RNA-Sequence-Structure Properties and Selenocysteine Insertion

Rolf Backofen*

Institut für Informatik, LMU München
Oettingenstraße 67, D-80538 München
backofen@informatik.uni-muenchen.de

Abstract. Selenocysteine (Sec) is a recently discovered 21st amino acid. Selenocysteine is encoded by the nucleotide triplet UGA, which is usually interpreted as a STOP signal. The insertion of selenocysteine requires an additional signal in form of a SECIS-element (**Se**leno**c**ysteine **I**nsertion **S**equence). This is an mRNA-motif, which is defined by both sequence-related and structure-related properties.

The bioinformatics problem of interest is to design new selenoproteins (i.e., proteins containing selenocysteine), since seleno variants of proteins are very useful in structure analysis. When designing new bacterial selenoproteins, one encounters the problem that the SECIS-element itself is translated (and therefore encodes a subsequence of the complete protein). Hence, changes on the level of mRNA made in order to generate a SECIS-element will also modify the amino acid sequence. Thus, one searches for an mRNA that is maximally similar to a SECIS-element, and for which the encoded amino acid sequence is maximally similar to the original amino acid sequence. In addition, it must satisfy the constraints imposed by the structure.

Though the problem is NP-complete if arbitrary structural constraints are allowed, it can be solved efficiently when we consider the structures as used in SECIS-elements. The remaining problem is to generate a description of the SECIS-element (and its diversity) based on the available data.

1 Selenocysteine Insertion

Selenocysteine (Sec) is a rare amino acid, which was recently discovered as the 21st amino acid [4]. Proteins containing selenocysteine are consequently called *selenoproteins*.

The discovery of selenocysteine was another clue to the complexity and flexibility of the mRNA translation mechanism. Since all 64 codons are either associated with one of the 20 standard amino acids or a STOP-signal, a codon has to be overloaded for encoding this 21st amino acid. Selenocysteine is encoded by the UGA-codon, which is usually a STOP-codon encoding the end of translation

* Partially supported by the DFG within the national program SPP 1087 “Selenoprotein – Biochemische Grundlagen und klinische Bedeutung”

by the ribosome (see Figure 1). It has been shown [4] that in the case of selenocysteine, termination of translation is inhibited in the presence of a specific mRNA sequence in the 3'-region after the UGA-codon that forms a hairpin like structure (called Sec insertion sequence (SECIS), see Figure 2). The interesting part is that in bacteria, the SECIS-element itself is translated into an amino acid sequence. Mutagenesis analysis has shown that the sequence, structure, and distance to the UGA-codon is important for selenocysteine insertion. Furthermore, they vary between different organisms [15, 23, 22].

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	$\boxed{\text{Stop}}$	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	⋮	⋮	⋮	⋮	

Fig. 1. Translation of RNA. Every amino acid is encoded by a nucleotide triple (called codon). The assignment of codons to amino acid is usually depicted by a three-dimensional array, the genetic code. The selenocysteine codon is encircled.

An mRNA sequence is an element of $\{A, C, G, U\}^*$; A, C, G, U are called *nucleotides*. An important property of the corresponding RNA-molecule is that it can form a distinguished structure. The structure of an RNA-molecule is formed by hydrogen bonds between complementary nucleotide pairs (namely A–U and C–G). Concerning structure, we distinguish three levels of abstractions:

1. primary structure: the mRNA sequence itself.
2. secondary structure: the set of hydrogen bonds that are realized by the corresponding RNA-molecule. Note that any nucleotide can have at most one bond to another nucleotide. The secondary structure of an mRNA-sequence can be determined efficiently by dynamic programming approaches [25, 24, 7, 14].
3. tertiary structure: the real three-dimensional structure of the corresponding RNA-molecule, which is very hard to determine.

Although the real three-dimensional structure of an RNA molecule is biologically important, it is generally agreed that for most purposes, it suffices to consider the secondary structure of the corresponding mRNA-sequence.

In the following, we will consider the problem of designing new selenoproteins. We will give a formal description of this problem. For this formalization, we have presented an efficient algorithm in another paper [2]. This algorithm requires a statistical representation of SECIS-elements. We will then review methods that are available to generate this description.

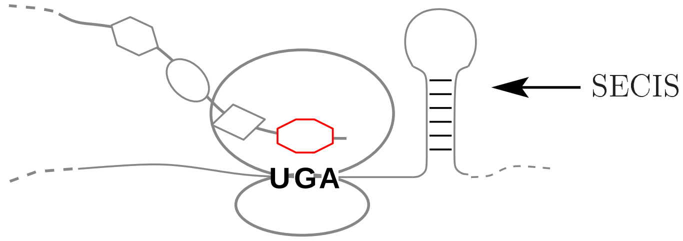

Fig. 2. SECIS-element and translation of mRNA by the ribosome in the case of selenocysteine.

2 Related Computational Problems

Concerning selenoproteins, bioinformatics currently faces the following challenges. An obvious problem is to search for existing but unknown selenoproteins. Since both sequence and structure is important for selenocysteine incorporation, and the different selenoproteins do not have a pronounced consensus sequence (on the amino acid level), the usually BLAST and FASTA search tools will not have enough selectivity. Therefore, new algorithms have to be devised. An example is SECISearch [20], which searches for mammalian selenoproteins by recognizing the corresponding SECIS-element using a hierarchical approach. First, one searches for sequences that satisfy the SECIS-consensus sequence, and allow for the secondary structure required for the SECIS-element. In a second step, the minimal free energy of the sequences is estimated by using the RNAfold program of the Vienna-RNA-Package [14].

However, we consider a different problem, namely modifying existing proteins such that selenocysteine is incorporated instead of another amino acid (e.g., instead of a cysteine). There are mainly two possible applications for newly generated selenoproteins. First, one can generate functionally modified proteins. Experimentally, this has been already done successfully for the bacterium *E. coli* [13], where the mRNA sequence of a protein not containing selenocysteine was engineered (by hand) such that it forms a *E. coli* SECIS (while preserving maximal similarity on the amino acid level). The engineered protein now had a selenocysteine in place of a catalytic cysteine (at position 41), and showed a 4-fold enhanced catalytic activity. Second, and more important, seleno-variants of proteins could be used for the phase determination in an X-ray crystallography (or in the NMR-analysis) of an given protein.

3 Finding New Insertion Sites

We consider the computational properties of substituting the amino acid at a given position of some protein by a selenocysteine. For this purpose, the mRNA-sequence of the protein has to be modified such that an appropriate SECIS-element is generated. Since the SECIS-element is later translated into an amino acid sequence, this modification implies that the amino acid sequence is modified, too. To keep the protein functional, we want to find a SECIS-element which minimizes the changes on the amino acid level.

For this purpose, we first need a description of the SECIS-element, generated from sequences known to be a SECIS-element.[1] For these sequences, the secondary structure usually has been determined. Since the structure of mRNA-motifs are more conserved than the sequence, we use the following two-level approach to describe the SECIS-element. For each possible secondary structure encountered in the set of SECIS-elements (only few), we collect all corresponding sequences into a so-called profile (or consensus sequence). A profile assigns every position of the considered structure a weight. This weight represents the frequence with which a nucleotide is observed at this position. In the following, we will call this *the profile for* a specific secondary structure.

For each combination of structure and corresponding profile, we search for an mRNA-sequence that allows the insertion of Selenocysteine, and satisfies the given structure. This (sub-)problem can be described formally as follows:

Input: An amino acid sequence $A_1 \dots A_n$, a graph $\Gamma = (V, E)$, $V = \{v_1, \dots, v_{3n}\}$ (the considered structure), functions $p_1, \dots, p_{3n}$ (the corresponding profile for this structure, where $p_i : \{A, C, G, U\} \to \mathbb{R}$), and a function sim : $\Sigma_{\text{Amino}} \times \Sigma_{\text{Amino}} \to \mathbb{R}$ measuring the similarity on the level of the amino acid sequence (e.g., given by the PAM250 similarity matrix [6]).

Output: Find mRNA-sequence $N_1, \dots, N_{3n} \in \{A, C, G, U\}^{3n}$ such that s.t. N_k is assigned to v_k and

1. $\{v_k, v_l\} \in E$ implies that N_k is the complement of N_l,
2. $\sum_{i=1}^{n} f_i(N_{3i-2}, N_{3i-1}, N_{3i})$ is maximized, where

$$\begin{aligned} f_i(N_{3i-2}, N_{3i-1}, N_{3i}) = {} & \alpha \times \text{sim}\left(A_i, \text{AMINO}(N_{3i-2}, N_{3i-1}, N_{3i})\right) \\ & + (1 - \alpha) \times \Big[p_{3i-2}(N_{3i-2}) + p_{3i-1}(N_{3i-1}) + p_{3i}(N_{3i})\Big]. \end{aligned}$$

α is a parameter for weighting the similarity on the amino acid level. $\text{AMINO}(N_{3i-2}, N_{3i-1}, N_{3i})$ is the amino acid encoded by the nucleotide codon $N_{3i-2}, N_{3i-1}, N_{3i}$. $\Gamma = (V, E)$ is called the *structure graph*. We denote the above problem by $MRSO(\Gamma, p_1, \dots, p_{3n}, \text{sim})$. An instance of this problem can be viewed by the following picture:

$$\begin{array}{rclllll}
\text{org. amino seq. } A & = & A_1 & \dots & A_i & \dots & A_n \\
 & & \wr & & \wr & & \wr \quad \leftarrow \text{sim}(A_i, A'_i) \\
\text{mod. amino seq. } A' & = & A'_1 & \dots & A'_i & \dots & A'_n \\
 & & \uparrow & & \uparrow & & \uparrow \\
\text{mod. mRNA seq. } N & = & \overbrace{N_1 N_2 N_3} & \dots & \overbrace{N_{3i-2} N_{3i-1} N_{3i}} & \dots & \overbrace{N_{3n-2} N_{3n-1} N_{3n}} \\
 & & \wr\ \wr\ \wr & & \wr\ \wr\ \wr & & \wr\ \wr\ \wr \quad \leftarrow p_j(N_j)\ (j \in [1 \dots 3n]) \\
\text{profile positions} & = & P_1\ P_2\ P_3 & \dots & P_{3i-2} P_{3i-1} P_{3i} & \dots & P_{3n-2} P_{3n-1} P_{3n}.
\end{array}$$

[1] Of course, the set of known SECIS-elements represent only a portion of set of all possible SECIS-elements.

It is easy to see that the problem cannot be simply solved by determining for each amino acid position $i \in [1..n]$ the codon $N_{3i-2}N_{3i-1}N_{3i}$ which is locally optimal (i.e., maximizes $f_i(N_{3i-2}, N_{3i-1}, N_{3i})$). The reason is that the choice of $N_{3i-2}, N_{3i-1}, N_{3i}$ may fix some other N_k by condition 1 of the problem description (depending on the given secondary structure). But the problem can be solved in linear time if we consider the typical hairpin structures of SECIS-elements. On the other hand, we have shown that the problem becomes NP-complete if we consider more complex structures having crossing edges in the linear embedding of the structure graph. Figure 3 shows two graphs associated two different types of mRNA structures. For the first type, namely the hairpin-like structure as encountered in SECIS-elements, the linear embedding of the structure graph does not contain crossing edges. The second type, namely pseudoknots, is important for programmed frameshifts which allow to encode two different amino acid sequences in one mRNA sequence. This mechanism has been identified in viruses as well as in prokaryotes and eukaryotes [9, 10].

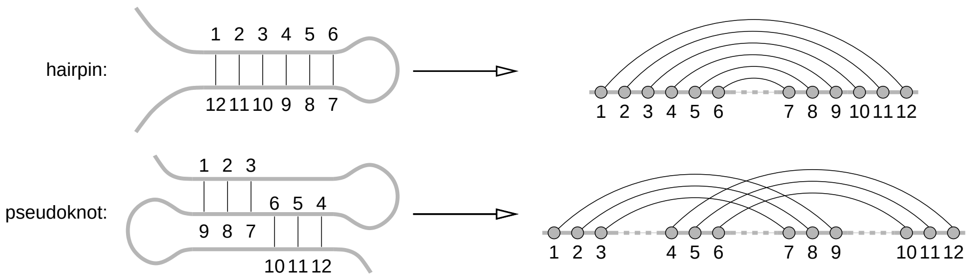

Fig. 3. Linear embedding of hairpin and pseudoknot. We have numbered only the positions that are part of a bond.

The computational properties of this problem are investigated in [2] and summarized in the theorem below. The linear time algorithm is a modified dynamic programming algorithm, and the NP-completeness is shown by reducing $3SAT$ to our problem.

Theorem 1. *Let* $(\Gamma, p_1, \ldots, p_{3n}, \mathrm{sim})$ *be an instance of* $MRSO$ *such that the structure graph* $\Gamma = (V, E)$ *is outer-planar (i.e., there exists a linear embedding that does not contain crossing edges. Then one can find an optimal mRNA-sequence* $N_1, \ldots, N_{3n}$ *in linear time. Furthermore, the decision version of* $MRSO$ *is NP-complete in the case of general structure graphs.*

For the general case, we have developed an approximation algorithm [1].

4 Determination of the Profile

There are several sources that provide data suitable for describing SECIS-Elements. The first source comes from so-called *mutagenesis experiments*. In these experiments, original SECIS-elements (called *wild-type*) are mutated, and the mutated variants are experimentally tested for their capability of inserting selenocysteine. On the one hand, this kind of experiments provide a valuable source for the descriptions of SECIS-elements, since the resulting sequences have been verified. On the other hand, usually only a small portion of the sequence space can be explored using this technique.

But there is another form of source that can be used to determine the consensus sequence(s) and structure(s) for SECIS-elements. The SECIS-element enforces incorporation not by their mere existence, but by the fact that there is a special protein which recognizes these elements. This protein is SelB, which is a special elongation factor for selenocysteine. An *elongation factor* is a protein which mediates the elongation of the amino acid sequence by the ribosome (see Figure 4). Thus, the capability of binding to the special elongation factor SelB is an important property that every SECIS-element has to satisfy. Now, there exists a high-throughput method called SELEX (**S**ystematic **E**volution of **L**igands by **Ex**pontential enrichment) [28]. In this method, a huge amount of RNA-sequences is generated and tested in parallel whether they bind to some ligand. Sequences that bind to SelB have been determined using this method [19, 18].

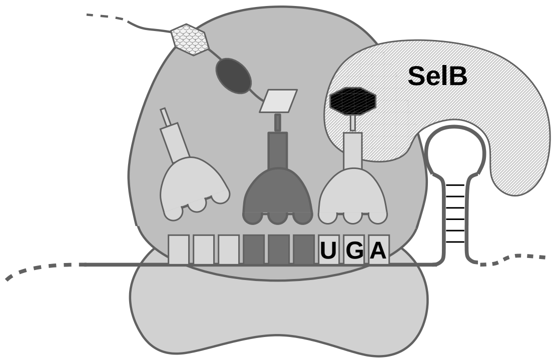

Fig. 4. A more pronounced view of selenocysteine insertion involving special elongation factor SelB.

The remaining problem is to determine the consensus sequence and structure from these data. The usual approach for determining the consensus sequence is to generate a multiple sequence alignment. A *multiple sequence alignment* is a two-dimensional representation of the evolutionary relationship between several sequences. The sequences are listed horizontally, with an additional gap symbols inserted at several places. Thus, every column contains for every sequence either a letter of this sequence, or a gap symbol (see Figure 5). A column in the multiple sequence alignment is called *highly conserved* if the column contains only a few gaps, and the majority of entries in this column are all equal. In Figure 5, columns 1,3,4,7,8,10 denote perfectly conserved columns.

a.)

$$\begin{array}{l}
\overbrace{A\ C\ A}^{M_{1,3}=0}\ C\ -\ A\ U\ U\ G\ U\\
A\ A\ A\ C\ C\ A\ U\ U\ U\ U\\
A\ U\ A\ C\ C\ A\ U\ U\ A\ U\\
A\ \underbrace{G\ A\ C\ C\ -\ U\ U\ C}_{M_{2,9}=2}\ U
\end{array}$$

b.)

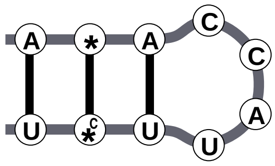

Fig. 5. Alignment and consensus structure for the sequences $s_1 = ACACAUUGU$, $s_2 = AAACCAUUUU$, $s_3 = AUACCAUUAU$, and $s_4 = AGACCUUCU$. a.) shows an alignment and the mutual information content for selected columns. b.) shows the consensus structure, where $\star$ denotes any nucleotide, and $\star^{\mathsf{C}}$ is the complement of the nucleotide assigned to $\star$.

The problem is that for RNA, the structure is usual more conserved than the sequence. An example are the two columns 2 and 9 in Figure 5, which are not conserved on the sequence level, but perfectly conserved in the corresponding consensus structure. The problem is that usual techniques for multiple sequence alignment will not align columns that are conserved only on the level of structure.

On the other hand, if one has given a correct multiple sequence alignment of RNA sequences, the consensus structure can be deduced from the multiple sequence alignment by evaluating the mutual information content of the columns in the alignment [8]. The mutual information content $M_{i,j}$ of two columns i and j represents the information gained when assuming that i and j depend on each other compared to the situation that they are independent. $M_{i,j}$ is calculated by

$$M_{i,j} = \sum_{x_i, x_j} f_{x_i x_j} \log \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}},$$

where x_i (resp. x_j) varies over all possible nucleotides that occur in column i (resp. j), f_{x_i} (resp. f_{x_j}) is the relative frequence of x_i (resp. x_j) in column i (resp. j). $f_{x_i x_j}$ is the relative frequence of occurrences of the pair (x_i, x_j) in columns i and j on the same row. $M_{i,j}$ has its maximal value of 2 if the symbols columns i and j are uniformly distributed (i.e., for all x_i, x_j we have $f_{x_i} = 0.25 = f_{x_j}$)[2], but are always forming a complementary base pair, i.e.,

$$f_{x_i x_j} = \begin{cases} 0.25 & \text{if } x_i \text{ and } x_j \text{ are complemetary base pairs} \\ 0 & else. \end{cases}$$

Then one can use a variant of a secondary structure prediction algorithm such as the basic Nussinov/Zuker algorithm [25, 24] (using the mutual information

[2] Recall that we are using the four letter alphabet $\{A, C, G, U\}$.

content as a score instead of energy or number of bonds) to calculate a secondary structure S that maximizes

$$\sum_{(i,j) \text{ form a bond in } S} M_{i,j}.$$

Of course, the problem is to generate the initial multiple sequence alignment. If the consensus structure were known, then one could use the apparatus of stochastic context-free grammars to generate good multiple sequence alignments [8, 12, 11, 26, 27, 5]. In this approach, one uses for every bond in the consensus structure a set of rules of the form $r : S \xrightarrow{p_r} xS'x^{\complement}$, where S, S' are non-terminals, x ranges over the terminal symbols $\{A, C, G, U\}$ (or over a subset of $\{A, C, G, U\}$), and $x^{\complement}$ is the nucleotide complementary to x. p_r is the probability assigned to the rule r. In addition, one needs rules to handle insertion to and deletions from the consensus structure [8]. Figure 6 shows an example of a SCFG for a small hairpin consensus structure. Once we have the SCFG, we can train it by using Expectation-Maximization (via the inside-outside algorithm[21, 3]) to re-estimate the probabilities p_r.

a.)

- $S \xrightarrow{0.75} a\ S_1\ u$
- $S_1 \xrightarrow{0.9} g\ S_2\ c$
- $\boldsymbol{S_2 \xrightarrow{0.5} g\ S_3\ c}$
- $\boldsymbol{S_3 \xrightarrow{1} aggucu}$
- $\boldsymbol{S \xrightarrow{0.25} c\ S_1\ g}$
- $\boldsymbol{S_1 \xrightarrow{0.1} c\ S_2\ g}$
- $S_2 \xrightarrow{0.5} c\ S_3\ g$

b.)

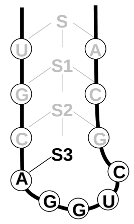

Fig. 6. Stochastic context-free grammar for RNAs. a.) shows a SCFG for a hairpin with 3 bonds, and b.) shows one parse and the corresponding RNA-sequence and structure.

The above described approaches show that we are in a perfect cycle. If we have a good multiple sequence alignment, then we can generate the consensus structure. But good multiple sequence alignments can usually only be generated if the consensus structure is known (via the SCFG-formulation). This is handled in [8] by simply starting with an initial random multiple sequence alignment, then iterating the cycle

$$\begin{array}{c} \textit{multiple sequence alignment} \\ \downarrow \\ \textit{consensus structure + SCFG + parameter re-estimation} \\ \downarrow \\ \textit{multiple sequence alignment} \end{array}$$

until a satisfying result is obtained. A similiar approach using the above described cycle is used in [17, 16], where the multiple sequence alignment is generated by a program that can first, align the sequence locally, and second, explore the structural properties of the aligned sequences.

5 Acknowledgment

I would like to thank Prof. Böck from the Institute of Microbiology, LMU München, for explaining him the biochemical background of selenocysteine, for many suggestion and the fruitful cooperation on the problems discussed in this paper. I would also like to thank N.S. Narayanaswamy, Firas Swidan and Sebastian Will for many helpful and enlightening discussions on related subjects. Furthermore, I would like to thank Andreas Abel, Jan Johannsen, Sven Siebert, and Slim Abdennadher for reading draft versions of this paper.

References

1. Rolf Backofen, N.S. Narayanaswamy, and Firas Swidan. On the complexity of protein similarity search under mRNA structure constraints. 2001. submitted.
2. Rolf Backofen, N.S. Narayanaswamy, and Firas Swidan. Protein similarity search under mRNA structural constraints: application to selenocysteine incorporation. 2001. submitted.
3. J. K. Baker. Trainable grammars for speech recognition. In *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, pages 547–550, 1979.
4. A. Böck, K. Forchhammer, J. Heider, and C. Baron. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem Sci*, 16(12):463–467, 1991.
5. Michael P.S. Brown. Small subunit ribosomal RNA modeling using stochastic context-free grammars. In *ISMB 2000*, pages 57–66, 2000.
6. M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, supplement 3, pages 345–352. National Biomedical Research Foundation, Washington, DC, 1978.
7. Mathews DH, Sabina J, Zuker M, and Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, 1999.
8. S. R. Eddy and R.Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.
9. P. J. Farabaugh. Programmed translational frameshifting. *Microbiology and Molecular Biology Reviews*, 60(1):103–134, 1996.
10. D. P. Giedroc, C. A. Theimer, and P. L. Nixon. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *Journal of Molecular Biology*, 298(2):167–186, 2000.
11. Leslie Grate. Automatic RNA secondary structure determination with stochastic context-free grammars. In *Proc. of the 3th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB'95)*, pages 136–144, Cambridge, England, 1995. AAAI Press.
12. Leslie Grate, Mark Herbster, Richard Hughey, David Haussler, Saira Mian, and Harry Noller. RNA modeling using gibbs sampling and stochastic context free grammars. In *Proc. of the 2th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB'94)*, pages 138–146, Stanford, California, 1994. AAAI Press.

13. Stephane Hazebrouck, Luc Camoin, Zehava Faltin, Arthur Donny Strosberg, and Yuval Eshdat. Substituting selenocysteine for catalytic cysteine 41 enhances enzymatic activity of plant phospholipid hydroperoxide glutathione peroxidase expressed in *escherichia coli*. *Journal of Biological Chemistry*, 275(37):28715–28721, 2000.
14. I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125:167–188, 1994.
15. Alexander Hüttenhofer and August Böck. RNA structures involved in selenoprotein synthesis. In R. W. Simons and M. Grunberg-Manago, editors, *RNA Structure and Function*, pages 903–639. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, 1998.
16. Gorodkin J, Heyer LJ, and Stormo GD. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*, 25(18):3724–32, 1997.
17. Gorodkin J, Stricklin SL, and Stormo GD. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res*, 29(10):2135–44, 2001.
18. Stefanie J. Klug, Alexander Hüttenhofer, and Michael Famulok. In vitro selection of RNA aptamers that bind special elongation factor SelB, a protein with multiple RNA-binding sites, reveals one major interaction domain at the carboxyl terminus. *RNA*, 5:1180–1190, 1999.
19. Stefanie J. Klug, Alexander Hüttenhofer, Matthias Kraomayer, and Michael Famulok. In vitro and in vivo characterization of novel mRNA motifs that bind special elongation factor SelB. *Proc. Natl. Acad. Sci. USA*, 94(13):6676–6681, 1997.
20. G. V. Kryukov, V. M. Kryukov, and V. N. Gladyshev. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J Biol Chem*, 274(48):33888–33897, 1999.
21. K. Lari and S.J. Young. Applications of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 5:237–257, 1991.
22. Z. Liu, M. Reches, I. Groisman, and H. Engelberg-Kulka. The nature of the minimal 'selenocysteine insertion sequence' (secis) in escherichia coli. *Nucleic Acids Research*, 26(4):896–902, 1998.
23. Susan C. Low and Marla J. Berry. Knowing when not to stop: selenocysteine incorporation in eukaryotes. *Trends in Biochemical Sciences*, 21(6):203–208, 1996.
24. Zuker M and Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48, 1981.
25. Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, July 1978.
26. Yasubumi Sakakibara, Michael Brown, Richard Hughey, I. Saira Mian, Kimmen Sjolander, Rebecca C. Underwood, and David Haussler. Recent methods for RNA modeling using stochastic context-free grammars. In *Proc. 5th Symp. Combinatorical Pattern Matching*, 1994.
27. Yasubumi Sakakibara, Michael Brown, Richard Hughey, I. Saira Mian, Kimmen Sjolander, Rebecca C. Underwood, and David Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22(23):5112–5120, 1994.
28. C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment - RNA ligands to bacteriophage-T4 DNA-polymerase. *Science*, 249(4968):505–510, 3 1990.