the CFC was used to model various defects in representing the real RNA structure defined. Specifically, the grammar underlying all the parse trees was in the form $v_0 \to v_0v_0|v_0v_1|v_1v_0|v_1v_1|v_1v_0v_1$ and $v_1 \to \mathsf{A}|\mathsf{C}|\mathsf{G}|\mathsf{U}$, where A, C, G and U represented the four bases of the RNA sequence. Scores of the consistency measures were calculated for the minimum separation of 3.

Proteins Eventually, the proposed measures were evaluated on a set of the maximum likelihood parse trees generated for three samples of protein sequences. The parse trees were obtained using probabilistic CFGs with rule probabilities automatically inferred solely from positive training sequences using a genetic algorithm-based framework (Dyrka and Nebel, 2009; Dyrka et al., 2013), as described in our recent work (Dyrka et al., 2018). Two variants of grammars were used: one with rules in CFC, and the other with the subset of rules conforming to CNF (no contact rules). The rule sets included all possible productions which can be generated for 3 lexical and 4 structural non-terminals in these forms ($|R_l| = 60, |R_b| = 196, |R_c| = 144$). The setup of the genetic algorithm is described in (Dyrka et al., 2018) and is similar to (Dyrka and Nebel, 2009). The processing was carried using a variant of the 8–fold Cross–Validation scheme in which 6 parts were used for training, 1 part was used for validation and parameter selection, and 1 part was used for final testing (the scheme resulted in 56 runs for each sample) (Dyrka et al., 2018). Within each sample, all sequences shared the same length, and for each sample, one experimentally solved spatial structure from the Protein Data Bank (PDB) (Berman et al., 2000) was selected as the representative. The samples were made non-redundant at the level of sequence similarity around 70%. They consisted of:

- *CaMn*: 24 sequences of a Calcium and Manganese binding site from the legume lectins (Sharon and Lis, 1990) collected according to the PROSITE PS00307 pattern (Sigrist et al., 2013) true positive and false negative hits, but extended to 27 residues to cover the entire binding site. The motif folds into a stem-like structure with 41 contacts, many of them forming nested dependencies (Fig. 7a based on pdb:2zbj (de Oliveira et al., 2008));
- *NAP*: 64 sequences of the Nicotinamide Adenine dinucleotide Phosphate binding site fragment from an aldo/keto reductase family (Bohren et al., 1989) collected according to the PS00063 pattern true positive and false negative hits (four least consistent sequences were excluded). The motif of length of 16 amino acids covers only a part of the binding site of the relatively large ligand. Eleven contacts within the motif seem

to be insufficient for fully defining the fold, which depends also on interactions with amino acids outside the motif;

- *HET-s*: 160 sequences of the HET-s-related motifs r1 and r2 involved in the prion-like signal transduction in fungi (Daskalov et al., 2015). The largest subset of motifs with length of 21 amino acids was used to avoid length effects on grammar scores. The beta-hairpin-like fold of the motif partially relies on interactions between hydrophobic amino acids at positions 5, 8, 10, 14, 16 and 18. There is also strong dependency between positions 17 and 21 (Daskalov et al., 2015). All 10 intra-motif contacts are shown in Fig. 9a, which is based on pdb:2kj3 (van Melckebeke et al., 2010).

5. Results

Simulated data Results of evaluation on simulated data are shown in Fig. 1, 2, and 3.

Values of the basic prediction-related measures such as precision, recall and F_1 score were completely determined by experiment parameters and as such they exhibited the single-point distribution (given the threshold $\delta = 4$). For F_1, it may be noticed that adding a pairing not present in the contact map to the parse tree (resulting in a false positive) and removing a pairing present in the contact map from the parse tree (resulting in a false negative) do not count equally. Indeed, increasing number of false positive pairings (which affects only precision) counts relatively less than increasing number of false negative pairings (which affects only recall).

Among the aggregated measures, both AP and AUPRC were quite sensitive to alterations: distributions for different levels of discrepancy between the contact map and the parse tree were virtually non-overlapping and had a narrow span. Out of these two, distribution of the average precision appeared to be more tight. In both cases, the probability masses tended to concentrate towards the lower bound at each level of discrepancy. In addition, for experiments with only false positives the measures had the single-point distribution as the value was determined by the level of discrepancy. Moreover, AP achieved consistently lower values than AUPRC when false positives were considered.

Characteristics of S_1, R_1 and AUROC were more complex. For experiments with false negatives, these measures behaved similarly to AP and AUPRC. Their mean value decreased with increasing discrepancy, however, variance was high so distributions for neighboring levels of alterations overlapped, yet were still discernible when comparing with more distant ones (see also next paragraph). For experiments with false positives alone, the measures only very slightly decreased with increasing number of added false positive pairings. Moreover, AUROC had the single-point distribution as the value was determined by the level of discrepancy. Values R_1 (S_1) for parse trees perfectly matching their respective contact maps ranged from 2.4 to about 3.0 (from 0.58 to 0.67).

RNA Results of evaluation of perfect and corrupted parse trees of an RNA molecule are shown in Tab. 1. Selected unlabeled parse trees are shown in Fig. 4 and 5. For the sake of clarity, edges representing lexical rules were removed.

The perfect tree (Fig. 4a) was one of the possible trees ideally representing contacts found in the RNA molecule. It achieved the perfect score of 1.0 for the thresholded measures (at $\delta = 4$), AP, AUPRC and AUROC, and set the reference level for R_1 and S_1 scores.

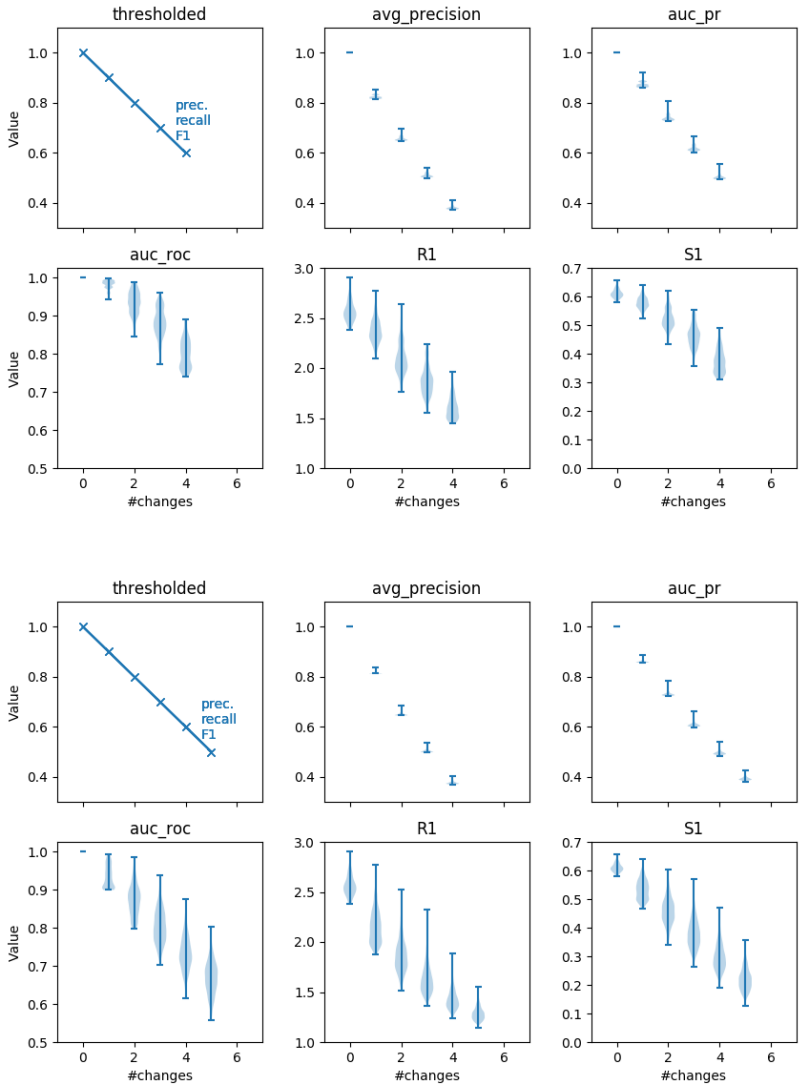


Figure 1: Parse tree quality measures in function of number of alterations for simulated maps with 10 contacts and the replacement operation (false positive and false negative pairings); non-crossing (up) or unconstrained (down).

The first test consisted on removing the four pairings making the outer stem (2-39, 3-38, 4-36, 5-35). Two unlabeled parse trees were created that accounted for this defect: one preserving the overall topology of the original tree (Fig. 4b) and one breaking the topology (Fig. 4c). The thresholded measures, which are based on counts of bases generated by the contact rules, evaluated both corrupted trees equally: the four false negatives resulted in the recall of 0.64 and F_1 of 0.78. The aggregated measures penalized the tree with broken topology more than the tree with preserved topology. Moreover, the latter was practically indistinguishable from the perfect tree in terms of AUROC, R_1 and S_1. Relatively large

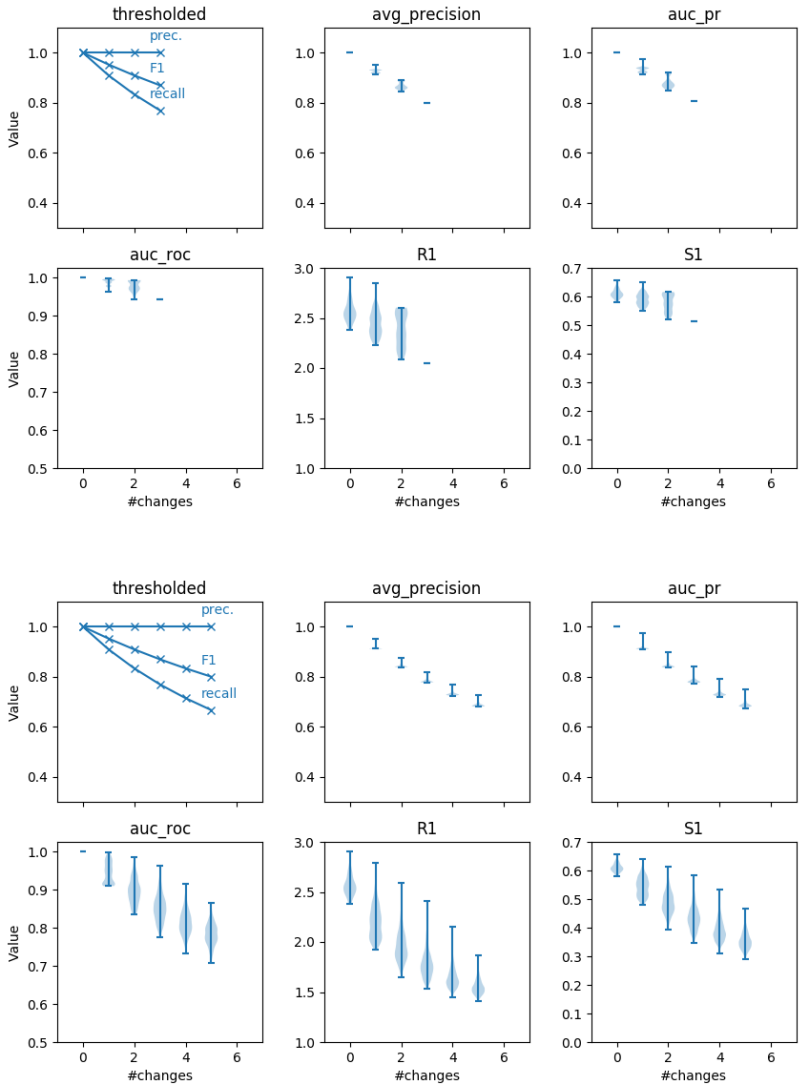


Figure 2: Parse tree quality measures in function of number of alterations for simulated maps with 10 contacts and the deletion operation (only false negative pairings); non-crossing (up) or unconstrained (down).

difference between R_1, S_1 scores for the two corrupted trees amounting to 50-55% of the reference was consistent with wide distributions of these measures observed for the simulated data (Fig. 1, 2, and 3).

The next test consisted on removing or adding three pairings in different ways. In the first attempt, all three pairings making the left-hand-side stem (7-15, 8-14, 9-13) were taken out. While the unlabeled parse tree created to represent this defect (Fig. 5a) broke the topology, the effect was visually weaker than previously. This is well reflected in intermediate values of virtually all aggregated measures, except AP, which is virtually equal

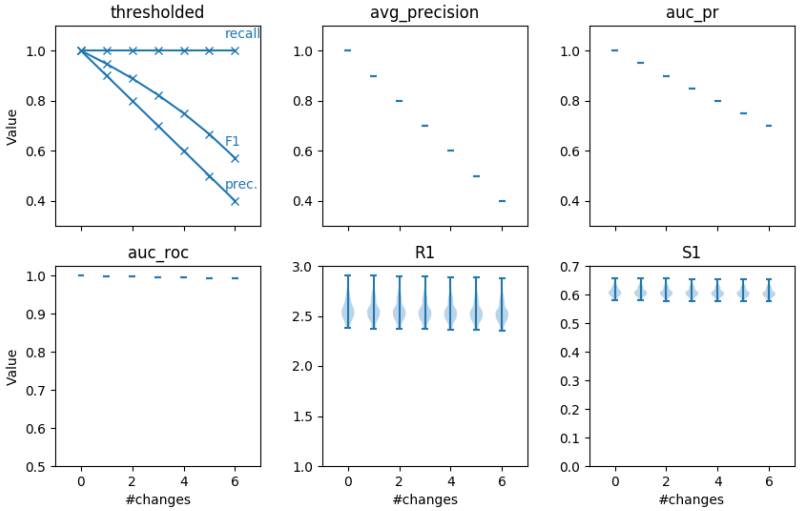


Figure 3: Parse tree quality measures in function of number of alterations for simulated maps with 10 contacts and the addition operation (only false positive pairings).

Table 1: Values of the evaluated measures for perfect and corrupted parse trees of a sample RNA molecule from Dowell and Eddy (2004).

Tree / defect	Fig.	F_1	Prec.	Recall	AP	AUPRC	AUROC	R_1	S_1
Perfect	4a	1.00	1.00	1.00	1.00	1.00	1.00	3.2	0.69
Missing outer pairings topology preserved	4b	0.78	1.00	0.64	0.76	0.88	0.99	3.2	0.69
Missing outer pairings topology broken	4c	0.78	1.00	0.64	0.64	0.64	0.76	1.6	0.38
Missing all pairings in a stem	5a	0.81	1.00	0.73	0.74	0.74	0.92	2.4	0.59
Missing a pairing in each stem	5b	0.81	1.00	0.73	0.80	0.84	0.99	3.0	0.67
Additional pairings in a stem	5c	0.88	0.79	1.00	0.79	0.89	0.998	2.9	0.66
Unfolded strand	–	0.00	0.00	0.00	0.02	0.23	0.48	0.9	-0.11
All residues in one stem	–	0.13	0.11	0.18	0.05	0.13	0.65	1.3	0.22

to that of the tree of preserved topology from the previous experiment. Subsequently, a single pairing was removed from each of the three stems (3-38, 8-14, 22-29), which led to the unlabeled parse tree topology only slightly diverging from the origin (Fig. 5b). Again, this topology-preserving discrepancy only marginally affected AUROC, and was within 90%

of R_1 and S_1 scores of the perfect tree. Quite similar was the effect of adding three false positive pairings (18-34, 21-30, 24-27) into the right-hand-side stem (Fig. 5c). Obviously, in this case the recall remained unaffected, while the precision dropped to 0.79 resulting in F_1 of 0.88. Finally, the measures were applied to trees intentionally unrelated to the real structure of the example RNA molecule. The tree made with a chain of rules producing a single terminal at each step, which represented the trivial topology of unfolded strand, achieved zero precision and recall, and slightly less than random scores or AUROC, R_1 and S_1. Eventually, the tree made with a chain of contact rules, which represented another trivial topology of a single long stem, mostly achieved slightly higher scores because of the two true positive contacts of the outer stem (2-39, 3-38). Interestingly, AUPRC was higher for the tree with unfolded topology (0.23) than for the tree with the single stem topology (0.13), see Discussion.

Proteins Distributions of values of the evaluated measures on protein data are shown in Fig. 6, and the average values are presented in Tab. 2. Sample parse trees are shown in Fig. 7, 8 and 9 with residues annotated with *local* AP calculated for contacts of each residue separately. The aim of this kind of presentation is to highlight residues correctly/incorrectly positioned in the tree according to the contact map.

Table 2

Dataset/grammar	Fig.	F_1	Prec.	Recall	AP	AUPRC	AUROC	R_1	S_1
CaMn/CFC	7d,7c	0.30	0.69	0.19	0.53	0.61	0.89	1.73	0.42
CaMn/CNF	8c,8b	n/a	n/a	n/a	0.24	0.38	0.65	1.35	0.21
NAP/CFC	–	0.11	0.14	0.09	0.12	0.11	0.44	0.97	-0.03
NAP/CNF	–	n/a	n/a	n/a	0.16	0.35	0.56	1.08	0.06
HET-s/CFC	9b	0.12	0.13	0.12	0.14	0.31	0.77	1.39	0.28
HET-s/CNF	9c	n/a	n/a	n/a	0.08	0.27	0.51	1.03	0.01

The results show considerable variation between samples and grammar forms. Grammar-wisely, the CFC grammars typically obtained higher scores than the CNF grammars for CaMn and HET-s datasets, apparently because the contact rules are particularly suited to model the stem-like structures of these protein fragments (Fig. 7a, 9a).

Sample-wisely, mean scores of the measures were highest for the CaMn dataset. Yet, even in the best scoring cases (CaMn/CFC at Fig. 7d and CaMn/CNF at Fig. 8c), values of the measures were only at the level of the RNA molecule with seriously corrupted topology due to the missing outer pairings (cf. Tab. 1 and Fig. 4c). Indeed, even though the best trees for CaMn correctly represented the main stem-like structure of the binding site, they could not represent many crossing interactions, e.g. between positions 17 and 21 (Fig. 7a). While in the CaMn case the highest scores were similar for the CFC and CNF trees, distributions were shifted towards higher values for CFC in comparison to CNF, especially for AP and the AUC measures (compare middle rank trees for CFC at Fig. 7c and CNF at Fig 8b). The worst scoring trees for CaMn had virtually no resemblance to the actual structure of the fold (Fig 7b and 8a).

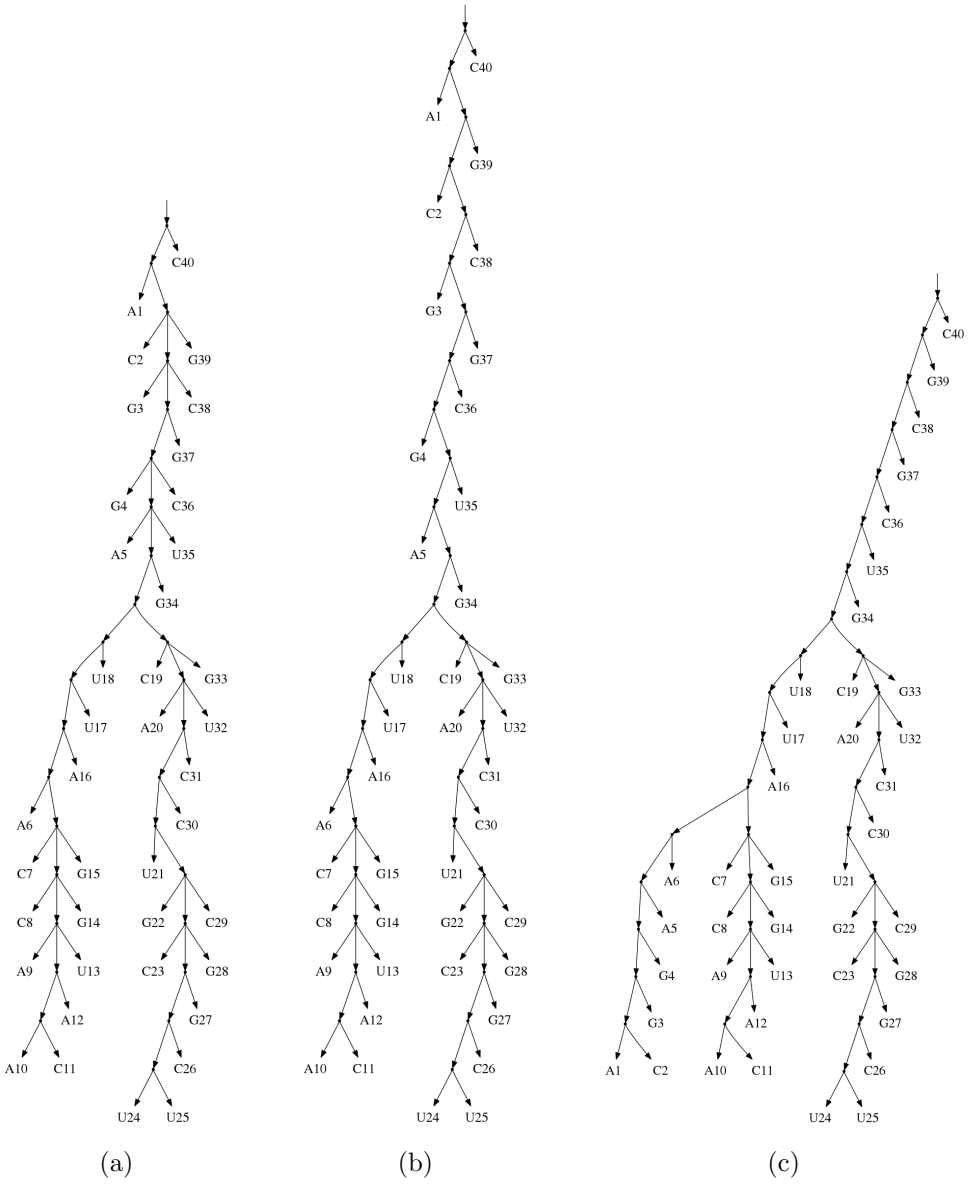


Figure 4: Unlabeled parse trees of an RNA sequence: (a) fully covering the reference contact map, (b,c) without four outermost pairings. The overall topology of the tree is preserved in (b) but broken in (c) for positions 1-6.

Parse trees generated for NAP were on average scored close to the theoretical or experimentally determined random levels (cf. Tab. 1). Similar was the case of parse trees for HET-s generated with the CNF grammars. Interestingly, the CFC parse trees for HET-s achieved relatively low scores of F_1, precision, recall, AP and AUPRC, but moderate scores of AUROC, R_1 and S_1. For example, the best scoring (in terms of AP) CFC tree for HET-s had a correct overall topology, which led to a considerable S_1 of 0.39, but it missed out

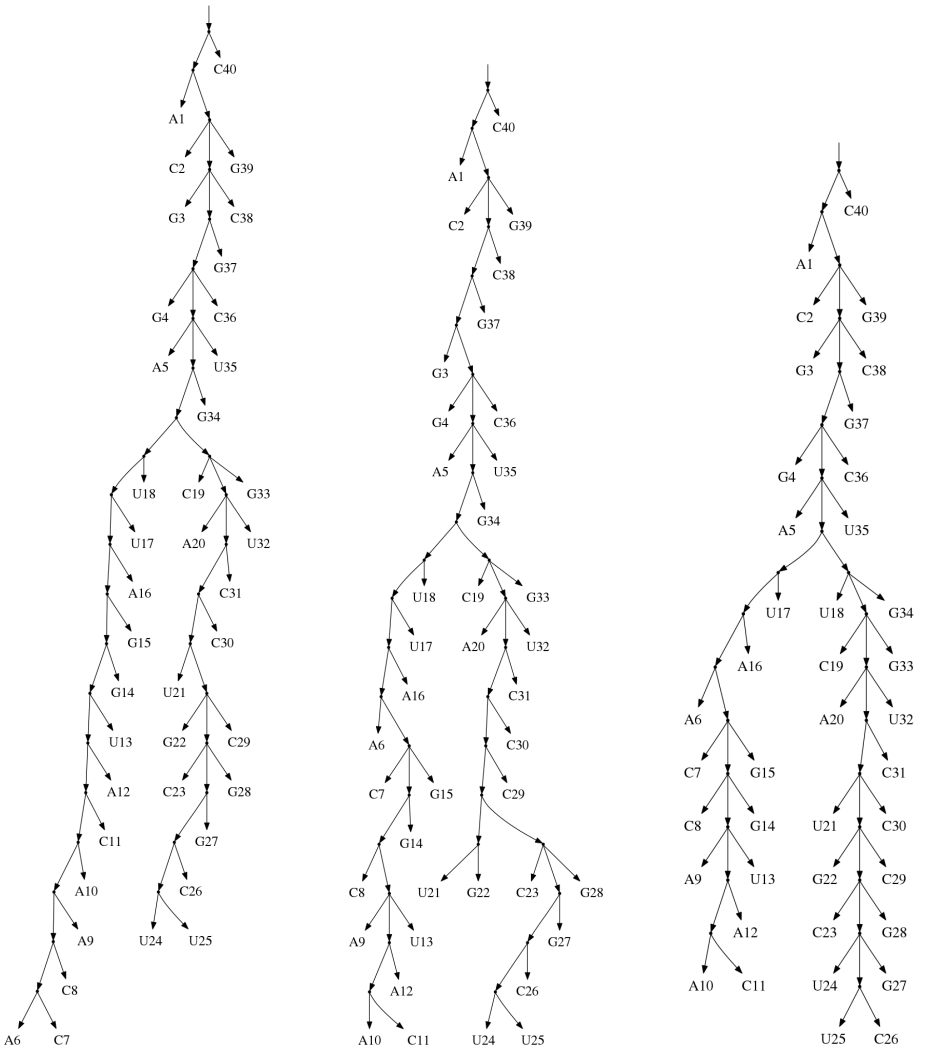


Figure 5: Unlabeled parse trees of an RNA sequence: (a) without all three contacts of the left-hand-side stem, (b) without one contact from each of the three stems, (c) with three additional (false positive) contacts in the right-hand-side stem.

some of the contacts overlapping those involved in the main stem (most notably between positions 17 and 21), and incorrectly added some false contacts (for example between positions 2 and 20), which led to AP as low as 0.31 (Fig. 9b). This inaccuracy in representing contacts despite generally correct topology resulted in AP not much higher than in the case of the best CNF tree with apparently corrupted topology (Fig. 9b).

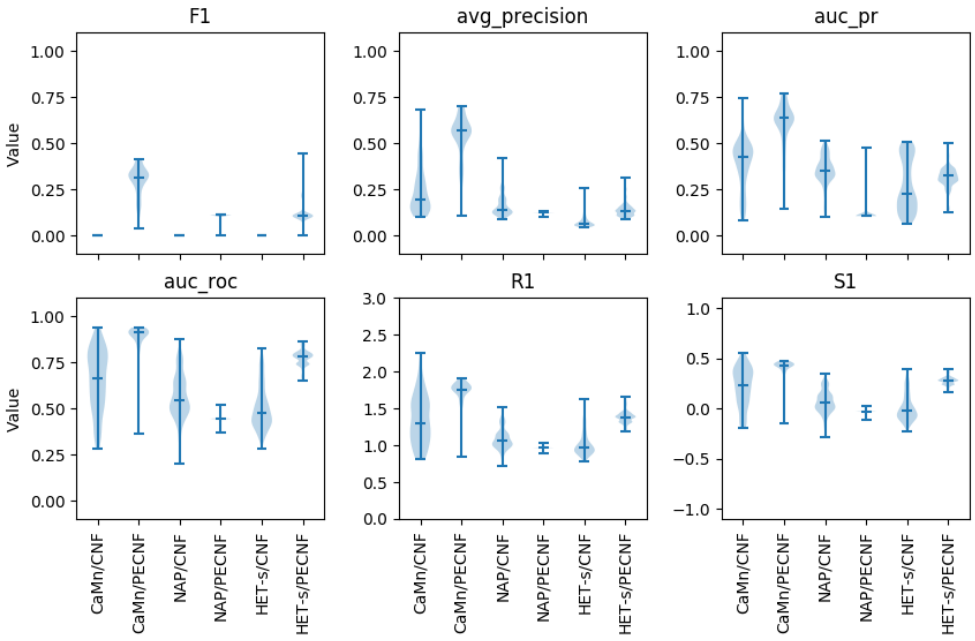


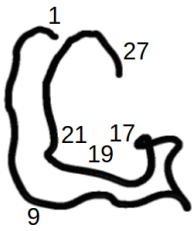
Figure 6: Distribution of scores of the evaluated measures on protein samples.

6. Discussion and conclusions

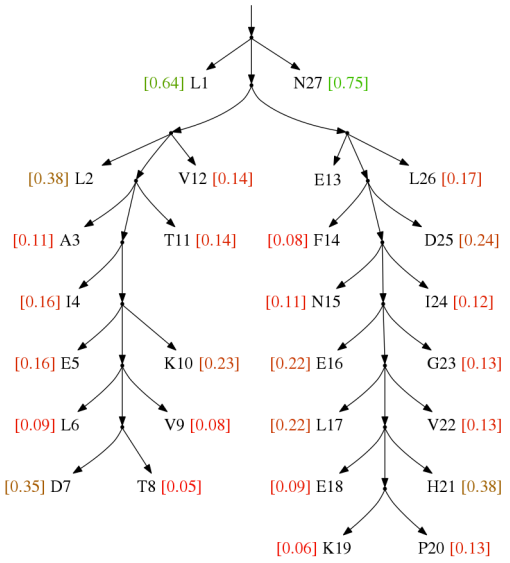
We investigated several quantitative measures of the consistency between the parse tree of context-free grammar and the contact map. Functionally, the measures can be divided into three groups: thresholded, contact-sensitive, topology-sensitive.

The thresholded measures (F_1, precision and recall) are best suited when a natural meaningful threshold can be easily defined. For the CFC grammars, it is the minimum possible distance between terminal symbols, which is achievable only by using the contact rules (δ=4). In such a case, the thresholded measures directly quantify the relative number of errors (missed and added contacts) in the tree. For this purpose, they were practical only when the overall topology of the parse tree was good enough to accommodate derivations with contact rules for considerable number of pairings (Tab. 1 and 2). Focusing on a single extreme threshold, the measures were agnostic to the overall shape of the tree, whether it was close or distant from the optimum (compare Tab. 1 and Fig. 4).

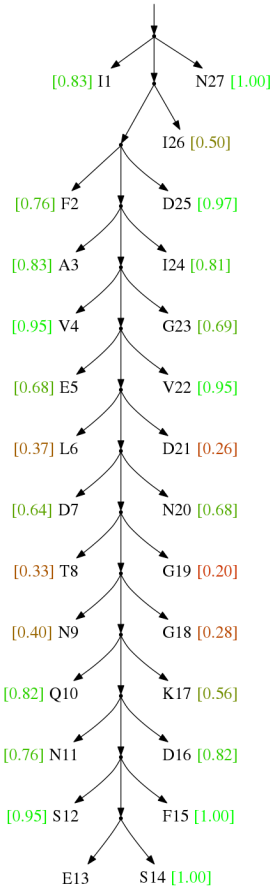
The contact-sensitive measures group is made of the aggregated precision measures, AP and AUPRC. They are characterized by a tight score distribution for a given number of errors in the tree, and good separation of scores for changing number of errors (Fig. 1). In addition, AP is more sensitive to false positive pairings than to false negative ones (Fig. 2 and 3). These features make AP especially suitable for detecting small defects in the trees of generally topologically correct structure (Tab. 1 and Fig. 7). Favorably, AP varied over virtually full range from 0.0 to 1.0 for the sample RNA molecule (Tab. 1), and from 0.1 to 0.7 for sample protein data (Fig. 6). Unexpectedly, the closely related AUPRC often scored the



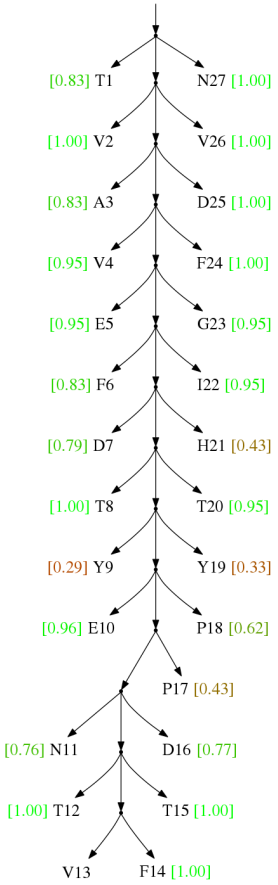
(a) pdb: 2zbj A4-30



(b) Worst AP
seq: NP_190127
AP=0.11,
S_1=-0.15



(c) Median AP
seq: NP_200734
AP=0.57,
S_1=0.43

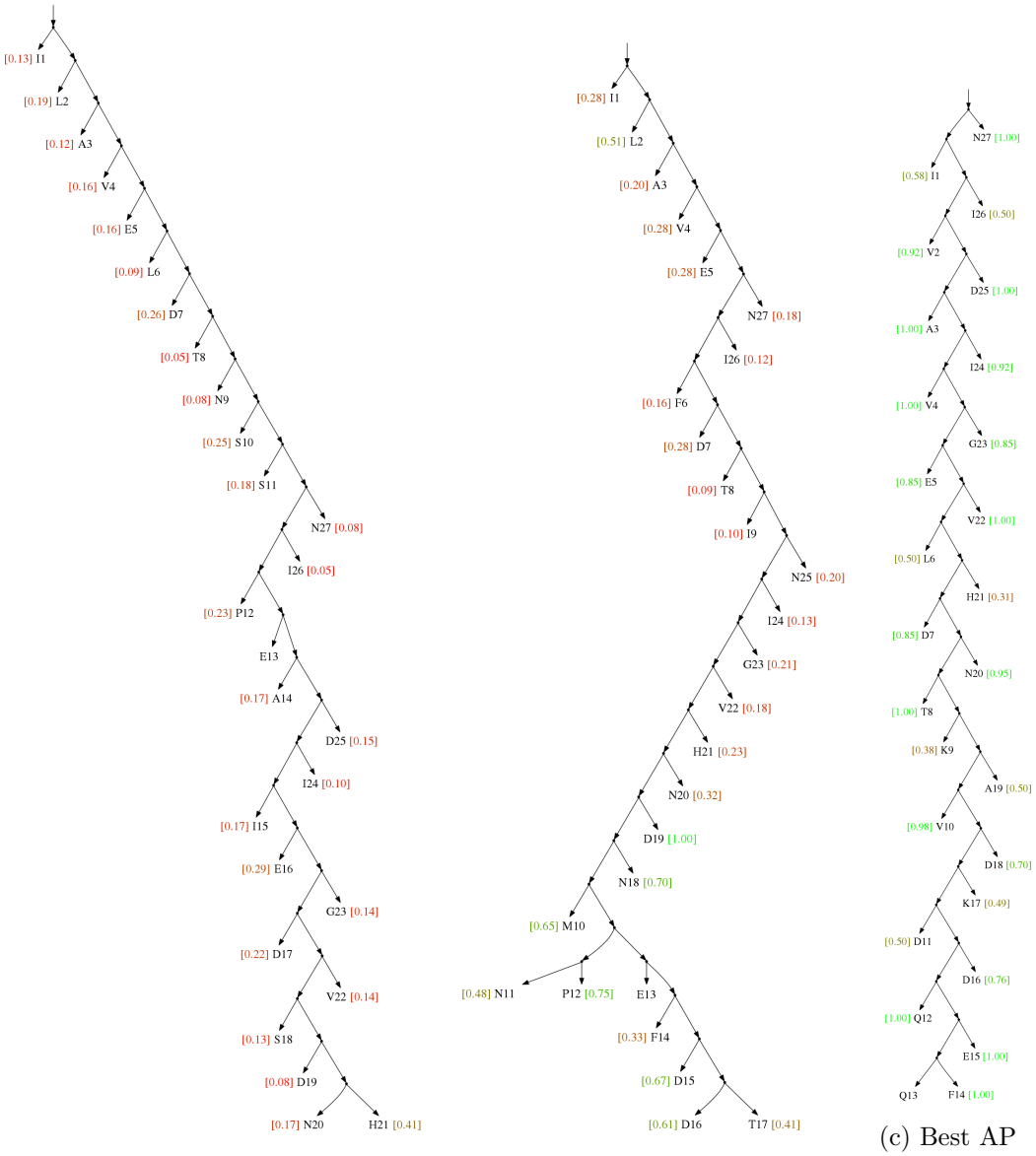


(d) Best AP
seq: P86352
AP=0.70,
S_1=0.46

Figure 7: Unlabeled parse trees obtained using the CFC grammars for CaMn ligand binding sequences (b-d). Leaves are annotated with *local* AP scores. The schematic representative spatial structure of the binding site is shown in (a). Endings and positions whose contacts are poorly modeled by the best AP tree are marked (see text).

same parse trees considerably higher than AP, for example in the presence of false positive pairings (Fig. 3). Also, it exhibited a peculiar bimodal distribution for some cases (see HET-s in Fig. 6). These are most likely artifacts of the implementation of AUPRC, which uses the linear interpolation and, therefore, can produce overly optimistic approximation of the area under the actual precision-recall curve.

The topology-sensitive measures group is made of AUROC and the mean-path measures, R_1 and S_1. They are characterized by wide and poorly separated score distributions given



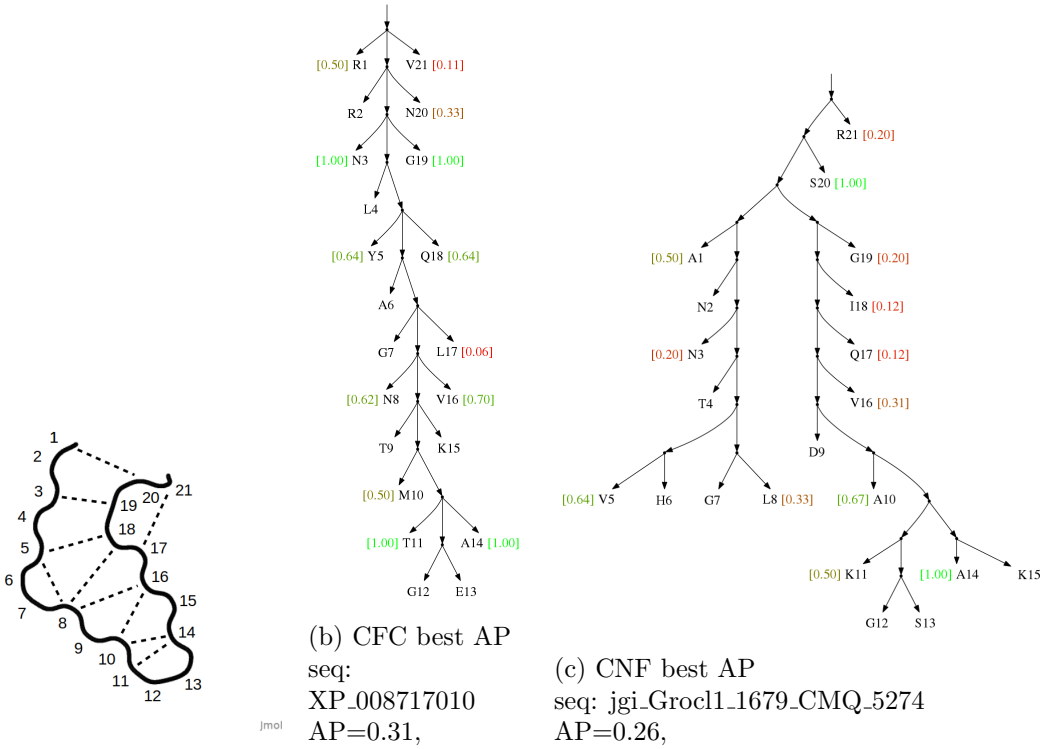
(a) Worst AP seq: Q9FIF0 AP=0.10, S_1=-0.16

(b) Median AP seq: NP_567233 AP=0.19, S_1=0.25

(c) Best AP seq: P_177168 AP=0.68, S_1=0.56

Figure 8: Unlabeled parse trees obtained using the CNF grammars for CaMn ligand binding sequences. Leaves are annotated with *local* AP scores.

a number of errors in the tree (Fig. 1). At the same time, qualitative analysis of the RNA and protein trees suggests that these measures are sensitive to errors breaking the overall topology of the tree (Tab. 1, Fig. 9). Therefore, they seem to be useful for discriminating between mediocre and moderately good parse trees, which are hardly discernible in terms



(a) pdb: 2kj3 A260-280

(b) CFC best AP seq: XP_008717010 AP=0.31, S_1=0.39

(c) CNF best AP seq: jgi_Grocl1_1679_CMQ_5274 AP=0.26, S_1=0.28

Figure 9: Unlabeled parse trees obtained for HET-s prion folding domain samples using the (b) CFC and (c) CNF grammars. Leaves are annotated with *local* AP scores. The schematic representative spatial structure of the domain is shown in (a) with all contacts with separation at least 3 marked using dashed lines.

of AP (Tab. 1 and Fig. 5, Fig. 9). The topology-sensitive measures have the interesting property of being hardly (AUROC) or only weakly (R_1 and S_1) sensitive to false positive pairings (Fig. 3, Tab. 1), which may be advantageous in situations where only a partial contact map is available as the ground truth. AUROC varied in the range of ca. 0.5 (0.2) to 1.0 (0.9) for the RNA (protein) sample data. The mean-path measures R_1 and S_1 do not admit fixed *right* value (even though the latter is scaled between -1 and 1), therefore, they have to be calibrated if they are intended to be used as absolute measures (Fig. 1). S_1 exhibited a strong almost linear correlation with AUROC (Pearson's r of 0.989 on the joint CaMn and HET-s protein sets). However, the optimal value of S_1 saturated slightly slower than for AUROC (cf. Tab. 1). R_1 was even more sensitive in the vicinity of optimal solutions. Despite discussed differences, the topology-sensitive measures were highly correlated with AP (Kendall's τ above 0.90 on the joint CaMn and HET-s protein sets).

Overall, our results support the thesis that the context-free parse trees of protein sequences can be robustly evaluated in the quantitative manner with reference to the protein contact map, despite the fact that a large share of protein contacts has the context-sensitive

character (crossing and overlapping). Thus, the study provides a solid ground for objective assessment of the explanatory power of the context-free grammars representing protein sequences. Further conclusions can be expected to arise when the proposed approach is applied in practice, for example to assessing inferred grammars against their predictive performance (Dyrka et al., 2018). As bioinformatic applications are often powered by the probabilistic CFGs (Knudsen and Hein, 1999; Dyrka and Nebel, 2009; Sciacca et al., 2011; Sükösd et al., 2012), the natural next step is to develop measures for assessing them. This may include modified versions of the parse tree measures analyzed in this study, and also measures aimed at assessing the probability distribution over all derivations (Knudsen and Hein, 2003). The latter would be a desirable complement to the former approach, since the maximum-likelihood trees of probabilistic CFG are not necessarily well approximates of the probability distribution over derivations with the contact rules (Dowell and Eddy, 2004). Another line of research may include using uncertain contact maps (Knudsen, 2005), which is especially tempting in the light of great progress in the protein contact prediction methods (Weigt et al., 2009; Wang et al., 2017).

Contributions WD, FC conceptualized the study and proposed the mean-path measures; MP, WD selected and implemented the measures, designed and conducted computational experiments; WD, MP analyzed the results; WD, MP, FC wrote the manuscript.

Acknowledgments

This research has been funded by National Science Centre, Poland [grant no 2015/17/-D/ST6/04054] and was supported by the E-SCIENCE.PL Infrastructure. Computational experiments have been partially carried out using resources provided by Wroclaw Centre for Networking and Supercomputing (http://wcss.pl) [grant no 98]. The authors would like to thank Olgierd Unold, Łukasz Culer and Agnieszka Kaczmarek for supporting early development of the project.

References

N Abe and H Mamitsuka. Predicting protein secondary structure using stochastic tree grammars. *Machine Learning*, 29:275–301, 1997.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acid Research*, 28:235–242, 2000.

K M Bohren, B Bullock, B Wermuth, and K H Gabbay. The aldo-keto reductase superfamily. cDNAs and deduced amino acid sequences of human aldehyde and aldose reductases. *Journal of Biological Chemistry*, 264(16):9547–51, 1989.

Anthony Bretaudeau, François Coste, Florian Humily, Laurence Garczarek, Gildas Le Corguillé, Christophe Six, Morgane Ratin, Olivier Collin, Wendy M Schluchter, and Frédéric Partensky. CyanoLyase: a database of phycobilin lyase sequences, motifs and functions. *Nucleic Acids Research*, 41:D396–D401, 2013.

François Coste and Goulven Kerbellec. Learning Automata on Protein Sequences. In Alain Denise, Pascal Durrens, Stéphane Robin, Eduardo Rocha, Antoine de Daruvar, and Alexis Groppi, editors, *JOBIM*, pages 199–210, Bordeaux, France, July 2006.

Jan Czekanowski. Zur differential Diagnose der Neandertalgruppe. Korrespondenzblatt der deutschen Gesellschaft für Anthropologie. *Ethnologie und Urgeschichte*, 40:44–47, 1909.

Asen Daskalov, Witold Dyrka, and Sven J. Saupe. Theme and variations: evolutionary diversification of the HET-s functional amyloid motif. *Scientific Reports*, 5:12494, 01 2015.

T.M. de Oliveira, P. Delatorre, B.A.M. da Rocha, E.P. de Souza, K.S. Nascimento, G.A. Bezerra, Tales R. Moura, R.G. Benevides, E.H.S. Bezerra, F.B.M.B. Moreno, V.N. Freire, W.F. de Azevedo, and B.S. Cavada. Crystal structure of Dioclea rostrata lectin: Insights into understanding the pH-dependent dimer-tetramer equilibrium and the structural basis for carbohydrate recognition in Diocleinae lectins. *Journal of Structural Biology*, 164(2): 177 – 182, 2008. ISSN 1047-8477.

Lee Raymond Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Robin D. Dowell and Sean R. Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5(1):71, Jun 2004. ISSN 1471-2105.

W. Dyrka. Probabilistic context-free grammar for pattern detection in protein sequences. Master's thesis, Faculty of Computing, Information Systems and Mathematics, Kingston University, London, 2007.

W Dyrka and J-C Nebel. A stochastic context free grammar based framework for analysis of protein sequences. *BMC Bioinformatics*, 10:323, 2009.

Witold Dyrka, Jean-Christophe Nebel, and Malgorzata Kotulska. Probabilistic grammatical model for helix–helix contact site classification. *Algorithms for Molecular Biology*, 8(1): 31, Dec 2013. ISSN 1748-7188.

Witold Dyrka, François Coste, and Juliette Talibart. Estimating probabilistic context-free grammars for proteins using contact map constraints. under review, preprint at arxiv.org, 05 2018.

S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

Sean R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10): e1002195, 10 2011.

Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655.

Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 2015.

E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, 30(11):1203–33, 2000.

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

B Knudsen and J Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15:446–54, 1999.

B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, 2003.

Martin Knudsen. Stochastic context-free grammars and RNA secondary structure prediction. Master's thesis, Aarhus University, Denmark, 2005.

H Mamitsuka and N Abe. Predicting location and structure of betasheet regions using stochastic tree grammars. In *Second International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA, USA*, pages 276–284. AAAI Press, 1994.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Soeding. HHblits: lightning-fast iterative protein sequence searching by HMM–HMM alignment. *Nature Methods*, 9(2):173–175, 2012.

Y Sakakibara, M Brown, R C Underwood, and I S Mian. Stochastic context-free grammars for modeling RNA. In *27th Hawaii Int Conf System Sciences*, pages 349–58, 1993.

Yasubumi Sakakibara. Efficient learning of context-free grammars from positive structural examples. *Information and Computation*, 97(1):23 – 60, 1992. ISSN 0890-5401.

Eva Sciacca, Salvatore Spinella, Dino Ienco, and Paola Giannini. Annotated stochastic context free grammars for analysis and synthesis of proteins. In Clara Pizzuti, Marylyn Ritchie, and Mario Giacobini, editors, *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, volume 6623 of *Lecture Notes in Computer Science*, pages 77–88. Springer Berlin / Heidelberg, 2011. ISBN 978-3-642-20388-6.

N Sharon and H Lis. Legume lectins–a large family of homologous proteins. *The FASEB Journal*, 4(14):3198–3208, 1990. PMID: 2227211.

Christian J. A. Sigrist, Edouard de Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(D1):D344–D347, 2013.

Johannes Soeding. Protein homology detection by HMM–HMM comparison. *Bioinformatics*, 21(7):951–960, 2005.

Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5(4):1–34, 1948.

Zsuzsanna Sükösd, Bjarne Knudsen, Jørgen Kjems, and Christian N.S. Pedersen. PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, 28(20):2691–2692, 2012.

Hélène van Melckebeke, Christian Wasmer, Adam Lange, Eiso AB, Antoine Loquet, Anja Böckmann, and Beat H. Meier. Atomic-resolution three-dimensional structure of HET-s(218–289) amyloid fibrils by solid-state NMR spectroscopy. *Journal of the American Chemical Society*, 132(39):13765–13775, 2010. PMID: 20828131.

Guido van Rossum and Jelke de Boer. Interactively testing remote servers using the Python programming language. *CWI Quarterly*, 4:283–303, 1991.

J Waldispuehl and J-M Steyaert. Modeling and predicting all-transmembrane proteins including helix-helix pairing. *Theoretical Computer Science*, 335:67–92, 2005.

J. Waldispuehl, B. Berger, P. Clote, and J.-M. Steyaert. Predicting transmembrane beta-barrels and interstrand residue interactions from sequence. *Proteins: Structure, Function and Genetics*, 65(1):61–74, 2006.

J. Waldispuehl, C.W. O'Donnell, S. Devadas, P. Clote, and B. Berger. Modeling ensembles of transmembrane beta-barrel proteins. *Proteins: Structure, Function and Genetics*, 71 (3):1097–1112, 2008.

Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*, 13 (1):1–34, 01 2017.

M Weigt, RA White, H Szurmant, JA Hoch, and T Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106:67–72, 2009.

Pawel P Wozniak and Malgorzata Kotulska. Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, 20(11): 2497, 2014.