Prediction of RNA Pseudoknotted Secondary Structure using Stochastic Context Free Grammars (SCFG)[1]

Rafael García
Politécnico Grancolombiano, Facultad de Ingeniería y Ciencias Básicas
Bogotá, Colombia
rgarcia@poligran.edu.co

Abstract

Pseudoknots are a frequent RNA structure that assumes essential roles for varied biocatalyst cell's functions. One of the most challenging fields in bioinformatics is the prediction of this secondary structure based on the base-pair sequence that dictates it. Previously, a model adapted from computational linguistics – Stochastic Context Free Grammars (SCFG) – has been used to predict RNA secondary structure. However, to this date the SCFG approach impose a prohibitive complexity cost [$O(n^4)$] when they are applied to the prediction of pseudoknots, mainly because a context-sensitive grammar is formally required to analyze them. Other hybrids approaches (energy maximization) give a $O(n^3)$ complexity in the best case, besides having several restrictions in the maximum length of the sequence for practical analysis.
Here we introduce a novel algorithm, based on pattern matching techniques, that uses a sequential approximation strategy to solve the original problem. This algorithm not only reduces the complexity to $O(n^2 \log n)$, but also widens the maximum length of the sequence, as well as the capacity of analyzing several pseudoknots simultaneously.

Keywords: pseudoknots, Stochastic Context Free Grammars (SCFG), secondary structure prediction, RNA, dynamic programming.

1. Introduction

Bioinformatics is an applied science where mathematical and computational theories and technologies are used in order to process, relate and derive predictions and inferences from data obtained in molecular biology. Bioinformatic's goal is to understand and analyze the information control and flow within different organisms. There is a synergic interaction between computer science, mathematics and biology, each with its own richness and limitations.
Within the bioinformatics scope, one of the most important fields is biological sequence analysis, which assumes the representation of the molecules that are essential for life (DNA, RNA, Proteins) as residues chains represented as symbols in a particular alphabet, thus allowing its study with grammatical analysis in order to find relations among these residues. These found relations determine, according to the central dogma[2] of molecular biology, the biological properties of such molecules [21][18]. This sequential representation of a molecule is known as its primary structure.
As many other proteins, RNA molecules presents distant pairings within a sequence. These bindings (secondary structure) manifest themselves as bonds within nucleotides located far and indistinctively in the sequence, resulting in a structural folding, known as Pseudoknot[3]. Since its discovery [17] these pseudoknots

[1] This paper is the english version of the paper *Predicción de Pseudonudos en la Estructura Secundaria de ARN vía Gramáticas Estocásticas Independientes del Contexto* [15] presented in CLEI 2005.
[2] The central dogma says the following:
a The DNA sequence determines the protein sequence.
b. The protein sequence determines the protein structure.
c. The protein structure determines the protein function.

[3] More formally, in a RNA sequence a pseudoknot occurs when a subsequence of a loop makes Watson-Crick pairs with a subsequence out of the loop. There exists two types of pseudoknots: H-pseudoknots (simples) and P-pseudoknots (recursives).

has drawn considerable attention because they give 3-D structure to the molecule, a structure that will determine in most cases its particular biological function[4].

Even though determining the molecular structure is of vital importance, it is also particularly hard and expensive to obtain structural data from RNA spectrometry and crystallography, so that an *a priori* prediction based on the residue sequence is an essential subject to bioinformatics [20]. We describe briefly current approaches to achieve these predictions.

In a molecular level it is plausible that RNA folding is dictated more by biophysical characteristics than by the mere count and relating of base pairs. Zuckers' minimization algorithm [27] assumes that the correct structural configuration is the one that presents the lowest equilibrium energy (ΔG^o). This prediction can be refined by combining it with experimental data, and statistical and thermodynamic information [24]. Nonetheless, given that the model for exact energy associated with the pseudoknots is not yet available, these existing models that use secondary structure energy approximations with relatively success [12] [9]. These approaches can't determine an optimal solution and most of the time can't tell how far from the optimal solution is the obtained one [20]. This is why nowadays no algorithm can predict in an accurate fashion pseudoknot classes; in fact the most important contributions present complexities that make them unusable[5].

There are several linguistic approaches to pseudoknots, most of them are variations of Stochastic Context Free Grammars (SCFG), which also aim to predict the structures and find them in databases. By its very definition, pseudoknots are described by a "copy language". This is the reason why formally it would be necessary a context sensitive grammar to analyze them, its corresponding problems would NP-Complete and their solutions have prohibitive complexity [6]. In order to avoid this problem, linguistic methods try the following approaches: intersections of associated SCFG [2]; special non terminal symbols and specific productions for each problem [19]; or parallel grammars that communicate between them [4]. Even so, its implementations require computational times that reach unfeasible levels [$O(n^6)$] and are not yet effective nor trustworthy [20].

In this article we propose a novel algorithm to solve the classical problems described above. This proposal consists in altering an existing effective algorithm known as Covariance Models (CM) [7], which is used to analyze secondary structure, in such a way that it is able to analyze pseudoknots. This is achieved by disaggregating pseudoknot prediction in two predictions of consecutive and related secondary structures, managing to divide the analysis in two iterations, thus avoiding simultaneous analysis, as well as its corresponding complexity and computational cost. This method yields a structure prediction in $O(n^2 \log n)$, improving the complexities in the algorithms developed so far and enlarging the maximum window of the sequence being analyzed.

This approach was inspired in the theory of information [22] and its handling of entropy, which states that the information is not evenly distributed across the message (sequence) but that there are regions or symbols that contain more information than others. In preserving or analyzing these "special" regions one can obtain a better comprehension of the whole message that most of the times is a cryptic message that encloses large quantities of unknown information – as in the case of the bioinformaticist's point of view –.

The article structure is as follows: in the second section the basic concepts of grammars that describe RNA sequences and basic concepts of SCFGs are introduced. In the third section the elements of the covariance model are exposed along with an innovative approach to predict pseudoknots in secondary structure as an extension to such models. In the fourth section an experimental evaluation of the proposed approach is presented, and in the last section conclusions and further work are presented.

[4] Virtually, pseudoknots are presents in all RNA types (ARNm, ARNt, lsuARN, ssuARN, 7-SL ARN, U2-snARN, I-group of introns, viral ARN) and possess many biological functions (see [12], [8], [16] or [5]).

[5] For example, the Rivas-Eddy algorithm [18] have temporal complexity $O(n^6)$, spatial complexity $O(n^4)$ and a constrained set of sequences to analyze (the sequence cannot contains more than 150 bases) [25]; the Lyngsø-Pedersen algorithm [14] have temporal complexity $O(n^4)$, spatial complexity $O(n^3)$ and make predictions about H pseudoknots; or another comparative divide and conquer algorithm with temporal complexity $O(n^2)$ and without constrains about the pseudoknot type. Though these *ad hoc* approaches needs manual and expert mediation [25]. At last, the ILM algorithm combines the thermodynamic and comparative approaches to obtain a more desirable and specific one than predecessors, but their final result is not optimal [19].

2. Secondary Structure Prediction via Stochastic Context Free Grammars.

2.1 Basic Concepts

One of the most promising techniques in bioinformatics is the analysis of stochastic grammars, since they allow the generation of sequence patterns in a natural way, besides having a broader range of action than other architectures [21].
Stochastic grammars have its origins in formal grammars that were developed as a model to analyze natural languages; in fact, these were developed at the same time that the double-helix model was developed by Watson and Crick [1]. Grammars are useful tools to model character sequences, in a certain way are useful to model molecular biology sequences [18] [1] [3]. Many bioinformatics problems can be reformulated in terms of formal languages, producing the corresponding grammar from the available data [1].
Among several utilities contributed by grammars, the main contribution is the ability to test by derivations if a sequence is syntactically correct, that is, if it belongs to a determined language. A derivation can be represented as a tree-like structure known as derivation tree. This tree reflects the syntactical structure of a sequence. It is possible that for a given sequence there are more than one derivation tree. In this case, we say that the grammar is ambiguous. In ambiguous grammars, complexity for the derivation rises given that the possible trees grows exponentially with the length of the sequence to be derived [1]. For a complete revision on basic concepts of formal language theory, including the study of grammars, reading of [10] or [13] is recommended.

2.2 Stochastic Context Free Grammars.

There is a wide variety of palindrome examples present in RNA/DNA. Biological palindromes have a different connotation than the usual one, for its letters are not identical but base-complementary[6], starting from the two ends. This is, in the same way that "a man, a plan, a canal, panama!" is usually seen as a palindrome (if the blanks and punctuation are removed), for its sequence mean the same thing even if we it is inverted, it is understood that AGAUUUCGAAUCU is a biological palindrome (given that if read from right to left a sequence of complementary bases to the original sequence). In other words, due to the complementary nature of the DNA double-helix structure, each half of a palindrome on a strand has its mirror image in the opposing strand[7] [1].
This distant pairings causes that the computational tools used commonly for protein analysis can't be used for RNA secondary structure. This is due to the nested structure observed frequently in RNA that can't be modeled in an efficient way using classic sequence correspondences (Neural Networks, regular languages or Hidden Markov Models (HMM)). For this reason an SCFG can be a better approach [6].
An SCFG is built by adding a probabilistic structure to the production rules of a given grammar [1]. In other words, each production rule $\varphi \rightarrow \delta$ has an assigned probability $P(\varphi \rightarrow \delta)$. $P(\varphi \rightarrow \delta)$ denotes the probability of using rule $\varphi \rightarrow \delta$ in a derivation, thus complying with the second axiom of probability $\Sigma_\delta(P_\delta(\varphi \rightarrow \delta))=1$.
A stochastic grammar is then characterized by a set of probabilistic models that generate the corresponding language.
In this way, to measure the probability $P(O|w)$ [8] of the stochastic derivation of a sequence O according to a grammar $M = M(w)$ with parameter w, is enough to compute:

$$P(O|w) = \Sigma P(S \rightarrow_\pi O|w)$$

where π is the sequence of derivations needed to produce O.

[6] The RNA bases **A** (Adenine) **C** (Cytosine) conform chemical links in complementary base-paring with **U** (Uracile) and **G** (Guanine), respectively.
[7] This structural ambiguity confers different roles to an element of the RNA sequence. In evolutive sense, this ambiguity gives a competitive adventage to the organism [23]. For example, secondary structure of a recursive palindrome is a stem with another side stem. Many structures associated to recursive palindromes are stems with several branches (orthodox secondary structures), such as the trefoil structure of tranferences RNA [1].
[8] This notation denote the probability of the occurrence of O given the occurrence of the parameters w.

3. A modification to the Covariance Models as an approach to the prediction of secondary structure

3.1 Covariance Model construction

Using an approach inherited from HMM, covariance models specify an SCFG architecture adequate to model consensus RNA secondary structures. Consensus structures are repetitive patterns in a RNA family that share various structural *motifs* among them[9] [6]. This method is inspired in the CYK dynamic programming algorithm, which has been the standard to analyze SCFG alignments. In synthesis, this algorithm finds an optimal derivation tree for a parameterized model in a sequence, being an optimization to the *inside/outside* algorithm [11]. At the same time, this algorithm computes the best score assigned to a given sequence from the various alternatives that a given SCFG may generate.
To describe with a SCFG this multiple alignment between RNA homologous sequences, various types of non terminal symbols are needed to model the different known structures.
The Non-terminals of the grammar included in the CM, and their semantic are: (see Figure 1) [6]:

- **P**: the paired columns in Watson-Crick bridges (bonds A-U C-G) are described by a non terminal that emits a *base pairing*.
- **L**: The non paired columns are described by a non terminal that emits *to the left* (direction 5´→ 3´)[10] whenever possible; that is, when no possible ambiguous sequences may arise.
- **R**: non terminal that emits *to the right* (direction 3' →5'). Case that can occasionally happen in protuberances between stems and loops in the right part of the structure (strand 3')[11]. It is used when ambiguous sequences emerge when **L** is used.
- **B**: Bifurcation non terminal used to split several stems or loops with various branches arising from it.
- **S**: Beginning non terminal that acts as immediate son to a bifurcation's derivation or a sequence start.
- **E**: Ending non terminal that finishes the derivation of sequences.
- **D**: Suppression non terminal that is used to describe a production that does not emit terminal symbols and does not describe one of the previous cases.

Each one of the non terminals has, by its stochastic characteristic, a probability $e_v(a,b)$ of emitting one or two pairs of bases. Here v is the non terminal and $a, b \in \{A,G,C,U\}$. The symbol $t_v(Y)$ represents the probability to go from state v to state Y.

Production			Description	P (Emission)	P(Transition)
P	→	aYb	Pair derivation (16 possbile emitting types)	$e_v(a,b)$	$t_v(Y)$
L	→	aY	Left derivation (4 possbile emitted symbols)	$e_v(a,b)$	$t_v(Y)$
R	→	Ya	Right derivation (4 possbile emitted symbols)	$e_v(a,b)$	$t_v(Y)$
B	→	SS	Bifurcation	1	1
D	→	Y	No symbol derivation	1	$t_v(Y)$
S	→	Y	Start	1	$t_v(Y)$
E	→	ϵ	End	1	1

Figure 1. Types of grammatical rules and its probabilities in a Covariance Model[12].

[9] *Homologous structures* in molecular vocabulary..
[10] In molecular biology, 5´ and 3´ are the ending points of the DNA/RNA sequences. The direction is associated with many celular functions such as duplication and replication.
[11] The secuence must be readed in 5´→ 3´ direction.
[12] Quoted from [6], pp 278.

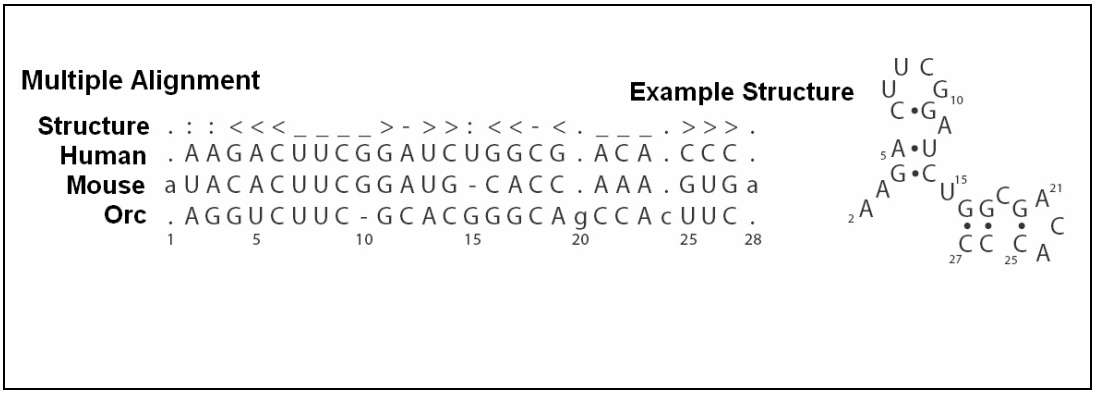

Figure 2. [13] A fictitious alignment of three sequences where 24 consensus columns are modeled out of 28 possible columns. The annotated structure sequence is consensus for the structure on the right, a structure that corresponds to the human sequence[14].

To build a CM it is necessary to begin with an alignment[15] of RNA sequences, with its correspondent annotation of the consensus secondary structure and the annotations of which columns must be considered insertions and which should be considered consensus columns (see Figure 2). From these columns a consensus structural tree is built.
The CM will be a directed graph of M states with transitions given by $t_v(y)$, with each state numbered in a way that $(y,z) \rightarrow v$. The CM can then be visualized as an array of transitions that run in only one direction, which guarantees two things: an iterative dynamic programming calculation through all the model states and that all transitions for state v can be maintained as a displacement and a count.

3.2 A Pseudoknot secondary structure prediction implementation as an extension of Covariance Models

Understanding CM, and generally speaking all SCFG, as computational tools that allow to probabilistically model distant relations between symbols in the molecular language, then it is possible to extend this concept to pseudoknots. This extension makes possible its prediction when in the original sequence the distant paired segments (copy language) are deleted, and then it is feasible to analyze the resulting sequence as usual.
Just as the relation that exist between residues of a stem is separated by non related residues, in the same way a pair of residues that form a pseudoknot will be separated by residues that are not necessarily related to the pseudoknot being analyzed[16]. Because of this, just as non related residues are set apart to determine the relation between base pairs that form the stem, in the same way it is possible to set apart existing relations between non related residues to determine the particular relation between the base pairs that form the pseudoknot.

[13] This alignment is presented in STOCKHOLM format. See details in: http://www.cgr.ki.se/cgb/groups/sonnhammer/Stockholm.html .
More information about formats used in molecular biology in: http://www.psc.edu/general/software/packages/hmmer/manual/node38.html
[14] Quoted from [6], pp 278.
[15] An alignment pretends to give a quantitative measure of similarity between sequences. To compare, an alignment assigns correspondences between symbols in the sequences.
[16] This is the SCFG approach. The SCFG approach is in the opposite direction of thermodynamic approach.

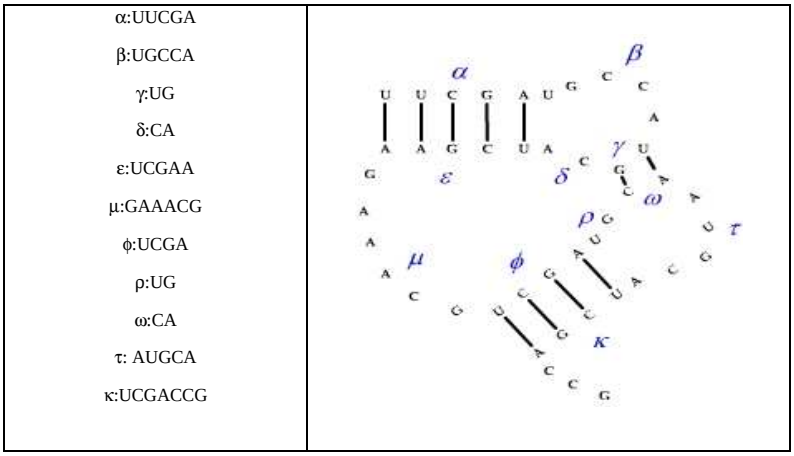

Figure 3. Canonical sketch of a pseudoknotted structure, divided in substring $\alpha\beta\gamma\delta\varepsilon\mu\phi\rho\omega\tau\kappa$, ready to subsequent abstraction.

Definition. If γ and δ are RNA sequences such that $|\gamma|=|\delta|$ and, always that $1\leq i\leq|\gamma|$ you have that γ_i y $(\delta^{-1})_i$ are base-complementary, where δ^{-1} is δ written backwards. Then it is stated that $\gamma\delta$ is a biological palindrome.

Note that if x is a RNA sequence that can be written as the concatenation of five subsequences this is $x=\pi\gamma\psi\omega\sigma$ (π y σ possibly empty) such that $\gamma\omega$ is a biological palindrome, then the CM model must predict the pairing (γ, ω^{-1}) as part of the secondary structure of x (see Figure 4).

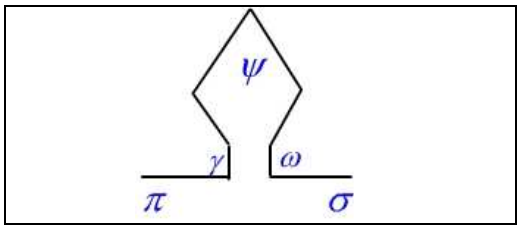

Figure 4. Abstraction sketch of the pseudoknotted structure $\pi\gamma\psi\omega\sigma$, ready to subsequent analysis.

Now, if $\gamma\omega$ is a biological palindrome and $\pi\psi\sigma$ can be decomposed as the concatenation of five substrings $\alpha\beta\mu\nu\rho$ such that $\beta\nu$ is a biological palindrome, then the pairing (β, ν^{-1}) is also predicted by the CM model as part of the secondary structure of x.

On the other hand, if $\gamma\omega$ is a biological palindrome and $\pi\psi\sigma$ can't be decomposed as in the last paragraph, then it is possible to infer that in $x=\pi\gamma\psi\omega\sigma$ one cannot find pairings that don't propose pseudoknots in the secondary structure of x.

Let $x'=\pi\psi\sigma$ be the sequences obtained of x by eliminating all paired base pairs according to the CM model. If there are sequences π', γ', ψ', ω' and σ' such that $x' = \pi\psi\sigma = \pi'\gamma'\psi'\omega'\sigma'$ and $\gamma'\omega'$ is a biological palindrome, then the pairing (γ', ω'^{-1}), that is part of the secondary structure of x', report a pseudoknot present in the secondary structure of x.

The previous statement can be concluded because only two cases make sense:

- $\gamma' \subseteq \pi \wedge \omega' \subseteq \psi$
- $\gamma' \subseteq \psi \wedge \omega' \subseteq \sigma$.

Thus, it is intended that πψσ preserve the information from the previous analysis while relations between γ and ω are analyzed. In this way the problem is divided and a second iteration can be made to find the two levels of relations in a separate way. After this step, the two analysis can be fused together and give the complete analysis at the end. In synthesis, instead of doing a parallel analysis as in [4], we propose a sequential analysis of two instances of the same problem, with evident gain in computational resources, which were before shared and now can be entirely dedicated to each of the two instances; besides the scale of structural complexity is reduced in both cases in a dramatic way.

3.3 General description for the validation of the proposed extension

To test the proposed algorithm a program was written in `C/C++`, which was executed in two systems, the result diverging only in the time of the required analysis. The fastest implementation was executed in an `INTEL/Linux` platform with 1.8GHz processor and 1024 Mbytes of RAM. The second implementation was executed in an `IRIX 6.5` platform (`SGI Origin 200`) with 4 parallel processors and 1024 Mbytes in RAM[17]. Such program analyze an alignment in STOCKHOLM[18] format, augmented with a marker *PK_cons*[19], which indicates the relations of pseudoknots according to the standard implemented by *PseudoBase* [26]. As an example figure 5 shows a theoretical alignment with a consensus pseudoknot for the sequences shown.

```
# STOCKHOLM 1.0
#=GF ID   trnaDummy
#=GF DE   taken from [Eddy 2002], with PK_cons indicating a fictitious pseudoknot

DF6280              GCGGAUUUAGCUCAGUUGGG.AGAGCGCCAGACUGAAGAUCUGGAGGUCC
DE6280              UCCGAUAUAGUGUAAC.GGCUAUCACAUCACGCUUUCACCGUGGAGA.CC
DD6280              UCCGUGAUAGUUUAAU.GGUCAGAAUGGGCGCUUGUCGCGUGCCAGA.UC
DC6280              GCUCGUAUGGCGCAGU.GGU.AGCGCAGCAGAUUGCAAAUCUGUUGGUCC
DA6280              GGGCACAUGGCGCAGUUGGU.AGCGCGCUUCCCUUGCAAGGAAGAGGUCA
#=GC SS_cons        <<<<<<<..<<<<.........>>>>.<<<<<.......>>>>>.....<
#=GC PK_cons        <<<[[<<..<<<<.........>>[>.<<<<<.......>>>>>.....]
#=GC RF             xxxxxxxxxxxxxxxxxxxx.xxxxxxxxxxxxxxxxxxxxxxxxxxxxx

DF6280              UGUGUUCGAUCCACAGAAUUCGCA
DE6280              GGGGUUCGACUCCCCGUAUCGGAG
DD6280              GGGGUUCAAUUCCCCGUCGCGGAG
DC6280              UUAGUUCGAUCCUGAGUGCGAGCU
DA6280              UCGGUUCGAUUCCGGUUGCGUCCA
#=GC SS_cons        <<<<.......>>>>>>>>>>>>.
#=GC PK_cons        <<<<.......>>>>>>>>]]>>.
#=GC RF             xxxxxxxxxxxxxxxxxxxxxxxx
//
```

Figure 5. Five RNA fragment alignment. Note that it is used `<>` to indicate relations between loops, and `[]` to indicate pseudoknots binding [20].

The written software abstracts the two levels of secondary structure and pseudoknots in two separate levels, one to indicate the secondary structure *SS_cons* (see Figure 6) and other to indicate the structure of pseudoknots *PK_cons* (see Figure 7) abstracting the one previously analyzed.
Once the information is processed in these two levels, the algorithm proceeds to analyze them separately to construct the consensus models in both structure levels. With these two models built it is possible to predict the structure of a nucleotide string without any included structure (see Figure 8).
The software will then compare separately the same string with the two consensus models, predicting the secondary structure that would probably have that sequence, as well as the probable pseudoknot that it would contain. The probable pseudoknot is abstracted in the secondary structure. After this step, the data is unified by building the pseudoknot structure to give the final result with the complete structure. The secondary structure and the pseudoknot structure (see Figure 8).

[17] The dicerences are due to two factors: the intensive use of the memory and the multyuser nature of the second computer.
[18] Full information about this format in: http://www.cgr.ki.se/cgb/groups/sonnhammer/Stockholm.html
[19] This is a non standard marker used only in this paper.
[20] For clarity, the Stockholm format has a header with relevant information followed by the name, the sequence and the consensus relations. The RF indicador announces that all columns in the alignment should be used in the model.

```
# STOCKHOLM 1.0
#=GF ID   trnaDummy
#=GF DE   taken from [Eddy 2002], with PK_cons indicating a fictitious pseudoknot

DF6280             GCGGAUUUAGCUCAGUUGGG.AGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCA
DE6280             UCCGAUAUAGUGUAAC.GGCUAUCACAUCACGCUUUCACCGUGGAGA.CCGGGGUUCGACUCCCCGUAUCGGAG
DD6280             UCCGUGAUAGUUUAAU.GGUCAGAAUGGGCGCUUGUCGCGUGCCAGA.UCGGGGUUCAAUUCCCCGUCGCGGAG
DC6280             GCUCGUAUGGCGCAGU.GGU.AGCGCAGCAGAUUGCAAAUCUGUUGGUCCUUAGUUCGAUCCUGAGUGCGAGCU
DA6280             GGGCACAUGGCGCAGUUGGU.AGCGCGCUUCCCUUGCAAGGAAGAGGUCAUCGGUUCGAUUCCGGUUGCGUCCA
#=GC SS_cons       <<<<<<<..<<<<.........>>>>.<<<<<.......>>>>>.....<<<<<.......>>>>>>>>>>>>.
#=GC RF            xxxxxxxxxxxxxxxxxxx.xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

Figure 6. First abstraction level. Note that the `[]` which indicated the pseudoknotted binding have been abstracted with `<>` to complete the existing loops.

```
# STOCKHOLM 1.0
#=GF ID   trnaDummy
#=GF DE   taken from [Eddy 2002], with PK_cons indicating a fictitious pseudoknot

DF6280              GCGGAUUUAGCUCAGUUGGG.AGAGCGCCAGACUGAAGAUCUGGAGGUCC
DE6280              UCCGAUAUAGUGUAAC.GGCUAUCACAUCACGCUUUCACCGUGGAGA.CC
DD6280              UCCGUGAUAGUUUAAU.GGUCAGAAUGGGCGCUUGUCGCGUGCCAGA.UC
DC6280              GCUCGUAUGGCGCAGU.GGU.AGCGCAGCAGAUUGCAAAUCUGUUGGUCC
DA6280              GGGCACAUGGCGCAGUUGGU.AGCGCGCUUCCCUUGCAAGGAAGAGGUCA
#=GC SS_cons        ...<<.................<.....................>
#=GC RF             xxxxxxxxxxxxxxxxxxx.xxxxxxxxxxxxxxxxxxxxxxxxxxx

DF6280              UGUGUUCGAUCCACAGAAUUCGCA
DE6280              GGGGUUCGACUCCCCGUAUCGGAG
DD6280              GGGGUUCAAUUCCCCGUCGCGGAG
DC6280              UUAGUUCGAUCCUGAGUGCGAGCU
DA6280              UCGGUUCGAUUCCGGUUGCGUCCA
#=GC SS_cons        ..................>>...
#=GC RF             xxxxxxxxxxxxxxxxxxxxxxx
//
```

Figure 7. Second abstraction level. Note that the `<>` which indicated the secondary relations have been abstracted with dots, and the `[]` which indicated the pseudoknotted binding have been abstracted to `<>`, thus allowing the secondary structure analysis.

```
Secuencia sonda en formato FASTA[21], con su identificador seguido de la
secuencia de nucleótidos.
>trna_yeast_phe
GCAGAUUUAGCUCAGUUGGCAGAGCGCCAGACUGAAGAUCUGGAGGUCCU
GUGUUCGAUCCACAGAAUACGCU

# STOCKHOLM 1.0
#=GF AU    Infernal 0.55

DF6280              GCGGAUUUAGCUCAGUuGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCU
#=GC SS_cons        (((((((,,<<<<___.____>>>>,<<<<<_______>>>>>,,,,,<<
#=GC PS_cons        (((([[((,,<<<<___.____>>[>,<<<<<_______>>>>>,,,,,]<
#=GC RF             gccgauaUagcgcAgU.GGuAgcgcgccacccUgucaagguggAGgUCcg

DF6280              GUGUUCGAUCCACAGAAUUCGCA
#=GC SS_cons        <<<_______>>>>>)))))))):
#=GC PS_cons        <<<_______>>>>>))))]]]]):
#=GC RF             gggUUCGAUuccccguaucggcg
//
//
```

Figure 8. Results of the secondary structure prediction in sequence *trna_yeast_phe.fa.* Note the secondary structure `<>` as well as the pseudoknot structure, displayed with `[]`[22].

[21] Full information about FASTA format in: http://www.ebi.ac.uk/help/formats_frame.html

4. Results and Experimental Evaluation

The set of source sequences for the test process has the same origin as the one used in [4][23]. This set, which can be found in the *tmRNA Database* [24], is already aligned and annotated for secondary structure and pseudoknots [28]. The tmRNA has at most 4 rightly annotated pseudoknots. From this database 35 sequences were downloaded. These sequences were organized in a phylogenic tree in such way that a test group for the model training and another group for the model testing were both in equilibrium at a phylogenic level, without any over represented class in the samples.
The source data were then aligned, deleting the non-shared columns between the bacteria sequences results in a complete alignment, obtaining a source suitable for bioinformatics analysis.
The appropriate model for the two abstracted structures was built from the 5 bacteria sequences that are in the phylum[25] of the bacteria to be analyzed. The following structure was derived from these two models starting with a sequence without structural annotation (see Figure 9).

```
Sequence without structural anotation (FASTA):
>Esch.coli v-2.       741 bp
GGGGCUGauucuggauucgaCGGGAUUUgcgaaAcCCaaGGUGAUGCCGAGGgGCGGUUggCCUCGua
aaaAGCCGCaaaaaauagucgcaaacgacgaaaacuacGCUUUAGCAGcuuaauaacCUGCuUAGAGC
CCUCUCUCCCUagCCUccgCUCUUAGGacGGGGaucaAGAGAGGucaaacccaaAAGAGaucGCGUgg
aAGCCcuGCCUGGGGUUgaaGCGUuaaaacuuaaUCAGGCuaGUUUGUUAGUgGcGuGUccGUCCaCa
GCUGGCAAGCgaauguaaaGACurACuaaGCAUGUaguACCgaGGaUguagAAAUUUCGgacGCGGGu
ucaacuCCCGCCAGCUCCacca

Fragment of the predicted structure (from base 01 to 100):

Esch.coli          GGGGCUGAUUCUGGAUUCGACGGGAUUUGCGAAACCCAAGGUGAUGCCGA
#=GC SS_cons       {{{{{{{,,,,,,,,,,,,[[[[[[[[-----[[[[[[[[[[[[[[<<
#=GC PS_cons       <<<<<<<,,,,,,,,,,,,<<<<<<<<-----<<<<<<<<<<<<<<[[[
#=GC RF            ggggcuguuuucGGuUUCGAcaggauuauCgAAaccaaaagugaugccga

Esch.coli          GGGGCGGUUGGCCUCGUAAAAAGCCGCAAAAAAUAGUCGCAAACGACGAA
#=GC SS_cons       <<-<<<<<<<_>>>>----->>>>>>>,,,,,,,,,,,,,,,,,,,,,,
#=GC PS_cons       [[-[[<<<[<_>>>>----]]]]]]],,,,,,,,,,,,,,,,,,,,,,
#=GC RF            ggGgcgguuggccucGuaAaaagccgcAaaAAAUAguCGCAAAcaAaaAA

(. . .)
```

Figure 9. A extract of the base sequence and corresponding predicted structure of *E. coli* built from an alignment retrieved from tmRNA DataBase[26].

If the predicted structure is compared to the structure provided by the *tmRNA Database* (see Figure 10), it is found that the algorithm predicts in general the whole structure, placing correctly the two pseudoknots that are found, even though it presents certain inconsistencies in the size of the pseudoknot or in the conforming bases in some region of the whole molecule. A false positive was never produced, that is, a pseudoknot placed by the algorithm where it should not be. This proves that the algorithm and it's prototype are able to analyze big quantities of data in a relatively trustworthy fashion, improving other algorithms which analyzed pseudoknots with similar results, but could only analyze little fragments, only one pseudoknot or totally ignoring the secondary structure that was not related to the pseudoknot [4].
It is quite important to point out that systems to date successfully identify pseudoknots simultaneously with sequences around 100 base pairs [4][5]. The sequence above has more than 740 base pairs, which is a noteworthy improvement in the practical field for pseudoknot analysis.

---

[22] Results presented in modifyed WUSS format. The original WUSS format not include the symbols [and] to mark pseudoknots. Punctuation symbols (.;,:) exhibit non pared remainder in the sequence.
[23] Software (GNU license) and biological sources used can be obtained by e-mailing to m.nino68@egresados.uniandes.edu.co.
[24] Online access: http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html
[25] Taxonomical division used by biologists to classify organisms.
[26] http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html.

```
Fragment of the predicted structure (from base 01 to 100):

Esch.coli        GGGGCUGauucuggauucgaCGGGAUUUgcgaaAcCCaaGGUGCAUGCCGAGGgGCGGUUggCCUCGuaaaa
#=GC SS_cons     <<<<<<<-------------<<<<<<<<-----<<<<<<<<<<<<<<<<<<<-<<<<<<<-<<->>---->
#=GC PS_cons     ------------------------------------------------((((((((((((udoknot-I))

Esch.coli        AGCCGCaaaaaauagucgcaaacgacgaaaacua
#=GC SS_cons     >>>>>>---------------------------
#=GC SS_cons     ))))))---------------------------
 (…)
//
```

Figure 10. Real structure of *E. coli* retrieved from tmRNA DataBase[27].

5. Conclusions

Given the rising amount of data the genome projects produce, the need for new sensible filters to find adequate sequences is preponderant. Besides, given that in some cases there is a lack of information to analyze the relations between the sequences, this filters need new characteristics to compare the sequences. One of the factors that are only starting to be exploited to such end is pseudoknots, that is why the algorithm proposed here is a contribution in that sense.

For the particular case of RNA and its secondary structural prediction, there is an attempt to decouple and disaggregate a complex problem by dividing it in sub problems of an easier analysis, paying special attention to preserving the coherence and in a rigorous way put the data back together. Because the results obtained were satisfactory, the success of this integrated approach is confirmed, approaches that have already been successful in other bioinformatics fields[28].

Acknowledgments

We gratefully acknowledge Carlos Jaramillo, for his multiple and patient conversations over molecular biology, and Rafael Gómez, for his inspiring suggestions about an integrated view of System Engineering and universal knowledge.

References

[1] Baldi, P.; Brunak, S. Bioinformatics: The machine learning approach. 2nd edition. Ed. *MIT Press*. (2001).

[2] Brown, M.; Wilson, C. RNA Pseudoknot Modeling Using Intersections Of Stochastic Context Free Grammars with Applications to Database Search. In Hunter, L. and Klein, T., editors. *Pacific Symposium on Biocomputing*, (1995), pp. 109-125.

[3] Brown, M. RNA Modeling Using Stochastic Context-Free Grammars. Phd´s Thesis. University of California, Santa Cruz. (1999).

[4] Cai, L; Russell, L; Wu, Y. Stochastic modelling of RNA pseudoknotted structures: a grammatical approach. *Bioinformatics* vol 19 (2003), pp. i66-i73.

[5] Deogun, J.S.; Donis, R., Komina, O. and Ma, F. RNA Secondary Structure Prediction with Simple Pseudoknots. In Proc. Second Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand. *CRPIT*, 29. Chen, Y.-P. P., Ed. ACS. (2004), pp. 239-246.

[6] Durbin, R; Eddy, S; Krogh, A; Mitchinson, G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. *Cambridge University Press*. (1998).

[7] Eddy, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a

[27] http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html.
[28] For example, protein structure prediction [6].

sequence to an RNA secondary structure. *BMC Bioinformatics*. (2002) Vol 3 pp. 18 http://www.biomedcentral.com/1471-2105/3/18.

[8] Felden, B.; Massire, C; Westhof, E.; Atkins, J.; Gesteland, R. Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature. *Nucleic Acids Res.,* vol 29 (2001), pp. 1602-1607.

[9] Gultyaev, A.P. van Batenburg. F., Pleij, C. The Computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, 250 (1995), pp. 37-51.

[10] Hopcroft, J. E. & Ullman, J. D. *Introduction to automata theory, lenguage and computation*. Addison-Wesley. (1979).

[11] Lari, K.; Young, S. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language* 4 (1990), pp. 35-56.

[12] Lee, D.; Han, K. Prediction of RNA Pseudoknots – Comparative Study of Genetic Algorithms. *Genome Informatics* 13 (2002), pp. 414-415.

[13] Lewis, H. R., Papadimitriou, Ch. H., *Elements of the theory of computation*. Prentice-Hall. (1981).

[14] Lyngsø, R. B. and C. N. S. Pedersen (2000) Pseudoknot Prediction in Energy Based Models. To appear in *Journal of Computational Biology.*

[15] Niño, M. and García, R., Predicción de Pseudonudos en la Estructura Secundaria de ARN vía Gramáticas Estocásticas Independientes del Contexto. In Proc. of XXXI Conferencia Latinoamericana de Informática (CLEI2005), Santiago de Cali, Colombia. Díaz, F., Rueda, C. and Buss, A. Ed. (2005), pp. 747 – 758.

[16] Paillart, J.; Skripkin, E.; Ehresmann, B.; Ehresmann, C.; Marquet, R. In vitro evidence for a long range pseudoknot in the 5'-untranslatd and matrix coding regions of HIV-1 genomic RNA. *J. Biol. Chem*., Vol 277 (2002), pp. 5995-6004.

[17] Pleij,CWA, Rietveld,K and Bosch,L. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Res*. Vol 13,5 (1985), pp. 1717-1731.

[18] Rivas E.; Eddy S.. A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots. *Journal of Molecular Biology*, Vol. 285, no. 5 (1999), pp. 2053-2068(16).

[19] Rivas, E. Eddy, S. The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics* 16 (2000), pp. 334-340

[20] Ruan, J.; Stormo, G.; Zhang, W. An Iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics.* Vol 20 (2004), pp.58-66.

[21] Searls, D. The Computational linguistics of Biological sequences. In: *Artificial Intelligence and Molecular Biology*, chap 2. *AAAI Press*. (1992).

[22] Shannon, C.E. A mathematical theory of communication. *Bell system Technical Journal*, Vol 27 (1948), pp. 379-423, 623-656.

[23] Strickberger, M. Evolution. 3rd Edition. Jones and Bartlett Publishers (2000).

[24] Tabaska, J.; Cary, R.; Gabow, H.; Stormo, G. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics Journal*,Vol 14 (1998), pp. 691-699.

[25] Tahi, F.; Engelen, S.; Régnier, M). Pcdfold-An algorithm for prediction of RNA including all kinds of Pseudoknots. Computers and Chemsitry, Vol 26 (2002), pp. 521-530.

[26] Van Batenburg, F.; Gultyaev, A.; Pleij, W.; Ng, J.; Oliehoek, J. PseudoBase, a database with RN A pseudoknots. *Nuc. Ac. Res.*, Vol 28 (2000), pp. 201-204.

[27] Zuker, M. On finding al suboptimal foldings of an RNA molecule. *Science*, 244 (1989), pp.48-52.

[28] Zwieb, C.; Gorodkin, J.; Knudsen, B.; *et al*., tmRDB (tmRNA database). *Nucleic Acids Research*, vol 31, No.1, (2003), pp. 446-447.