REVIEW Genomics & Informatics Vol. 9(1) 5-11, March 2011

Recent Progresses in the Linguistic Modeling of Biological Sequences Based on Formal Language Theory

Hyun-Seok Park[1]*, Bulgan Galbadrakh[1,2] and Young-Mi Kim[2]

[1]Bioinformatics Laboratory, School of Engineering, Ewha Womans University, Seoul 120-750, Korea, [2]Natural Language Processing Laboratory, School of Natural Science, Huree Institute of Information and Communication Technology, Ulaanbaatar, P.O.Box 780, Mongolia

Abstract

Treating genomes just as languages raises the possibility of producing concise generalizations about information in biological sequences. Grammars used in this way would constitute a model of underlying biological processes or structures, and that grammars may, in fact, serve as an appropriate tool for theory formation. The increasing number of biological sequences that have been yielded further highlights a growing need for developing grammatical systems in bioinformatics. The intent of this review is therefore to list some bibliographic references regarding the recent progresses in the field of grammatical modeling of biological sequences. This review will also contain some sections to briefly introduce basic knowledge about formal language theory, such as the Chomsky hierarchy, for non-experts in computational linguistics, and to provide some helpful pointers to start a deeper investigation into this field.

Keywords: natural language processing, Chomsky hierarchy, bioinformatics, formal language theory

Introduction

In formal language theory, a language is simply a set of strings of characters drawn from some alphabet, where the alphabet is a set of symbols. The challenge of computational linguistics is to find concise ways of specifying a given language L, preferably in a way that reflects some underlying model of the source of that language (Searls, 1993). For example, we can use informal descriptions of the language L_{GENE} that make use of a natural description, as follows:

L_{GENE} = { w $\in$ {a, t, g, c}* | w begins with "atg" }

However, simply exhaustively enumerating languages as below is imopssible:

L_{GENE} = { atg, atga, atgt, atgg, atgc, atgaa, atgtt, atggg, atgcc, ⋯}

Regular expression, a widely-used method for specifying simple languages, can be used to define L_{GENE}, in a more concise way, as atg(alt|glc)*. Alternatively, the same language can be defined in formal grammar rules (Hopcroft and Ullman, 1979), as in Fig. 1(a), and as a finite state automata as in Fig. 1(b).

Formal grammar (*N*, *T*, *P*, *S*) consists of: a finite set of terminal symbols (T: usually represented by lowercase letters), a finite set of non-terminal symbols (*N*: usually represented by uppercase letters), a finite set of production rules with a left and a right-hand side consisting of a sequence of these symbols (P), and a start symbol (S). Those readers requiring a more detailed introduction to formal language theory and bioinformatics are referred to chapter 2 of the book, "Artificial intelligence and molecular biology" (Searls, 2002; Searls, 2003).

A derivation is a rewriting of a string using the rules of the grammar. Thus, a rule may be applied to a sequence of symbols by replacing an occurrence of the symbols on the left-hand side of the rule with those that appear on the right-hand side. For example, by applying the production rules in Fig. 1(a), the string "atgcca" can be derived from the non-terminal S, by applying a series of derivations: $S \rightarrow aA \rightarrow$ atB $\rightarrow$ atgC $\rightarrow$ atgcC $\rightarrow$ atgccC $\rightarrow$ atgccaC $\rightarrow$ atgcca.

The simple grammar topologies or even less expressive formalisms can be sufficient to characterize biological sequences in many cases. But they cannot model long-term dependencies such as contacts of amino acids that are far in the sequence but close in the physical folding of the protein. In order to model higher-order structures of biological sequences, we need more powerful grammatical systems based on formal language theory, as a biological sequence can be thought of as a richly-expressive language for specifying the structures and processes of life. Searls (1988) initiated pioneering works to view biological sequences

*Corresponding author: E-mail neo@ewha.ac.kr
Tel +82-2-3277-2831, Fax +82-2-3277-2306
Accepted 2 March 2011

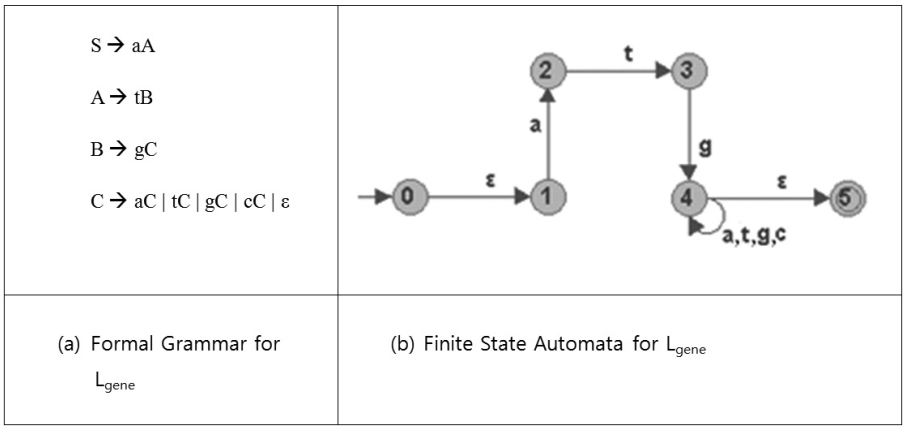

Fig. 1. (a) Formal Language and (b) Finite State Automaton, generating the language L_{GENE} = { w ∈ {a, t, g, c}* | w begins with "atg" }. The grammar represents a set of finite-length sequences of symbols that may be constructed by repeatedly applying production rules to the start symbol S.

simply as linguistic sentences (Searls, 1988). When we view these sequences just as strings on alphabets, a grammatical representation based on formal language theory can be applied to various problems for biological sequence analyses. Indeed, linguistic grammars have been used to model and predict multiple sequence alignments, transcription binding sites, RNA folding and secondary structures, integrons, insertion sequences, genes, and gene cassettes.

The remainder of this paper introduces the Chomsky hierarchy (Chomsky, 1957), and surveys bioinformatics approaches based on regular grammars, context-free grammars, and context-sensitive grammars, to model various types of biological sequences.

Modeling of Biological Sequences Categorized by the Chomsky Hierarchy

In Transformational Analysis (Chomsky, 1955) and Syntactic Structures (Chomsky, 1957), Chomsky initiated the theory of generative grammar and of the theory of formal languages as a branch of mathematical logic. The Chomsky hierarchy refers to a containment hierarchy of classes of formal grammars. Fig. 2 summarizes each of four types of grammar.

In increasing complexity and power, they are called type-3, type-2, type-1, and type-0 - each one a subclass of the next. Each type can be defined by a class of grammars, as indicated in Fig. 2. Exactly how much linguistic power is actually required to model biological sequences is one of the ultimate questions in bioinformatics.

The following sections survey various approaches based on grammar formalisms, ordered by type-3 (regular grammars), type-2 (context-free grammars), and type-1 (context-sensitive grammars) languages.

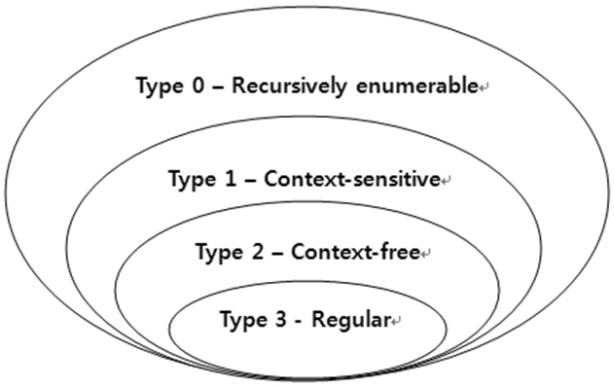

Fig. 2. The Chomsky Hierarchy is a series of increasingly complex classes of formal languages. The simplest are regular languages, followed by context-free, context-sensitive, and recursively enumerable languages.

Models based on regular grammars

Regular grammars, the first level of Chomsky's hierarchy, in Fig. 2 (Type-3 grammars), generate regular languages. Regular grammar is a formal grammar (N, T, P, S), such that all of the production rules in P are of one of the following forms:

1. *A → a - where A is a non-terminal in N and a is a terminal in T*
2. *A → Ba (or A → aB) - where A and B are in N and a is in T*
3. *A → ε - where A is in N and ε is the empty string.*

Regular languages can be described by regular expressions, and they are commonly used to define search patterns and the lexical structure of languages. Head (Head, 1987) initiated a formal analysis of the generative power of recombinatorial behaviors in biological sequences. His persistent splicing languages are shown to coincide with a class of regular languages. Brazma,

Jonassen, Eidhammer, and Gilbert (Brazma *et al*., 1998) surveyed approaches and algorithms used for the automatic discovery of patterns with expressive power in the class of regular languages.

An early work toward learning grammars based on regular expressions is the work of Yokomori (Yokomori, 1994) on learning a special type of regular language called a locally testable language from positive data, and its application in identifying the protein α-chain region in amino acid sequences. Peris's group used a grammatical approach to predict coiled-coil proteins (Peris *et al*., 2006) and transmembrane domains in proteins (Peris *et al*., 2008).

Actually, if we extend our scope, and consider the fact that most of the works on patterns (Liew *et al*., 2005) can be represented by regular grammars, the fields of so-called motif bioinformatics belong to type-3 grammar. Also, many of the current motif databases are based on the expressive power of regular grammar. To mention one among a few, the Prosite (Hulo *et al*., 2006) and ProRule (Sigrist *et al*., 2005) databases, one of the most successful databases, define signatures of known families of amino acid sequences that are expressed in sub-regular expressions.

On the other hand, a problem faced in these kinds of large-scale realistic grammars is that more than one production rule may apply to a structure. Naturally, probabilistic grammars have often been used to circumvent these ambiguities by using a probabilistic model consisting of a non-probabilistic model plus some numerical quantities. Among probabilistic grammars, the profile Hidden Markov Model, or pHMM (Eddy, 1998; Krogh *et al*., 1994) is most closely related to regular grammars, because an *n*-gram is a subsequence of *n* items from a given sequence, and language models built from n-grams are actually (*n*-1)-order Markov models.

Coste and Kerbellec. (2005) showed a successful application of the classical state merging framework developed in grammatical inference, to learning automata on selection and ordering of similar fragments to be merged, and on physico-chemical property identification (Coste and Kerbellec., 2005). Their work offers the opportunity to learn more expressive topologies than those of pHMMs, while still benefiting from the weighting schemes developed for pHMM.

Recently, Tsafnat *et al*. (2011) used computational grammar inference methods to automate LGS (Larger than Gene Structures) discovery (Tsafnat *et al*., 2011). The authors compared the ability of six algorithms to infer LGS grammars from DNA sequences annotated with genes and other short sequences.

As we have discussed, regular grammars (including motif bioinformatics) are the most prevalently used formalism in bioinformatics. Still, the limitations of regular grammar are that regular grammar can only model the primary structures of biological sequences and cannot explicitly model higher-order structures such as secondary structures of RNAs and tertiary structures of proteins. In order to model higher-order structures of biological sequences, many researchers have used more powerful grammatical systems.

Models based on context-free grammars

As stated in the previous section, regular grammar has its limitations, and the approaches based on type-3 language have been criticized, especially by Chomskyans, because they lack any explicit representation of long-range dependency.

In the Chomsky hierarchy in Fig. 2, type-2 grammar, or context-free grammar generates context-free language, which is more powerful than regular grammar. In terms of production rules, every production of a context-free grammar is of the form:

$A \rightarrow w$

where A is a single non-terminal symbol, w is a string, and the left-hand side of a production rule is always a single non-terminal symbol.

We can specify a simple grammar representing an RNA palindrome in the following way:

S → aSu
S → uSa
S → gSc
S → cSg
…

The grammar above captures long-range dependency. For example, the palindrome string "aug…cau" can be derived from the non-terminal S, by applying a series of derivations: $S \rightarrow$ aSu → auSau → augScau …

Numerous attempts have been made to solve the problems of modeling of families of homologous RNA sequences, and predicting RNA secondary structure prediction techniques, since computational recognition based on type-2 grammar has been shown to perform in polynomial time.

In the 1990's, DCGs (definite clause grammars) (Pereira and Warren, 1980), a kind of logic programming-based grammars, were adopted to study so-called "DNA linguistics" (Searls, 1993). Later, DCGs and the Prolog programming language were used in modeling gene regulation (Collado-Vides, 1992; Rosenblueth *et al*., 1996), benefiting from features such as parameter-passing and arbitrary Prolog code embeddings. Basic Gene Grammar, an attempt to simplify the representations of DNA sequences, and with expressive power equivalent to that of DCG, has been used to model and predict

transcription binding sites (Leung *et al*., 2001).

Nevill-Manning and Whitten (Neville-Manning and Whitten, 1997) initiated an attempt to produce context-free grammars of biological sequences, in an automatic way. Later, similar attempts have been made by other researchers, along this line (Apostolico and Lonardi, 2000; Carrascosa *et al*., 2011; Cherniavsky and Ladner, 2004; Lanctot *et al*., 2000), generating grammars based on repeated phrases. This task can be formalized as the problem of finding the smallest context-free grammar by recursively replacing the repeats by non-terminals.

Muggleton *et al*. (2001) investigated whether Chomsky-like grammar representations are useful for learning cost-effective, comprehensible predictors of members of biological sequence families (Muggleton *et al*., 2001). As a case study, they proved that the most cost-effective, comprehensible multi-strategy predictor of human neuropeptide precursors employ context-free grammar.

Like the cases for type-3 grammars, many of the attempts based on type-2 grammars also used stochastic methods, especially to resolve difficulties that arise because longer sentences are highly ambiguous when processed with realistic grammars. Here, a stochastic context-free grammar (SCFG) can be obtained by specifying a probability for each production in a context-free grammar. SCFGs extend context-free grammars in the same way that HMMs extend regular grammars. It is a more expressively powerful class of stochastic grammars than the HMMs.

A pioneering work was performed by Sakakibara *et al*. (1994), extending the notion of profile HMMs (Eddy, 1998; Krogh *et al*., 1994) to profile SCFG. They assessed the ability of trained SCFGs to perform three tasks: to discriminate transfer RNA (tRNA) sequences from nontRNA sequences, to produce multiple alignments, and to ascertain a secondary structure of new sequences. Knudsen and Hein (1999) also suggest SCFGs as an alternative probabilistic methodology for modeling RNA structure (Knudsen and Hein, 1999).

For representative databases based on SCFG, RFAM (Gardner *et al*., 2009; Griffiths-Jones *et al*., 2003) of modeling common non-coding RNA families by stochastic context-free grammars called covariance models, can be cited (Eddy and Durbin, 1994).

SCFGs have also been used for alignments of sequences. Pair stochastic context-free grammars (PSCFGs) have been studied for alignments of a pair of RNA sequences without any prior information about their secondary structures (Holmes and Rubin, 2002; Rivas and Eddy, 2001). PSCFGs are a generalization of stochastic context-free grammars and can generate an aligned pair of sequences. Later, the notion of pair HMMs defined on alignments of linear sequences is extended to pair stochastic tree automata, called Pair HMMs on Tree Structures (PHMMTSs) (Sakakibara, 2003), defined on alignments of trees. PSCFGs are used by the multiple structural alignment softwares such as Stemloc (Holmes, 2005).

In (Chuong *et al*., 2006), the authors used a secondary structure prediction method based on conditional log-linear models (CLLMs), a flexible class of probabilistic models that generalize upon SCFGs by using discriminative training and feature-rich scoring.

Dowell and Eddy (2004) studied the tradeoffs between model complexity and prediction accuracy, by comparing nine different small SCFGs, and concluded that SCFG designs have prediction accuracies near the performance of free energy minimization models (Dowell and Eddy, 2004); still, probabilistic methods have not replaced free energy minimization methods as the tool of choice for secondary structure prediction, as the accuracies of the best SCFGs have yet to match those of the best physics-based models.

Recently, Dyrka and Nebel (2009) a framework, based on the combination of stochastic context-free grammars related to different physico-chemical properties of amino acids and on genetic algorithms, which was shown to produce relevant protein binding site descriptors (Dyrka and Nebel, 2009).

Models based on context-sensitive grammars

The presence of pseudoknot secondary structures of noncoding RNA molecules and the presence of repeats of many varieties in biological sequences indicate the need to use more powerful grammar formalism. Modeling various repeat sequences, or the pseudoknot structures of RNAs is beyond the generative power of context-free grammars and inevitably involves the complexity of context-sensitivity.

Type-1 grammars (context-sensitive grammars), in Fig. 2, generate the context-sensitive languages. Context-sensitive grammar is a formal grammar (N, T, P, S) such that all of the production rules are of the following forms:

$\alpha A \beta \rightarrow \alpha \gamma \beta$

where $A \in N$, $\alpha, \beta \in (N \cup T)^*$ and $\gamma \in (N \cup T)+$ are applied.

Actually, the term, “context-sensitive” comes from the fact that a symbol can have different interpretations, depending on *where* it appears in the input language. Thus, unlike context-free grammars, more than one symbol can appear on the left-hand side of the grammars. For example, we can specify a grammar representing $\{a^n t^n g^n \mid n \rangle 0\}$ in the following way:

A → aATG
A → aTG
aT → at
tT → tt
GT → TG
G → g

Using the rules above, the string "aattgg" can be derived from the non-terminal A, by applying a series of derivations: A → aATG → aaTGTG → aatGTG → aatTGG → aattGG → aattgG → aattgg.

A few attempts have been made to represent pseudoknots. Rivas and Eddy (Rivas and Eddy, 1999) suggested a formal transformational grammar that avoids the use of general context-sensitive rules by introducing a small number of auxiliary symbols used to reorder the strings generated by otherwise context-free grammar.

Sometimes, special grammar formalisms, classified as mildly context-sensitive grammars, are used (Joshi *et al*., 1988). Uemura *et al*. (1999) defined two subclasses of tree-adjoining grammar (Joshi *et al*., 1975) called sl-tag and esl-tag, and argued that esl-tag is appropriate for representing RNA secondary structures including pseudoknots (Uemura, 1999). Matsui *et al*. (2005) proposed the pair stochastic tree-adjoining grammars (PSTAGs) for modeling pseudoknots, showing that their method significantly improves the prediction accuracies of RNA secondary structures (Matsui *et al*, 2005).

Abe and Mamitsuka (Abe and Mamitsuka, 1999) studied a more powerful class of grammar, called stochastic ranked node rewriting grammars, than SCFGs and applied them to the problem of secondary structure prediction of proteins, concentrating on the problem of predicting β-sheet regions, to capture the parallel and anti-parallel dependencies and their combinations.

Rivas and Eddy (2000) introduced a new class of grammars for deriving RNA secondary structure by a sequence with a single hole (Rivas and Eddy, 2000). The grammar is based on a number of auxiliary symbols used to reorder the strings. Cai, Malmberg, and Wu (2003) described a formal transformational grammar that extends context-free grammar, based on parallel communicating grammar systems (Cai *et al*., 2003). The key feature is the use of special non-terminal symbols that dictate specific rearrangements of substrings in a derivation.

Context-sensitive grammar formalisms have also been used to model non-coding RNAs. To model ncRNA precursors, Yoon and Vaidyanathan (2004) proposed a context-sensitive HMM (CSHMM), which is an extension of the idea of HMMs by introducing a memory, in the form of a stack or a queue (Yoon and Vaidyanathen, 2004). Later, Agarwal *et al*. (2011) extended the idea slightly and proposed a CSHMM structure with two context-sensitive states to model miRNA sequences (Agarwal *et al*., 2011).

Patridge *et al*. (2009) used similarities between DNA and natural languages (Baquero, 2004) to develop a context-sensitive grammar to define cassette arrays. Also, Tsafnat *et al*. (2009) presented a method to discover higher-order DNA structures, using a context-sensitive deterministic grammar (Tsafnat *et al*., 2009). These grammars have been applied to the discovery of gene cassettes associated with integrons.

On the other hand, there has been an attempt to model RNA structures with pseudoknots, using just context-free grammars by adding four building blocks of genus to the conventional secondary structures (Reidys *et al*., 2011). Reidys *et al*. (2011) used the natural topological classification of RNA structures, resulting in corresponding unambiguous multiple context-free grammars to provide an efficient dynamic programming approach.

Conclusion

In this review, we tried to gather a list of published works, categorized by the expressive power of formal grammars. The lists might not be exhaustive, and there is a possibility that some of the works, presented here, could have been misclassified, because not all the works fall within the traditions of the Chomsky hierarchy.[1)]

The genome may not be just a molecule with patterns. It may be a language, and an information storage mechanism. Precisely constructed sequence models for linguistic structure can play an important role in the process of biological discovery itself. In this respect, grammatical representations have increasing importance in the field of bioinformatics for biological sequence analyses.

It is noteworthy that a recent comparison of the predictive power of learned grammars against an expert-developed grammar shows the possibility that an inferred grammar can represent a general model that accurately identifies structures without referring to prior knowledge about them (Tsafnat et al., 2011). More positively, a linguistically modeled biological sequences may automatically provide solutions for thorny biological problems and thus provide us a deeper understanding of genome.

1) Those readers requiring a more detailed introduction to a number of works for applications of profile HMMs to the problems of gene finding and promoter analyses are referred to (Durbin *et al*., 1998). Grammatical inference methods may find grammatical structures hidden in biological sequences, and those readers who are interested in automatic syntax acquisition or grammatical inferences are referred to (Coste, 2010; Sakakibara, 2005). Those works would be worth reading further to gain a more comprehensive understanding of this field.

Acknowledgments

We would like to thank Dr. Philippe Meunier, for his valuable comments. This work was partly sponsored by the Korean government scholarship program of NIIED (National Institute of International Education) and the industry-university cooperation program of SMBA (Small and Medium Business Administration).

References

Abe, N., and Mamitsuka, H. (1999). A New Method for Predicting Protein Secondary Structures Based on Stochastic Tree Grammars. *Proc. 11th Int'l Conf. Machine Learning* 3-11.

Agarwal, S., Vaz, C., Bhattacharya, A., and Srinivasan, A. (2010). Prediction of novel precursor miRNAs using context-sensitive hidden Markov model (CSHMM). *BMC Bioinformatics* 11(suppl 1), S29.

Apostolico, A., and Lonardi, S. (2000). Off-line compression by greedy textual substitution. *Proceedings of the IEEE.* 88, 1733-1744.

Apostolico, A., and Lonardi, S. (2000). Compression of biological sequences by greedy off-line textual substitution. In: Data Compression Conference. 143-153.

Baquero, F. (2004). From pieces to patterns: evolutionary engineering in bacterial pathogens. *Nat. Rev. Microbiol.* 2, 510-518.

Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *J. Computat. Biol.* 5, 279-305.

Cai, L., Malmberg, R.L., and Wu, Y. (2003). Stochastic Modeling of RNA Pseudoknotted Structures: A Grammatical Approach. *Bioinformatics* 19, 66-73.

Carrascosa, R., Coste, F., GallŽ, M., and Infante-Lopez, G. (2011). Searching for smallest grammars on dna sequences. *Journal of Discrete Algorithms, Elsevier* (to be published).

Cherniavsky, N., and Ladner, R.E. (2004). Grammar-based compression of DNA Sequences. UW CSE Technical Report (TR2007-05-02), presented at the DIMACS Working Group on the Burrows-Wheeler Transform.

Chomsky, N. (1957). Syntactic Structures. Mouton and Co.

Chuong, B.D., Daniel, A.W., and Serafim, B. (2006). RNA secondary structure prediction without physics-based models. *Bioinformatics* 22, e90-e98.

Collado-Vides, J. (1992). Grammatical model of the regulation of gene expression. *Proc. Natl. Acad. Sci. USA,* 89, 9405-9409.

Coste, F., and Kerbellec, G. (2005). A similar fragments merging approach to learn automata on proteins. In Gama, J., Camacho, R., Brazdil, P., Jorge, A., Torgo, L., eds. ECML. Volume 3720 of Lecture Notes in *Computer Science*., Springer. 522-529.

Coste, F. (2010). Biological Sequences by Grammatical Inference, author manuscript, published in ICGI 2010 Tutorial Day, Valencia: Espagne *(http://www.irisa.fr/symbiose/francois_coste)*

Dowell, R.D., and Eddy, S.R. (2004). Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* 5,

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge Univ. Press.

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

Eddy, S.R., and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucl. Acids Res.* 22, 2079-2088.

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

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Gri_ths-Jones, S., Eddy, S.R., and Bateman, A. (2009). Updates to the RNA families database. *Nucl. Acids. Res.* 37, 136-140.

Griffiths-Jones, S., Bateman, A., Marshall, MI, Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. *Nucl. Acids Res.* 31, 439-441.

Head, T. (1987). Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors. *Bull. math. Biol.* 49, 737-759.

Holmes, I., and Rubin, G. (2002). Pairwise RNA structure comparison with stochastic context-free grammars. In *Proceedings of 5th Pacific Symposium on Biocomputing.* World Scientific Press, Singapore, pp. 163-174.

Holmes, I. (2005). Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics* 24, 6-73.

Hopcroft, J.E., and Ullman, J.D. (1979). Introduction to Automata Theory, Languages, and Computation, Addison-Wesley Publishing, Reading Massachusetts, ISBN 0-201-029880-X.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro E., Langendijk-Genevaux, P.S., Pagni, M., and Sigrist, C.J.A. (2006). The PROSITE database. *Nucl. Acids. Res.* 34, D227-D230.

Joshi, A.K., Levy, L.S., and Takahashi, M. (1975). Tree adjunct grammars. *J. Computer & System Sciences* 10, 136-163.

Leung, S.W., Mellish, C., and Robertson, D. (2001). Basic Gene Grammars and DNA-Chart Parser for language processing of Escherichia coli promoter DNA sequences, *Bioinformatics* 17, 226-236.

Knudsen, B., and Hein, J. (1999). RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* 15, 446-454.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling, *J. Mol. Biol.* 235, 1501-1531.

Lanctot, J.K., Li, M., and Yang, E.H. (2000). Estimating DNA sequence entropy. In *ACMSIAM Symposium on Discrete Algorithms*. 409-418.

Liew, A.W., Yan, H., and Yang, M. (2005). Pattern recognition techniques for the emerging field of bioinformatics.

A review, *Pattern Recognition* 38, 2055-2073.

Matsui, H., Sato, K., and Sakakibara, Y. (2005). Pair Stochastic Tree Adjoining Grammars for Aligning and Predicting Pseudoknot RNA Structures. *Bioinformatics* 21, 2611-2617.

Nevill-Manning, C.G., and Witten, I.H. (1997). Compression and explanation using hierarchical grammars. *The Computer Journal* 40, 103-116.

Partridge, S.R., Tsafnat, G., Coiera, E., and Iredell, J. (2009). Gene cassettes and cassette arrays inmobile resistance integrons, *FEMS Microbiol. Rev.* 33, 757-784.

Pereira, F., and Warren, D. (1980). Definite clause grammars for language analysis. *Artif. Intell.,* 13, 231-278.

Peris, P., L'opez, D., Campos, M., and Sempere, J.M. (2006). Protein motif prediction by grammatical inference. In *LNCS (LNAI).* Sakakibara, Y., Kobayashi, S., Sato, K., Nishino, T., Tomita, E., eds. (Springer: Heidelberg) 4201, pp. 175-187.

Peris, P., L'opez, D., and Campos, M. (2008). IgTM: An algorithm to predict transmembrane domains and topology in proteins. *BMC Bioinformatics* 9, 367.

Reidys, M., Huang, W.D., Andersen, E., Penner, C., Stadler, F., and Nebel, E. (2011). Topology and prediction of RNA pseudoknots. Bioinformatics advance access, doi:10.1093/bioinformatics/btr090.

Rivas, E., and Eddy, S. (2000). The Language of RNA: A Formal Grammar That Includes Pseudoknots. *Bioinformatics* 16, 334-340.

Rivas, E., and Eddy, S. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2, 8.

Rosenblueth, D., Thieffry, D., Huerta, A., Salgado, H., and Collado-Vides, J. (1996). Syntactic recognition of regulatory regions in Escherichia coli. *Comput. Appl. Biosci.* 12, 415-422.

Sakakibara, Y. (2003). Pair Hidden Markov Models on Tree Structures. *Bioinformatics* 19, 232-240.

Sakakibara, Y. (2005). Grammatical Inference in Bioinformatics. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27, 1051-1062.

Searls, D.B. (1988). Representing Genetic Information with Formal Grammars. In *Proceedings of the 7th National Conference on Artificial Intelligence.* 386-391.

Searls, D. (1993). The computational linguistics of biological sequences. In Artificial Intelligence and Molecular Biology, chapter 2, Hunter, L., ed. (MIT Press: Boston, MA), pp. 47-120.

Searls, D.B. (2002). The language of genes. *Nature* 420, 211-217.

Sigrist, C.J.A., De Castro, E., Langendijk-Genevaux, P.S., Le Saux, V., Bairoch, A., and Hulo, N. (2005). ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics* 21, 4060-4066.

Tsafnat, G., Coiera, E., Partridge, S.R., Schaeffer, J., and Iredell, J.R. (2009). Context-driven discovery of gene cassettes in mobile integrons using a computational grammar. *BMC Bioinformatics* 10, 281.

Tsafnat, G., Schaeffer, J., Clayphan, A., Iredell, J.R., Partridge, S.R., and Coiera, E. (2011). Computational inference of grammars for larger-than-gene structures from annotated gene sequences. *Bioinformatics* 27, 791-796.

Uemura, Y., Hasegawa, A., Kobayashi, S., and Yokomori, T. (1999). Tree-Adjoining Grammars for RNA Structure Prediction. *Theoretical Computer Science* 10, 277-303.

Yokomori, T., Ishida, N., and Kobayashi, S. (1994). Learning local languages and its application to protein α-chain identification. In: System Sciences, vol.5: Biotechnology Computing, Proceedings of the Twenty-Seventh Hawaii International Conference. 113-122.

Yoon, B.J., and Vaidyanathan, P.P. (2004). RNA secondary structure prediction using context-sensitive hidden Markov models. Proceedings of IEEE International Workshop on Biomedical Circuits and Systems (BioCAS): Dec. 2004, Singapore.