ALGORITHMS FOR THE PREDICTION OF RNA SECONDARY STRUCTURES

M. Elloumi

Computer Science Department, Faculty of Economic Sciences and Management of Tunis, Tunisia

e-mail: Mourad.Elloumi@fsegt.rnu.tn

В данной работе задача прогнозирования вторичной структуры макромолекул ДНК решается с помощью “энергетических расчетов”. Излагаются алгоритм динамического программирования для вычисления свободных энергий стабильных вторичных структур и отслеживающий алгоритм для прогнозирования этих структур. Свободные энергии стабильных вторичных структур рассчитываются с использованием нового подхода, получившего название “m-многоконтурный подход” (m-MA), $m > 1$. Вычисление выполняется за время, пропорциональное n^4, и требует объема памяти, пропорционального n^2. Прогонозирование стабильных вторичных структур выполняется за время, пропорциональное $n^3 * \log_3(n)$. В сравнении с другими подходами алгоритм (m-MA) позволяет улучшить оценку минимальных энергетических вкладов множества контуров, что уточняет оценку свободных энергий стабильных вторичных структур.

Introduction

In molecular biology, a macromolecule can be coded by a string called *primary structure*. Each character in this string codes a constituent of the macromolecule. For the *RiboNucleic Acid* (RNA), these constituents, called *bases*, are *Adenine*, *Cytosine*, *Guanine* and *Uracil*. They are coded respectively by the characters A, C, G and U. Under some thermodynamic conditions, some regions of the macromolecule interact, thus creating folds within the macromolecule. These interactions are expressed at the level of the primary structure by pairings between the different substrings coding the regions that interact. The primary structure of the macromolecule provided with these pairings is called *secondary structure*. It is easy then to imagine that a macromolecule, represented by its primary structure, can have many secondary structures. However, only one of these structures is *stable*: it is the one that has the minimum *free energy*. The knowledge of this structure plays an important role, not only, to determine the interactions of the macromolecule with the *DesoxyriboNucleic Acid* (DNA) and the proteins, but also, to know its functions and its *biochemical* activities [9, 24].

Purely experimental methods, such as *X-ray diffraction* and *Nuclear Magnetic Resonance* (NMR) [18, 19], used to determine the secondary structures of RNA macromolecules are costly, require a long experimentation time and are practicable only for small molecules (few tens of bases). We resort then to a technique, called *prediction by energy computation*, which is both *experimental* and *algorithmic*: the estimation of the energetic contribution generated by a given pairing or by a *loop* is made from experimental results [11, 12, 26, 27, 21, 16], whereas, the choice of the pairings to keep in order to have the stable secondary structure is made through an algorithm [20, 30, 14, 17, 2, 6, 23, 31, 13, 7, 8]. We distinguish two types of algorithms to predict secondary structures of RNA macromolecules:

(*i*) Either algorithms adopting a *regions* approach: we establish the list of all the substrings (*regions*) that can be paired with each other, while respecting the thermodynamic laws. Then, from the different combinations of the unoverlapped pairings, we establish the list of all the possible secondary structures. For each secondary structure, we compute its free energy by using the experimental results [11, 12, 26, 27, 21, 16]. The structure that has the minimum free energy is the *true* secondary structure of the macromolecule.

The algorithms that adopt this approach are, unfortunately, costly in computing time. Among these algorithms, we cite the one of Pipas and Mc Mahon [20], the one of Studnicka et al. [25], the one of Martinez [14] and the one of Dumas and Ninio [6]. The algorithm of Pipas and Mc Mahon is the first algorithm to be used to predict secondary structures of RNA macromolecules. Its computing time complexity is $O(2^n)$, where n is the size of the string [22].

(*ii*) Or algorithms adopting a *dynamic programming* approach [3, 4]. These algorithms have been developed either under the *Hypothesis of Linearity of Energy* (HLE) or under the *Hypothesis of Loops Dependent Energy* (HLDE) [23, 31]. Using these algorithms, we proceed in two steps:

During the first step, we compute the energy of the stable secondary structure associated with the concerned string (primary structure): the computation of the energies of the stable secondary structures associated with longer substrings is made by using the computations results of the energies of the stable secondary structures associated with shorter substrings. We reiterate this process until the energy of the stable secondary structure associated with the whole string is computed.

During the second step, we *predict* the pairings that generate the stable secondary structure associated with the concerned string: the prediction of the pairings that generate the stable secondary structures associated with shorter substrings is made according to the pairings that generate the stable secondary structures associated with longer substrings. We reiterate this process until the prediction of the stable secondary structure associated with the whole string is ended.

The algorithms that adopt this approach are less costly. Among these algorithms, we cite the one of Waterman and Smith [30], the one of Nussinov and Jacobson [17], the one of Auron et al. [2], those of Sankoff et al. [23] and those of Elloumi [7, 8]. The order of computing time complexity of these algorithms varies between $O(n^3)$ and $O(n^4)$, where n is the length of the string.

In this paper, we present under the HLDE our dynamic programming algorithm to compute the free energies of the stable secondary structures and our traceback algorithm to predict these structures. We compute the free energies of the stable secondary structures thanks to a new approach called *m-Multiloop Approach* (m-MA), where $m > 1$. This computation is achieved within a time proportional to n^4 and using a memory space proportional to n^2. The prediction of the stable secondary structures is achieved within a time proportional to n^3*$\log_3(n)$. Compared to other approaches, the m-MA enables us to improve the estimation of the minimum energetic contributions of the *multiloops.* And hence, it enables us to improve the estimation of the free energies of the stable secondary structures. The other approaches, either ignore the energetic contributions of the multiloops, or compute these contributions under the HLE.

In the first section of this paper, we present, on one hand, a formal definition of a *secondary structure* and of its different kinds of *loops*, on the other hand, we define the *free energy* and the *loop energy* associated with a substring.

In the second section, we show how we represent a secondary structure and its different loops.

In the third section, we present the different equations of energies computation.

In the fourth section, we present, under the HLDE, our dynamic programming algorithm to compute the free energies of the stable secondary structures and our traceback algorithm to predict these structures.

Finally, in the last section, we present our conclusion.

1. Definitions and notations

Let $\mathcal{A}$ be a finite alphabet, a *string* is an element of $\mathcal{A}^*$, it is a concatenation of elements of $\mathcal{A}$. The *length* of a string w, denoted by $|w|$, is the number of the characters that constitute this string. By convention, the null length string will be denoted by ε. A string w of length n will be denoted by $w_{1,n}$ and the i^{th} character of w, $1\leq i \leq n$, will be denoted by w^i. A portion of w that begins at the position i and ends at the position j, $1 \leq i \leq j \leq n$, is called *substring* of w and will be denoted by $w_{i,j}$. By convention, when $j < i$ we will set $w_{i,j} = \varepsilon$. When $i = 1$ and $1 \leq j \leq n$ then the substring $w_{1,j}$ is called *prefix* of w and when $1 \leq i \leq n$ and $j = n$ then the substring $w_{i,n}$ is called *suffix* of w. The *primary structure* of an RNA macromolecule is a string which characters belong to the alphabet $\mathcal{A}_{RNA} = \{A, C, G, U\}$.

Let w be a primary structure of an RNA macromolecule, the set $\{w^i, w^{i+1}, \ldots, w^j\}$, $0 < i \leq j \leq |w|$, of the characters making up a substring $w_{i,j}$ of w will be denoted by $\mathcal{C}(w_{i,j})$. We define on $\mathcal{C}(w)$ a *pairing* relation, denoted by $\leftrightarrow$, satisfying the following properties:

(i) If $w^i \leftrightarrow w^j$ then $(j - i) \geq 4$.

(ii) If $w^i \leftrightarrow w^j$ then $w^i = A$ and $w^j = U$, or $w^i = U$ and $w^j = A$, or $w^i = C$ and $w^j = G$, or $w^i = G$ and $w^j = C$, or $w^i = G$ and $w^j = U$, or $w^i = U$ and $w^j = G$. The pair $\{w^i, w^j\}$ is called *Watson-Crick Pair* (WCP).

(iii) If $w^i \leftrightarrow w^j$ then for any k, $k \in [1..i-1] \bigcup [i+1..j-1] \bigcup [j+1..|w|]$, we can have neither $w^i \leftrightarrow w^k$ nor $w^j \leftrightarrow w^k$.

(iv) For any couples (i, i') and (j, j'), $i' \in]i..j[$ and $j' \in [1..i[\bigcup]j..|w|]$, if we have $w^i \leftrightarrow w^j$ then we cannot have $w^{i'} \leftrightarrow w^{j'}$.

A *secondary structure* associated with a primary structure w and a pairing relation $\leftrightarrow$, defined on $\mathcal{C}(w)$, is the set $S(w, \leftrightarrow) = \{(w^i,\ w^j)|w^i \leftrightarrow w^j \text{ and } 0 < i < j \leq |w|\}$. The empty secondary structure will be denoted by ω. A subset $S(w_{i,j}, \leftrightarrow, 0 < i < j \leq |w|$, of $S(w, \leftrightarrow)$ such that $S(w_{i,j}, \leftrightarrow) = \{(w^p, w^q)|w^p \leftrightarrow w^q \text{ and } 0 < i \leq p < q \leq j \leq |w|\}$ is called *substructure* of $S(w, \leftrightarrow)$.

With each secondary structure $S(w, \leftrightarrow)$ we associate a negative weight, denoted by $E(w, \leftrightarrow)$, called *free energy* of the structure $S(w, \leftrightarrow)$. The function E is called *energetic function.* The secondary structure for which this energy is minimum is called *stable* secondary structure of the macromolecule. It will be denoted by $S_{\min}(w)$ and its free energy will be denot ed by $E_{\min}(w)$:

$$E_{\min}(w) = \begin{cases} \min_{\leftrightarrow}\{E(w, \leftrightarrow)\}, & \text{if } \exists \leftrightarrow \text{ on } \mathcal{C}(w), \\ E''(w) & \text{else.} \end{cases} \tag{1}$$

The function E'' is an energetic function dependent solely on the nature of the bases that constitute the string w. By convention, we will set $E''(\varepsilon) = 0$.

Let us consider now a substring $w_{i,j}$, $0 < i < j \leq |w|$, the *loop energy*, denoted by $E_{\text{loop}}(w_{i,j})$, associated with the substring $w_{i,j}$ is the minimum free energy that can have a secondary structure of $w_{i,j}$ containing the couple $(w^i$,$w^j)$:

$$E_{\text{loop}}(w_{i,j}) = \begin{cases} \min_{\leftrightarrow|w^i \leftrightarrow w^j}\{E(w_{i,j} \leftrightarrow)\}, & \text{if } \exists \leftrightarrow \text{ on } \mathcal{C}(w_{i,j})|w^i \leftrightarrow w^j, \\ +\infty & \text{else.} \end{cases} \tag{2}$$

Each secondary structure $S(w, \leftrightarrow)$ can be subdivided in a unique way in a certain number of *loops*. We distinguish five types of loops:

(*i*) If $w^i \leftrightarrow w^j$ and the bases $w^{i+1}, w^{i+2}, \ldots, w^{j-1}$ are not paired then the singleton $\eta_{i,j}(w) = \{(w^i, w^j)\}$ is called *hairpin loop.*

(*ii*) If $w^i \leftrightarrow w^j$, $w^{i+1} \leftrightarrow w^{j-1}, \ldots, w^{i+k} \leftrightarrow w^{j-k}$, with $k \geq 1$, then the set $\sigma^k_{i,j}(w) = \{(w^i, w^j), (w^{i+1}, w^{j-1}), \ldots, (w^{i+k}, w^{j-k})\}$ is called *stack.*

(*iii*) If $w^i \leftrightarrow w^j$ and $w^{i+k} \leftrightarrow w^{j-1}$ (resp. $w^i \leftrightarrow w^j$ and $w^{i+1} \leftrightarrow w^{j-k}$), with $i + 1 < i + k < j - 1$ (resp. $i + 1 < j - k < j - 1$), and the bases $w^{i+1}, w^{i+2}, \ldots, w^{i+k-1}$ (resp. $w^{j-k+1}, w^{j-k+2}, \ldots, w^{j-1}$) are not paired then the pair $\lambda^k_{i,j}(w) = \{(w^i, w^j), (w^{i+k}, w^{j-1})\}$ (resp. $\rho^k_{i,j}(w) = \{(w^i, w^j), (w^{i+1}, w^{j-k})\}$) is called *left bulge loop* (resp. *right bulge loop*).

(*iv*) If w^i, w^j and $w^{i+l} \leftrightarrow w^{j-m}$, with $i + 1 < i + l < j - m < j - 1$, and the bases $w^{i+1}, w^{i+2}, \ldots, w^{i+l-1}$ and $w^{j-m+1}, w^{j-m+2}, \ldots, w^{j-1}$ are not paired then the pair $\zeta^{l,m}_{i,j}(w) = \{(w^i, w^j), (w^{i+l}, w^{j-m})\}$ is called *interior loop.*

(*v*) If $w^i \leftrightarrow w^j$, $w^{i+k_1} \leftrightarrow w^{i+l_1}$, $w^{i+k_2} \leftrightarrow w^{i+l_2}, \ldots, w^{i+k_m} \leftrightarrow w^{i+l_m}$, with $i < i+k_1 < i+l_1 < i+k_2 < i+l_2 < \ldots < i+k_m < i+l_m < j$, and for any k, $k \in]i..i+k_1[\bigcup]i+l_1..i+k_2[\bigcup\ldots\bigcup]i+l_m..j[$, we have w^k is not paired then the set $\mu^{k_1,l_1,\ldots,k_m,l_m}_{i,j}(w) = \{(w^i, w^j), (w^{i+k_1}, w^{i+l_1}),$ $(w^{i+k_2}, w^{i+l_2}), \ldots, (w^{i+k_m}, w^{i+l_m}\}$ is called *multiloop.* The couples (k_1, l_1), (k_2, l_2), $\ldots$, (k_m, l_m) generate together m branches that is why $\mu^{k_1,l_1,\ldots,k_m,l_m}_{i,j}(w)$ is also called *m-multiloop.*

Each one of these loops is said to be *closed* by the couple (w^i, w^j). The set of loops that constitute a secondary structure $S(w, \leftrightarrow)$ will be denoted by $\mathcal{L}(w, \leftrightarrow)$ and the set of loops that constitute the stable secondary structure $S_{\min}(w)$ will be denoted by $\mathcal{L}_{\min}(w)$.

Let l_i be one of the loops of a secondary structure $S(w, \leftrightarrow)$. An unpaired base w^k of w is said to be *accessible* from l_i, if and only if, there is a couple $(w^p$,$w^q)$ of l_i such that $p < k < q$ and there is no other couple (w^l, w^m) belonging to $S(w, \leftrightarrow)$ but not belonging to l_i such that $p < l < k < m < q$.

The set of non accessible bases from any loop of $\mathcal{L}(w, \leftrightarrow)$ is called the *tail* of the secondary structure $S(w, \leftrightarrow)$ and will be denoted by $\tau(w, \leftrightarrow)$. The tail of the stable secondary structure $S_{\min}(w)$ will be denoted by $\tau_{\min}(w)$.

Let $w_{i,j}$ be a substring and let $\mathcal{S}_\eta(w_{i,j})$ be the set of the hairpin loops that can be defined thanks to the bases of $w_{i,j}$. We define on $\mathcal{S}_\eta(w_{i,j})$ a partial order relation, denoted by $\rhd$, such that for a couple $(\{(w^p$,$w^q)\}$, $\{(w^r$,$w^s)\})$ of $\mathcal{S}_\eta(w_{i,j}) \times \mathcal{S}_\eta(w_{i,j})$, we have $\{(w^p, w^q)\} \rhd \{(w^r, w^s)\}$, if and only if $p < q < r < s$. A list of hairpin loops $[\{(w^{p_1}, w^{q_1})\}, \{(w^{p_2}, w^{q_2})\}, \ldots, \{(w^{p_m}, w^{q_m}\}]$ ordered by the partial order relation $\rhd$ will be denoted by $\{(w^{p_1}, w^{q_1})\} \rhd \{(w^{p_2}, w^{q_2})\} \rhd \ldots \rhd \{(w^{p_m}, w^{q_m})\}$.

The *hairpin loops graph*, denoted by $G_\eta(w_{i,j}) = (V_\eta(w_{i,j}), E_\eta(w_{i,j}))$, associated with the substring $w_{i,j}$ is the directed graph such that $V_\eta(w_{i,j}) = \{(p - i + 1, q - i + 1)|\{(w^p, w^q)\} \in \mathcal{S}_\eta(w_{i,j})\}$ and $E_\eta(w_{i,j}) = \{((p - i + 1, q - i + 1), (r - i + 1, s - i + 1))|\{(w^p, w^q)\} \rhd \{(w^r, w^s)\}\}$.

2. Representation of a secondary structure and its loops

A secondary structure $S(w, \leftrightarrow)$ can be represented by an undirected graph G, $G = (V, E)$, such that $V = \mathcal{C}(w)$ and $E = S(w, \leftrightarrow)$. A loop l of $S(w, \leftrightarrow)$ defined on $\mathcal{C}(w_{i,j}), 0 < i < j \leq |w|$, can be represented by a subgraph G', G'=(V',E'), of G such that $V' = \mathcal{C}(w_{i,j})$ and $E' = l$. When the edges of G are represented by segments with equal lengths, we say that we have a *normal* representation of the structure $S(w, \leftrightarrow)$.

Fig. 1 is a normal representation of a secondary structure with its different loops.

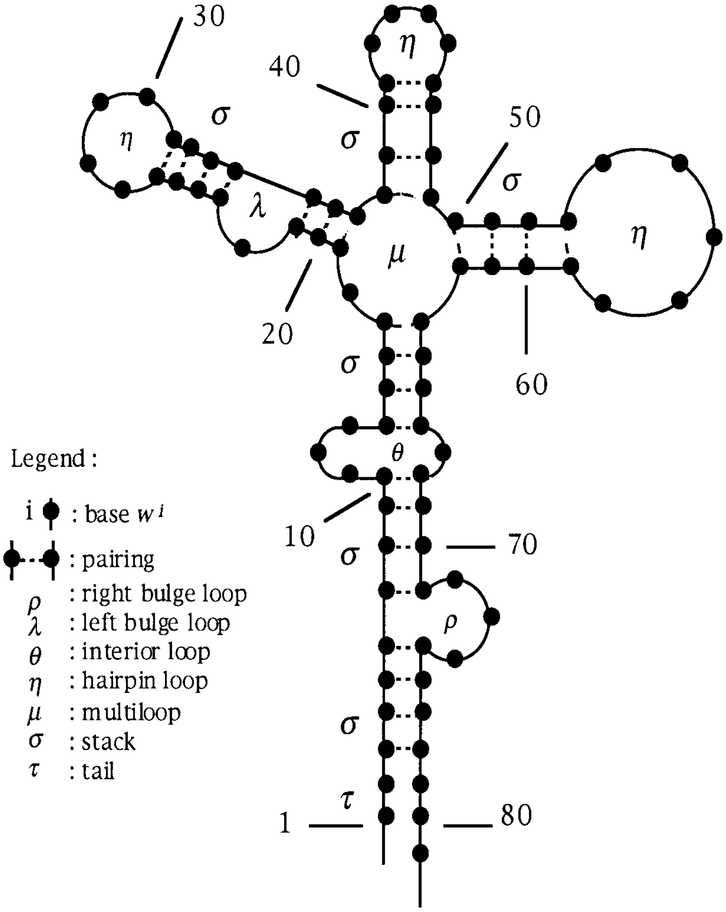

Fig. 1. Normal representation of a secondary structure.

3. Equations of energies computation

As we have explained in the introduction, to predict the stable secondary structure of an RNA macromolecule, we are brought to compute the minimum free energy that can have a secondary structure of this macromolecule. Unfortunately, the computing methods based on principles of thermodynamic do not permit to compute this energy. Whereas, it is experimentally possible to determine the energetic contribution of a WCP or a loop [11, 12, 26, 27, 21, 16]. A first hypothesis introduced by biochemists consists then in supposing that the energy of a secondary structure depends only on WCPs that constitute this structure [17]. We will call this hypothesis,

Hypothesis of Pairs Dependent Energy (HPDE):

$$E_{\min}(w) = \begin{cases} \min_{\leftrightarrow}\{ \sum\limits_{w^i \leftrightarrow w^j} e(w^i, w^j)\}, & \text{if } \exists \leftrightarrow \text{ on } \mathcal{C}(w), \\ 0 & \text{else,} \end{cases} \tag{3}$$

where e is a negative energetic function dependent solely on the nature of the concerned WCP.

Actually, the HPDE is a hypothesis that is far from being realistic. In fact, it ignores a very important fact: only stacks contribute with negative energies in the computation of the free energies of the secondary structures, the other loops contribute with positive energies [10]. And hence, only stacks tend to stabilize the secondary structures of the macromolecules, the other loops tend to destabilize them. Starting from this fact, biochemists have introduced later a better (more realistic) hypothesis. This hypothesis consists in supposing that the free energy of a secondary structure depends not only on the pairs that constitute this structure, but also, on the other unpaired bases of the macromolecule [31]. We will call this hypothesis, *Hypothesis of Loops Dependent Energy* (HLDE):

$$E_{\min}(w) = \begin{cases} \min_{\leftrightarrow} \left\{ \sum\limits_{l_i \in \mathcal{L}(w,\leftrightarrow)} E'(l_i) + \sum\limits_{w^r \in \tau(w,\leftrightarrow)} e'(w^r) \right\}, & \text{if } \exists \leftrightarrow \text{ on } \mathcal{C}(w), \\ E''(w) = \sum\limits_{s=1}^{|w|} e'(w^s) & \text{else.} \end{cases} \tag{4}$$

Function E' is an energetic function that depends on both the pairs that constitute the loop l_i and the accessible bases from this loop. The values of the energetic function E' are negative for the stacks and positive for the other loops [11, 12, 26, 27, 21, 16]. The function e' is a positive energetic function dependent only on the nature of the base in question.

In the particular case where for any loop l_i, $l_i \in \mathcal{L}(w, \leftrightarrow)$, we have:

$$E'(l_i) = \sum_{\substack{(w^p,w^q) \in l_i \\ \text{and}(w^p,w^q) \notin l_h, (h<i)}} e(w^p, w^q) + \sum_{\substack{w^s \text{ accessible} \\ \text{from } l_i}} e'(w^s), \tag{5}$$

we say that the function E is *linear*. We will call this hypothesis, *Hypothesis of Linearity of Energy* (HLE) [23].

It is easy to remark that the equation of $E_{\min}(w)$ under the HPDE is nothing else except a particular case of the one of $E_{\min}(w)$ under the HLE. Indeed, in the equation of $E_{\min}(w)$ under the HLE, we only have to set $e'(w^i) = 0$, for any i, $0 < i \leq |w|$, to find again the one of $E_{\min}(w)$ under the HPDE.

Our dynamic programming algorithm to compute the free energies of the stable secondary structures, and our traceback algorithm to predict these structures, under the HLDE, use the following theorems:

Lemma 1. For any substring $w_{i,j}$, $0 < i \leq j - 4 \leq |w|$, of a primary structure w and for any k, $i < k \leq j$, if $(w^i, w^k) \in S_{\min}(w_{i,j})$ then under the HLDE we have:

$$E_{\min}(w_{i,j}) = E_{\text{loop}}(w_{i,k}) + E_{\min}(w_{k+1,j}).$$

Proof. By definition, we have $E_{\min}(w_{i,j}) = \min_{\leftrightarrow}\{E(w_{i,j},\leftrightarrow)\}$. Considering that the couple $(w^i,w^k) \in S_{\min}(w_{i,j})$, we eliminate then all the substructures that do not contain this couple. Therefore, we have:

$$E_{\min}(w_{i,j}) = \min_{\leftrightarrow | w^i \leftrightarrow w^k} \{E(w_{i,j}, \leftrightarrow)\}.$$

Since the concerned substructures are those that contain the couple (w^i,w^k) then thanks to the HLDE we have:

$$E_{\min}(w_{i,j}) = \min_{\leftrightarrow|w^i\leftrightarrow w^k}\{E(w_{i,k},\leftrightarrow) + E(w_{k+1,j},\leftrightarrow)\}$$

i. e.:

$$E_{\min}(w_{i,j}) = \min_{\leftrightarrow|w^i\leftrightarrow w^k}\{E(w_{i,k},\leftrightarrow)\} + \min_{\leftrightarrow}\{E(w_{k+1,j},\leftrightarrow)\}.$$

Finally, thanks to Equations (1) and (2), we have:

$$E_{\min}(w_{i,j}) = E_{\text{loop}}(w_{i,k}) + E_{\min}(w_{k+1,j}).$$

■

Theorem 1. For any substring $w_{i,j}$, $0 < i \leq j-4 \leq |w|$, of a primary structure w, we have under the HLDE:

$$E_{\min}(w_{i,j}) = \begin{cases} \min\Big\{e'(w^i)+E_{\min}(w_{i+1,j}),\ \min\limits_{\substack{i+4\leq k\leq j\\ \text{such that}\\ w^i\leftrightarrow w^k}}\{E_{\text{loop}}(w_{i,k})+E_{\min}(w_{k+1,j})\}\Big\} & \text{if } \exists \leftrightarrow \text{ on } \mathcal{C}(w_{i,j}),\\ \sum\limits_{s=i}^{j} e'(w^s) & \text{else.}\end{cases}$$

Proof. If there are pairings on $\mathcal{C}(w_{i,j})$ then the secondary structures of $w_{i,j}$ are in one of the following cases:

(a) either there is no base w^k, $0 < i < k \leq j$, such that (w^i, w^k) belongs to this structure,

(b) or there is a base w^k, $0 < i < k \leq j$, such that (w^i,w^k) belongs to this structure.

If the stable secondary structure $S_{\min}(w_{i,j})$ is in the case (a) then we have obviously $S_{\min}(w_{i,j}) = S_{\min}(w_{i+1,j})$. Then we have too $\mathcal{L}_{\min}(w_{i,j}) = \mathcal{L}_{\min}(w_{i+1,j})$. On the other hand, under the HLDE we have:

$$E_{\min}(w_{i,j}) = \sum_{l_k\in\mathcal{L}_{\min}(w_{i,j})} E'(l_k) + \sum_{w^r\in\tau_{\min}(w_{i,j})} e'(w^r).$$

The base w^i is unpaired, then it belongs to $\tau_{\min}(w_{i,j})$. We can then rewrite $E_{\min}(w_{i,j})$ as follows:

$$E_{\min}(w_{i,j}) = \sum_{l_k\in\mathcal{L}_{\min}(w_{i,j})} E'(l_k) + \sum_{\substack{w^r\in\tau_{\min}(w_{i,j})\\ \text{and}\\ w^r\neq w^i}} e'(w^r) + e'(w^i).$$

Or, thanks to the equality between $\mathcal{L}_{\min}(w_{i,j})$ and $\mathcal{L}_{\min}(w_{i+1,j})$:

$$E_{\min}(w_{i,j}) = \sum_{l_k\in\mathcal{L}_{\min}(w_{i+1,j})} E'(l_k) + \sum_{w^r\in\tau_{\min}(w_{i+1,j})} e'(w^r) + e'(w^i).$$

Hence, thanks to Equation (4):

$$E_{\min}(w_{i,j}) = E_{\min}(w_{i+1,j}) + e'(w^i),$$

On the other hand, if the structure $S_{\min}(w_{i,j})$ is in the case (b) then let us call k_0 the position in $w_{i,j}$ such that $(w^i, w^{k_0}) \in S_{\min}(w_{i,j})$. According to Lemma 1, we have:

$$E_{\min}(w_{i,j}) = E_{\text{loop}}(w_{i,k_0}) + E_{\min}(w_{k_0+1,j}).$$

Since $E_{\min}(w_{i,j})$ is minimum, we have then:

$$E_{\min}(w_{i,j}) = \min_{i+4\leq k\leq j}\{E_{\text{loop}}(w_{i,k}) + E_{\min}(w_{k+1,j})\}.$$

Considering both cases (a) and (b), since we seek to minimize the value of the energy $E_{\min}(w_{i,j})$, we have then:

$$E_{\min}(w_{i,j}) = \min\{e'(w^i) + E_{\min}(w_{i+1,j}), \min_{i+4\leq k\leq j}\{E_{\text{loop}}(w_{i,k}) + E_{\min}(w_{k+1,j})\}\}.$$

Finally, from Equation (4), if there are no pairings on $\mathcal{C}(w_{i,j})$ then we have:

$$E_{\min}(w_{i,j}) = \sum_{s=i}^{j} e'(w^s).$$

■

Theorem 2. Let $w_{i,j}$, $0 < i \leq j-4 \leq |w|$, be a substring of a primary structure w such that the bases w^i and w^j can be paired with each other. Under the HLDE, we have:

$$E_{\text{loop}}(w_{i,j}) = \min\Big\{E'(\eta_{i,j}(w)), \min_{l}\{E'(\sigma^l_{i,j}(w)) + E_{\text{loop}}(w_{i+l,j-l})\},$$

$$\min\Big\{\min_{l}\{E'(\lambda^l_{i,j}(w)) + E_{\text{loop}}(w_{i+l,j-1})\}, \min_{l}\{E'(\rho^l_{i,j}(w)) + E_{\text{loop}}(w_{i+1,j-l})\}\Big\},$$

$$\min_{(l,m)}\{E'(\zeta^{l,m}_{i,j}(w)) + E_{\text{loop}}(w_{i+l,j-m})\}, \min_{(k_1,l_1,\ldots,k_m,l_m)}\{E'(\mu^{k_1,l_1,\ldots,k_m,l_m}_{i,j}(w)) + \sum_{s=1}^{m} E_{\text{loop}}(w_{i+k_s,i+l_s})\}\Big\}.$$

Proof. Let us denote $\leftrightarrow_{\min}$ the pairing on $\mathcal{C}(w_{i,j})$ such that $E(w_{i,j},\leftrightarrow_{\min})$ is minimum and $w^i \leftrightarrow_{\min} w^j$. We have $E_{\text{loop}}(w_{i,j}) = E(w_{i,j},\leftrightarrow_{\min})$. The loop q_0 of $S(w_{i,j},\leftrightarrow_{\min})$ that contains the couple $(w^i$,$w^j)$ is in one of the following cases:

(a) either it is a hairpin loop,
(b) or it is a stack,
(c) or it is a bulge loop,
(d) or it is an interior loop,
(e) finally, or it is a multiloop.

Let us examine each one of these cases.

Case (a):

In this case, we have obviously:

$$E_{\text{loop}}(w_{i,j}) = E'(\eta_{i,j}(w)).$$

Case (b): Let us call l_0 the position in $w_{i,j}$ such that $\sigma^{l_0}_{i,j}(w) = q_0$. Under the HLDE, we have:

$$E_{\text{loop}}(w_{i,j}) = E'(\sigma^{l_0}_{i,j}(w)) + E(w_{i+l_0,j-l_0},\leftrightarrow_{\min}).$$

Considering that the base w^{i+l_0} is paired with the base w^{j-l_0} (since $(w^{i+l_0}, w^{j-l_0}) \in \sigma^{l_0}_{i,j}(w)$) and the energy $E(w_{i+l_0,j-l_0},\leftrightarrow_{\min})$ is minimum (otherwise the energy $E_{\text{loop}}(w_{i,j})$ will not be minimum), we have then:

$$E(w_{i+l_0,j-l_0},\leftrightarrow_{\min}) = E_{\text{loop}}(w_{i+l_0,j-l_0}).$$

Hence, we have:

$$E_{\mathrm{loop}}(w_{i,j}) = E'(\sigma_{i,j}^{l_0}(w)) + E_{\mathrm{loop}}(w_{i+l_0,j-l_0}).$$

The energy $E'(\sigma_{i,j}^{l_0}(w)) + E_{\mathrm{loop}}(w_{i+l_0,j-l_0})$ is minimum, we have then:

$$E_{\mathrm{loop}}(w_{i,j}) = \min_{l}\{E'(\sigma_{i,j}^{l}(w)) + E_{\mathrm{loop}}(w_{i+l,j-l})\}.$$

The cases (c), (d) and (e) are processed in the same way as the case (b) and we have: Case (c):

$$E_{\mathrm{loop}}(w_{i,j}) = \min\{\min_l\{E'(\lambda_{i,j}^{l}(w)) + E_{\mathrm{loop}}(w_{i+l,j-1})\}, \min_l\{E'(\rho_{i,j}^{l}(w)) + E_{\mathrm{loop}}(w_{i+1,j-l})\}\}.$$

Case (d):

$$E_{\mathrm{loop}}(w_{i,j}) = \min_{(l,m)}\{E'(\zeta_{i,j}^{l,m}(w)) + E_{\mathrm{loop}}(w_{i+l,j-m})\}.$$

Case (e):

$$E_{\mathrm{loop}}(w_{i,j}) = \min_{(k_1,l_1,\ldots,k_m,l_m)}\{E'(\mu^{k_1,l_1,\ldots,k_m,l_m}(w)) + \sum_{s=1}^{m} E_{\mathrm{loop}}(w_{i+k_s,j+l_s})\}.$$

Considering all these cases together, since we seek to minimize the value of the energy $E_{\mathrm{loop}}(w_{i,j})$, we have then:

$$E_{\mathrm{loop}}(w_{i,j}) = \min\Big\{E'(\eta_{i,j}(w)), \min_{l}\{E'(\sigma_{i,j}^{l}(w)) + E_{\mathrm{loop}}(w_{i+l,j-l})\},$$

$$\min\Big\{\min_{l}\{E'(\lambda_{i,j}^{l}(w)) + E_{\mathrm{loop}}(w_{i+l,j-1})\}, \min_{l}\{E'(\rho_{i,j}^{l}(w)) + E_{\mathrm{loop}}(w_{i+1,j-l})\}\Big\},$$

$$\min_{(l,m)}\{E'(\zeta_{i,j}^{l,m}(w)) + E_{\mathrm{loop}}(w_{i+l,j-m})\}, \min_{(k_1,l_1,\ldots,k_m,l_m)}\{E'(\mu_{i,j}^{k_1,l_1,\ldots,k_m,l_m}(w)) + \sum_{s=1}^{m} E_{\mathrm{loop}}(w_{i+k_s,i+l_s})\}\Big\}.$$

■

We present now, under the HLDE, our dynamic programming algorithm, guided by a base-to-base pairing, to compute the energy of the stable secondary structure, then, we present our algorithm that predicts this structure by basing itself on tracing back the matrix filled by the previous algorithm. We use a new approach, called *m-Multiloop Approach* (m-MA), $m > 1$, that enables us to determine the m-multiloop that has the minimum energetic contribution. Thanks to this approach, the complexity of our prediction algorithm under the HLDE is reduced to a polynomial order.

In [8], we have also presented, under the HLE, our other algorithms of energies computation and prediction.

4. Energies computation and prediction of the stable secondary structure

According to Theorems 1 and 2, the computation of an energy $E_{\min}(w)$ under the HLDE depends on the one of the energetic contributions of the different loops defined on $\mathcal{C}(w)$.

The computation of the energetic contributions of stacks, hairpin, bulge and interior loops do not cause problems. In fact, there are experimental results [11, 12, 26, 27, 21, 16] concerning the energetic contributions of these loops. On the other hand, for a substring $w_{i,j}$ such that the bases w^i and w^j can be paired with each other, the number of these loops do not exponentially increase with the size of the substring $w_{i,j}$. Indeed, the number of stacks closed by the couple (w^i,w^j) is $O(j-i)$, the ones of bulge and interior loops are $O((j-i)^2)$, and finally, there is only one hairpin loop closed by the couple (w^i,w^j).

Concerning the computation of the energetic contributions of multiloops, two problems arise:

(i) The first one concerns the measure of the energetic contributions of these loops. In fact, the measures that have been made concern only multiloops with at most 8 *residues* [16]. Those which concern multiloops with more than 8 residues are very difficult to accomplish.

(ii) The second one concerns the number of possible multiloops associated with a given substring. Indeed, for a substring $w_{i,j}$, such that the bases w^i and w^j can be paired with each other, the number of these loops grows exponentially with the size of the substring $w_{i,j}$ [7] (Theorem 2.5, p25).

To overcome these two problems, we suggest the following approach, which we will call *m-Multiloop Approach* (m-MA), $m > 1$, that enables us to determine the m-multiloop that has the minimum energetic contribution. In what follows, we describe our approach for $m = 3$:

If the concerned substring, let us call it $w_{i,j}$, is of a length more than a certain threshold t_{sz}, $t_{sz} << |w|$, we operate by a *divide-and-conquer* strategy [1] we locate a couple (w^l, w^m), $i < l < m < j$, such that, on one hand, the bases w^l and w^m can be paired with each other and, on the other hand, the loop energy $E_{\text{loop}}(w_{l,m})$ is minimum. This couple divides the substring $w_{i+1,j-1}$ in two other smaller substrings: $w_{i+1,l-1}$ and $w_{m+1,j-1}$. Then we process the substring $w_{i+1,l-1}$ (resp. $w_{m+1,j-1}$), in the same way as we have processed the substring $w_{i,j}$: we locate a couple (w^p,w^q) (resp. (w^r,w^s)), $i+1 \leq p < q \leq l-1$ (resp. $m+1 \leq r < s \leq j-1$), such that, on one hand, the bases w^p and w^q (resp. w^r and w^s) can be paired with each other and, on the other hand, the loop energy $E_{\text{loop}}(w_{p,q})$ (resp. $E_{\text{loop}}(w_{r,s})$) is minimum.

The pairings between the bases w^p and w^q, w^l and w^m, w^r and w^s, and w^i and w^j, generate together a 3-multiloop.

If the substring $w_{i,j}$ is of a length less than the threshold t_{sz}, we identify all the multiloops defined on $\mathcal{C}(w_{i,j})$ and closed by the couple (w^i,w^j). The energetic contribution of one of these loops is estimated thanks to Ninio's experimental results [16].

By adopting the 3-MA, the loop energy $E_{\text{loop}}(w_{i,j})$ is defined by the following equation:

$$E_{\text{loop}}(w_{i,j}) = \min \Big\{ E'\eta_{i,j}(w)), \min_{l}\{E'(\sigma^l_{i,j}(w)) + E_{\text{loop}}(w_{i+l,j-l})\},$$

$$\min \Big\{ \min_{l}\{E'(\lambda^l_{i,j}(w)) + E_{\text{loop}}(w_{i+l,j-1})\}, \min_{l}\{E'(\rho^l_{i,j}(w)) + E_{\text{loop}}(w_{i+1,j-l})\} \Big\}, \tag{6}$$

$$\min_{(l,m)}\{E'(\zeta^{l,m}_{i,j}(w)) + E_{\text{loop}}(w_{i+l,j-})\}, E'_{\mu}(w_{i,j}) \Big\},$$

where $E'_{\mu}(w_{i,j})$ is the minimum energy that can have a substructure associated with the substring $w_{i,j}$, containing a multiloop closed by the couple (w^i, w^j):

$$E'_{\mu}(w_{i,j}) = \begin{cases} E_{\text{loop}}(w_{l_0,m_0}) + E_{\text{loop}}(w_{p_0,q_0}) + E_{\text{loop}}(w_{r_0,s_0}) + \sum\limits_{\substack{s \in]i..j[\setminus \\ ([p_0..q_0] \bigcup [l_0..m_0] \bigcup [r_0..s_0])}} e'(w^s) + e(w^i, w^j) \text{ if } (j-i) > t_{sz}, \\ \min_{(k_1,l_1,\ldots,k_m,l_m)}\{E'(\mu^{k_1,l_1,\ldots,k_m,l_m}(w)) + \sum_{s=1}^{m} E'(\eta_{i+k_s,i+l_s}(w))\} \text{ else.} \end{cases} \tag{7}$$

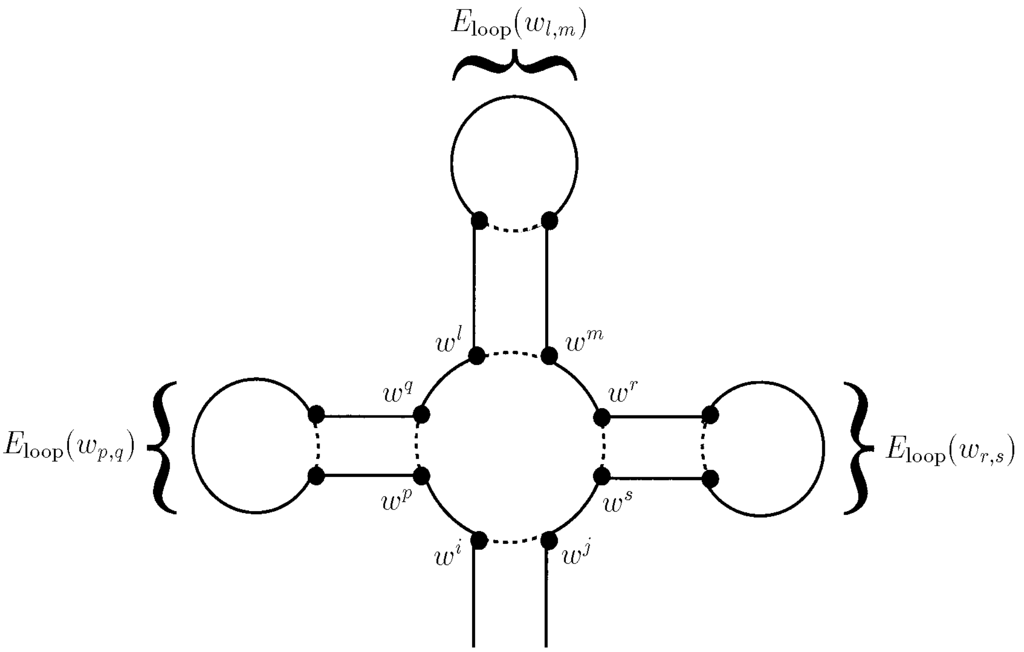
$E_{\text{loop}}(w_{l,m})$
w^l
w^m
w^q
w^r
$E_{\text{loop}}(w_{p,q})$
$E_{\text{loop}}(w_{r,s})$
w^p
w^s
w^i
w^j

Fig. 2. A secondary structure with a 3-multiloop.

where l_0, m_0, p_0, q_0, r_0 and s_0 are positions in $w_{i,j}$ such that:

$$\begin{cases} E_{\text{loop}}(w_{l_0,m_0}) = \min\limits_{i<l<m<j} \{E_{\text{loop}}(w_{l,m})\}, \\ E_{\text{loop}}(w_{p_0,q_0}) = \min\limits_{i<p<q<l_0} \{E_{\text{loop}}(w_{p,q})\}, \\ E_{\text{loop}}(w_{r_0,s_0}) = \min\limits_{m_0<r<s<j} \{E_{\text{loop}}(w_{r,s})\}. \end{cases} \tag{8}$$

Compared to other approaches, we mention among others the one of Waterman [28, 29], the one of Zuker and Stiegler [33] and the one of Sankoff et al. [23], the m-MA enables one to improve the estimation of the minimum energetic contributions of the multiloops. In fact:

(i) When the concerned substring is long, we compute in a polynomial time the minimum energy that can have an m-multiloop, where $m = 2$ for 5S rRNA, $m = 3$ for ^{t}RNA, 12S rRNA and 16S rRNA, and $m = 4$ for 23S rRNA [9, 24].

(ii) When the concerned substring is short, we compute the minimum energy that can have a multiloop, basing ourselves on Ninio's experimental results [16]. This enables us to compute more accurately the minimum free energy of the whole macromolecule.

The other approaches are:

(i) Either ignore the energetic contributions of multiloops, i. e. suppose that $E'_\mu(w_{i,j}) = 0$ for any couple (i,j), $0 < i \le j - 4 < |w|$. It is the case of Waterman's approach [28, 29] and the one of Zuker and Stiegler [33].

(ii) Or compute these contributions under the HLE. These approaches suppose, implicitly, that we have: $E'_\mu(w_{i,j}) = E_{\min}(w_{i+1,j-1}) + e(w^i,w^j)$ for any couple (i,j), $0 < i \le j - 4 < |w|$. It is the case of the approach of Sankoff *et al.* [23].

Certainly, the second type approaches are better than the first type ones, but they remain nevertheless unrealistic. Indeed, the linear functions that approximate the energetic contributions of the multiloops, give too large values in the case where these loops are too long or too short, whereas, the values that they give are too small in the case where these loops are of moderate lengths [23]. Our approach by its search of m-multiloops, in the case of long substrings, and its use of experimental results [16], in the case of short substrings, gives a more accurate estimation of the minimum energetic contributions of the multiloops.

On the other hand, by using the m-MA, the computation of the free energies of the stable secondary structures is achieved within a time proportional to n^4 and using a memory space proportional to n^2. These complexities are known to be the best existing complexities, to solve the problem of the prediction of the stable secondary structures of RNA macromolecules under the HLDE [16, 31, 29, 32, 15].

We present now our dynamic programming algorithm that computes the loop energies by using the 3-MA. We will need a half of a matrix M, of size $(|w|*|w|)$, to store the energetic values of the function E_{loop}. For any couple $(i{,}j)$, $0 < i < j \le |w|$, we will set:

$$M[j, i] := E_{\text{loop}}(w_{i,j}).$$

The other half of the matrix will be used to store the energetic values of the function $E_{\min}$. For any couple $(i{,}j)$, $0 < i < j \le |w|$, we will set:

$$M[i, j] := E_{\min}(w_{i,j}).$$

Algorithm 1

```
(i) (i.a) Construct a matrix $M$ of size $(|w|*|w|)$, such that, for any couple $(i, j)$, $0 < (j - i) < 4$, we have $M[j, i] := +\infty$;
(ii) for $j := 5$ to $|w|$ do
  for $i := j - 4$ downto $1$ do
    if $\{w^i{,}w^j\}$ is a WCP then
      (ii.a) $m_1 := E'(\eta_{i,j}(w))$;                                   {hairpin}
      (ii.b) $m_2 := \min_{\substack{1 \le l \le \lfloor (j-i-4)/2 \rfloor \\ \text{and} \\ \{w^{i+l}, w^{j-l}\} \text{ is a WCP}}} \{E'(\sigma^l_{i,j}(w)) + M[j - l, i + l]\}$;   {stack}
      (ii.c)
        (ii.c') $m_3 := \min_{\substack{2 \le l \le (j-i-5) \\ \text{and} \\ \{w^{i+l}, w^{j-1}\} \text{ is a WCP}}} \{E'(\lambda^l_{i,j}(w)) + M[j - 1, i + l]\}$;   {left bulge}
        (ii.c'') $m'_3 := \min_{\substack{2 \le l \le (j-i-5) \\ \text{and} \\ \{w^{i+1}, w^{j-l}\} \text{ is a WCP}}} \{E'(\rho^l_{i,j}(w)) + M[j - l, i + 1]\}$;   {right bulge}
        (ii.c''') $m_3 := \min\{m_3, m'_3\}$
      (ii.d) $m_4 := \min_{\substack{2 \le l \le (j-i-6) \\ 2 \le m \le (j-i-l-4) \\ \text{and} \\ \{w^{i+l}, w^{j-m}\} \text{ is a WCP}}} \{E'(\zeta^{l,m}_{i,j}(w)) + M[j - m, i + l]\}$;   {interior}
      (ii.e)                                                               {multiloop}
        if $(j - i) > t_{sz}$ then                                         {3-multiloop}
        (ii.e')
          (ii.e'.a.a) $m_5 := \min_{\substack{(i+1) \le l \le (j-5), \\ (l+4) \le m \le (j-1), \\ \text{and} \\ \{w^l, w^m\} \text{ is a WCP}}} \{M[m, l]\}$
          (ii.e'.a.b) give values to $l_0$ and $m_0$ such that $M[m_0, l_0] = m_5$
          (ii.e'.b.a)$m'_5 := \min_{\substack{(i+1) \le p \le (l_0-5), \\ (p+4) \le q \le (l_0-1), \\ \text{and} \\ \{w^p, w^q\} \text{ is a WCP}}} \{M[q, p]\}$   {$\{(w^p, w^q)\}$ is on the left of$\{(w^l, w^m)\}$}
```

```
                $(ii.e'.b.b)$ give values to $p_0$ and $q_0$ such that $M[q_0,p_0]{=}m'_5$
                $(ii.e'.c.a) m''_5 := \min_{\substack{(m_0+1)\leq r\leq(j-5),\\ (r+4)\leq s\leq(j-1),\\ \text{and}\\ \{w^r,w^s\} \text{ is a WCP}}} \{M[s,r]\};$      $\{\{(w^r,w^s)\}$ is on the right of$\{(w^l,w^m)\}\}$
                $(ii.e'.c.b)$ give values to $r_0$ and $s_0$ such that $M[s_0,r_0] = m''_5$
              $(ii.e'.d)$
                $m_5 := m_5 + m'_5 + m''_5 + \sum_{\substack{s\in]i..j[\backslash\\ ([p_0..q_0]\bigcup[l_0..m_0]\bigcup[r_0..s_0])}} e'(w^s) + e(w^i,w^j)$
            else                                                   {multiloop closed by$(w^i,w^j)$}
              $(ii.e'')\ m_5 := \min_{(k_1,l_1,\ldots,k_m,l_m)} \left\{E'(\mu_{i,j}^{k_1,l_1,\ldots,k_m,l_m}(w)) + \sum_{s=1}^{m} E'(\eta_{i+k_s,i+l_s}(w))\right\}$
            endif
            $(ii.f)\ M[j,i] := \min\{m_1,m_2,m_3,m_4,m_5\};$
            else $M[j,i] := +\infty$
        endif
    endfor
endfor
```

During the step $(ii.e'')$, we seek to construct all the multiloops defined thanks to the bases of the substring $w_{i+1,j-1}$ through searching for all the lists of hairpin loops defined, too, thanks to the bases of the substring $w_{i+1,j-1}$. The search of all these lists is made by identifying all the paths that exist in the graph $G_\eta(w_{i+1,j-1})$. Indeed, each vertex (k_s,l_s), in this graph, represents a hairpin loop $\{(w_{i+k_s},w_{i+l_s})\}$. A path $[(k_1,l_1),(k_2,l_2),\ldots,(k_m,l_m)]$, $m\geq 2$, represents then an ordered list $\{(w^{i+k_1},w^{i+l_1})\} \rhd \{(w^{i+k_2},w^{i+l_2})\} \rhd \ldots \rhd \{(w^{i+k_m},w^{i+l_m})\}$ made up by m hairpin loops. If the bases w_i and w_j can be paired with each other then this list and the couple (w^i,w^j) constitute together the multiloop $\mu_{i,j}^{k_1,l_1,\ldots,k_m,l_m}(w)$.

The search of all the paths linking two vertices in a graph is a well-known problem. In [5], Berge describes an algorithm that solves this problem.

Proposition 1. Let w be the primary structure of an RNA macromolecule. Algorithm 1 computes the loop energies $E_{\text{loop}}(w_{i,j})$, $0<i<j\leq|w|$, by using the 3-MA and we have $M[j,i]=E_{\text{loop}}(w_{i,j})$.

Proof. From Theorem 2 and Equations (6) and (7), Algorithm 1 computes the loop energies $E_{\text{loop}}(w_{i,j})$, $0<i<j\leq|w|$, by using the 3-MA and for any couple (i,j), $0<i<j\leq|w|$, if the bases w^i and w^j can be paired with each other then $M[j,i]=E_{\text{loop}}(w_{i,j})$, else $M[j,i]=+\infty$. ■

Proposition 2. Algorithm 1 is of complexities $O(|w|^4)$ in computing time and $O(|w|^2)$ in memory space.

Proof. For each couple (i,j), $0<i<j\leq|w|$, we make at most 1 iteration during the step $(ii.a)$, $\lfloor(j-i-4)/2\rfloor$ iterations during the step $(ii.b)$, 2*$(j-i-6)$ iterations during the step $(ii.c)$ and $(j-i-7)^2$ iterations during the step $(ii.d)$. The computations of the energetic contributions of a hairpin and a bulge loop are of complexity $O(|w|)$ in computing time. The ones of the energetic contributions of a stack and an interior loop are of complexity $O(1)$ in computing time [16]. Therefore, the computing time complexity of the steps $(ii.a)$ and $(ii.b)$ is $O(|w|)$, and the one of the steps $(ii.c)$ and $(ii.d)$ is $O(|w|^2)$.

Let us consider now the step $(ii.e)$:

(i) When $(j-i) > t_{sz}$, we go into the step $(ii.e')$: we make at most $(j-i-5)^2$ iterations during the step $(ii.e'.a)$, $(l_0-i-5)^2$ iterations during the step $(ii.e'.b)$ and $(j-m_0-5)^2$ iterations during the step $(ii.e'.c)$ $(i < l_0 < m_0 < j)$. Therefore, the step $(ii.e')$ is of complexity $O(|w|^2)$ in computing time.

(ii) When $(j-i) \leq t_{sz}$, we look for all the multiloops defined thanks to bases of the substring $w_{i,j}$ and closed by the couple (w^i,w^j). The computation of the energetic contribution of one of these multiloops is achieved within a time proportional to t_{sz} [16], $t_{sz} << |w|$. Then, the computing time of this step is bounded by the constant $v_\mu(t_{sz})^* t_{sz}$, where $v_\mu(t_{sz})$ is the maximum number of multiloops that can be defined thanks to bases of a substring $x_{1,t_{sz}}$ and closed by the couple $(x_1, x_{t_{sz}})$ [7] (Theorem 2.5, p25). Then, the step $(ii.e)$ is of complexity $O(|w|^2)$ in computing time.

We have $(|w|-4)^2/2$ couples (i,j) to process, since we must have $(j-i) \geq 4$, therefore, Algorithm 1 is of complexity $O(|w|^4)$ in computing time.

Algorithm 1 uses a memory space equal to $|w|^2/2$, then, it is of complexity $O(|w|^2)$ in memory space.

■

We present now our dynamic programming algorithm that computes the energies $E_{\min}(w_{i,j})$, $0 < i < j \leq |w|$, under the HLDE. This algorithm uses the matrix M whose second half has been filled thanks to Algorithm 1 ($M[j,i] := E_{\text{loop}}(w_{i,j})$, for any (i,j), $0 < i < j \leq |w|$).

Algorithm 2

```
(i) (i.a) Construct a matrix $M$ of size $(|w|^*|w|)$;
    (i.b) for any $(i,j)$, $0 \leq j-i < 4$, do
          $s := 0$;
          for $k := i$ to $j$ do $s := s + e'(w^k)$ endfor;
              $M[i,j] := s$
          endfor;

(ii) for $j := 5$ to $|w|$ do
     for $i := j-4$ downto 1 do

     (ii.a) $m_1 := e'(w^i) + M[i+1,j]$; $m_2 := +\infty$;
     (ii.b) for any $k$, $i+4 \leq k \leq j$, do
            if $\{w^i,w^k\}$ is a WCP then
                $m_2 := \min\{m_2, M[k,i] + M[k+1,j]\}$
            endif;
            endfor;
     (ii.c) $M[i,j] := \min\{m_1, m_2\}$;

     endfor;
     endfor;

(iii) $E_{\min}(w) := M[1,|w|]$.
```

Proposition 3. Let w be the primary structure of an RNA macromolecule. For any couple (i, j), $0 < i \leq j \leq |w|$, Algorithm 2 computes the energy $E_{\min}(w_{i,j})$ under the HLDE, by using the 3-MA, and we have $M[i, j] = E_{\min}(w_{i,j})$.

Proof. According to Theorem 1 and Proposition 1, for any couple (i, j), $0 < i \leq j \leq |w|$, Algorithm 2 computes the energy $E_{\min}(w_{i,j})$ under the HLDE, by using the 3-MA, and if the bases w^i and w^j can be paired with each other then $M[i, j] = E_{\min}(w_{i,j})$, else $M[i, j] = \sum\limits_{s=i}^{j} e'(w^s)$. ■

Proposition 4. Algorithm 2 is of complexities $O(|w|^3)$ in computing time and $O(|w|^2)$ in memory space.

Proof. For each couple (i, j), $0 < i < j \leq |w|$, we search for the position k_0, $i + 4 \leq k_0 \leq j$, such that $\{w^i, w^{k_0}\}$ is a WCP and $M[k_0, i] + M[k_0 + 1, j] = \min_{i+4 \leq k \leq j}\{M[k, i] + M[k + 1, j]\}$. This search is made linearly by incrementing the position k, $i + 4 \leq k \leq j$. Then, for a couple (i,j), this search is of complexity $O(|w|)$ in computing time. We have $(|w| - 4)^2/2$ couples (i, j) to process (since we must have $(j - i) \geq 4$). Therefore, Algorithm 2 is of complexity $O(|w|^3)$ in computing time.

Algorithm 2 uses a memory space equal to $|w|^2/2$ then it is of complexity $O(|w|^2)$ in memory space.

We present now Algorithm 4 which is our prediction algorithm under the HLDE. This algorithm traces back the matrix M filled thanks to Algorithms 1 and 2.

To construct the stable secondary structure $S_{\min}(w_{i,j})$ associated with a substring $w_{i,j}$, Algorithm 4 operates in the following way:

(*i*) When the base w^i is not paired with any other base of the substring $w_{i,j}$ then the secondary structure $S_{\min}(w_{i,j})$ is equal to the secondary structure $S_{\min}(w_{i+1,j})$. And we search then for the couples that constitute this structure by a recursive call to Algorithm 4.

(*ii*) Whereas, when the base w^i is paired with a base w^{k_0}, $0 < i < k_0 \leq j$, then, from Lemma 1, energy $E_{\min}(w_{i,j})$ satisfies the equation:

$$E_{\min}(w_{i,j}) = E_{\text{loop}}(w_{i,k_0}) + E_{\min}(w_{k_0+1,j}).$$

And then, the couple (w^i, w^{k_0}) belongs too to the substructure associated with the substring w_{i,k_0} and having a free energy equal to $E_{\text{loop}}(w_{i,k_0})$. The search of the couples that constitute this substructure is made by making a recursive call to Algorithm 3. This algorithm uses the 3-MA but in the *opposite direction*: we deduce the pairings associated with shorter substrings according to the pairings associated with the ends of longer substrings. The construction of the structure $S_{\min}(w_{k_0+1,j})$, associated with the substring $w_{k_0+1,j}$, is made by a recursive call to Algorithm 4.

We begin then by describing Algorithm 3, then, we describe Algorithm 4. The list L, used by Algorithm 3, represents the set of the couples that make up the stable secondary structure associated with the substring $w_{i,j}$. It is initialized to the empty list at the first call to this algorithm.

Algorithm 3 (i,j)

```
(i) (i.a) $m_2 := \min_{\substack{1 \leq l \leq \lfloor (j-i-4)/2 \rfloor \\ \text{and} \\ \{w^{i+l}, w^{j-l}\} \text{ is a WCP}}} \{E'(\sigma^l_{i,j}(w)) + M[j - l, i + l]\}$;   {stack}
    (i.b) give a value to $l_\sigma$ such that $E'(\sigma^{l_\sigma}_{i,j}(w)) + M[j - l_\sigma, i + l_\sigma] = m_2$;
(ii) (ii.a) $m_3 := \min_{\substack{2 \leq l \leq (j-i-5) \\ \text{and} \\ \{w^{i+l}, w^{j-1}\} \text{ is a WCP}}} \{E'(\lambda^l_{i,j}(w)) + M[j - 1, i + l]\}$;   {left bulge}
```

```
  (ii.b) give a value to $l_\lambda$ such that $E'(\lambda_{i,j}^{l_\lambda}(w)) + M[j-1, i+l_\lambda] = m_3$;
(iii) (iii.a) $m'_3 := \min_{\substack{2\le l\le(j-i-5)\\ \text{and}\\ \{w^{i+1},w^{j-l}\}\text{ is a WCP}}} \{E'(\rho_{i,j}^{l}(w)) + M[j-l, i+1]\}$;      {right bulge}
  (iii.b) give a value to $l_\rho$ such that $E'(\rho_{i,j}^{l_\rho}(w)) + M[j-l_\rho, i+1] = m'_3$;
(iv) (iv.a) $m_4 := \min_{\substack{2\le l\le(j-i-6)\\ 2\le m\le(j-i-l-4)\\ \text{and}\\ \{w^{i+l},w^{j-m}\}\text{ is a WCP}}} \{E'(\zeta_{i,j}^{l,m}(w)) + M[j-m, i+l]\}$;      {interior}
  (iv.b) give values to $l_\zeta$ and $m_\zeta$ such that $E'(\zeta_{i,j}^{l_\zeta,m_\zeta}(w)) + M[j-m_\zeta, i+l_\zeta] = m_4$;
(v)      { multiloop }
  (v.a) if $(j-i) > t_{sz}$, then      { 3-multiloop }
    (v.a.a) $m_5 := \min_{\substack{(i+1)\le l\le(j-5),\\ (l+4)\le m\le(j-1),\\ \text{and}\\ \{w^l,w^m\}\text{ is a WCP}}} \{M[m,l]\}$
    (v.a.b) give values to $l_\mu$ and $m_\mu$ such that $M[m_\mu, l_\mu] = m_5$;
    (v.a.c) $m'_5 := \min_{\substack{(i+1)\le p\le(l_\mu-5),\\ (p+4)\le q\le(l_\mu-1),\\ \text{and}\\ \{w^p,w^q\}\text{ is a WCP}}} \{M[q,p]\}$      $\{\{(w^p, w^q)\}$ is on theleft of$\{(w^l, w^m)\}\}$
    (v.a.d) give values to $p_\mu$ and $q_\mu$ such that $M[q_\mu, p_\mu] = m'_5$
    (v.a.e) $m''_5 := \min_{\substack{(m_\mu+1)\le r\le(j-5),\\ (r+4)\le s\le(j-1),\\ \text{and}\\ \{w^r,w^s\}\text{ is a WCP}}} \{M[s,r]\}$;      $\{\{(w^r, w^s)\}$ is on the right of$\{(w^l, w^m)\}\}$
    (v.a.f) give values to $r_\mu$ and $s_\mu$ such that $M[s_\mu, r_\mu] = m''_5$;
    (v.a.g) $m_5 := m_5 + m'_5 + m''_5 + \sum_{\substack{s\in]i..j[\setminus\\ ([p_\mu..q_\mu]\bigcup[l_\mu..m_\mu]\bigcup[r_\mu..s_\mu])}} e'(w^s) + e(w^i, w^j)$

  (v.b) else      { multiloop closed by $(w^i,w^j)$ }

    (v.b.a)

        $m_5 := \min_{(k_1,l_1,\ldots,k_m,l_m)} \left\{ E'(\mu_{i,j}^{k_1,l_1,\ldots,k_m,l_m}(w)) + \sum_{s=1}^{m} E'(\eta_{i+k_s,i+l_s}(w)) \right\}$

    (v.b.b) give values to $k_1^0, l_1^0, k_2^0, l_2^0, \ldots, k_m^0, l_m^0$ such that

        $E'(\mu_{i,j}^{k_1^0,l_1^0,\ldots,k_m^0,l_m^0}(w)) + \sum_{s=1}^{m} E'(\eta_{i+k_s^0,i+l_s^0}(w)) = m_5$

    endif;

  (vi) $L := L\bigcup\{w^i, w^j)\}$;
  (vii) case $M[j,i]$ of
    $m_2$ : $L := L\bigcup\{(w^{i+1}, w^{j-1})\}\bigcup\{(w^{i+2}, w^{j-2})\}\bigcup\ldots\bigcup\{(w^{i+l_\sigma-1}, w^{j-l_\sigma+1})\}$
        Algorithm 3 $(i+l_\sigma, j-l_\sigma)$;
    $m_3$: Algorithm 3 $(i+l_\lambda, j-1)$;
    $m'_3$: Algorithm 3 $(i+1, j-l_\rho)$;
    $m_4$: Algorithm 3 $(i+l_\zeta, j-m_\zeta)$;
    $m_5$: if $(j - i) > t_{sz}$ then
        Algorithm 3 $(l_\mu, m_\mu)$;
        Algorithm 3 $(p_\mu, q_\mu)$;
        Algorithm 3 $(r_\mu, s_\mu)$
    else
        $L := L \bigcup\{(w^{i+k_1^0}, w^{i+l_1^0})\} \bigcup\{(w^{i+k_2^0}, w^{i+l_2^0})\} \bigcup \ldots \bigcup\{(w^{i+k_m^0}, w^{i+l_m^0})\}$
  endif

endcase
```

Proposition 5. Let $w_{i,j}$ be a substring of the primary structure w such that the bases $w^i$and w^j can be paired with each other. And let M be the matrix filled thanks to Algorithm 1 ($M[j,i] := E_{\text{loop}}(w_{i,j})$ for any (i,j), $0 < i < j \leq |w|$). Algorithm 3 gives the substructure, associated with the substring $w_{i,j}$, containing the couple (w^i, w^j) and having a free energy equal to $E_{\text{loop}}(w_{i,j})$.

Proof. From Proposition 1, Equations (6) and (7), and Theorem 2, Algorithm 3 gives the substructure, associated with the substring $w_{i,j}$, containing the couple (w^i, w^j) and having a free energy equal to $E_{\text{loop}}(w_{i,j})$.

■

Proposition 6. Algorithm 3 is of complexity $O(|w|^2 * \log_3(|w|))$ in computing time.

Proof. We demonstrate in the same way as in the proof of Proposition 2 that:

(i) the step (i) of Algorithm 3 is of complexity $O(|w|)$ in computing time,

(ii) the steps (ii), (iii), (iv) and (v) are of complexity $O(|w|^2)$ in computing time.

Therefore, each call to Algorithm 3 is of complexity $O(|w|^2)$ in computing time. On the other hand, each call to Algorithm 3 generates, at most, three other recursive calls to Algorithm 3. Therefore, for a primary structure w, we have, at most, $\log_3(|w|)$ recursive levels. Hence, Algorithm 3 is of complexity $O(|w|^2 * \log_3(|w|))$ in computing time.

■

```
    Algorithm 4 $(i, j)$
$(i)$ if $(j - i) \geq 4$ then
    $(i.a)$ if $M[i,j] = e'(w^i) + M[i+1,j]$ then Algorithm 4 $(i+1, j)$
        else
    $(i.b)$   $k := i + 4$;
        while $(M[i,j] \neq M[k,i] + M[k+1,j])$ do $k := k + 1$ endwhile;
    $(i.c)$ Algorithm 3 $(i,k)$; Algorithm 4 $(k+1, j)$;
        endif;
  endif;
$(ii)$ $S_{\min}(w_{i,j}) := L$.
```

Proposition 7. Let $w_{i,j}$ be a substring of the primary structure w and let M be the matrix filled thanks to Algorithms 1 and 2. Algorithm 4 gives the substructure $S_{\min}(w_{i,j})$, under the HLDE, by using the 3-MA.

Proof. Propositions 1 and 3, and Theorem 1 guarantee that Algorithm 4 gives the substructure $S_{\min}(w_{i,j})$, under the HLDE, by using the 3-MA.

■

Proposition 8. Algorithm 4 is of complexity $O(|w|^3 * \log_3(|w|))$ in computing time.

Proof. Each call to Algorithm 4 generates an other call to this algorithm. Therefore, for a primary structure w, we make, at most, $|w|$ calls. During a call concerning a couple (i,j), $0 < i \leq j - 4 \leq |w|$:

(i) We look for the position k_0, $i+4 \leq k_0 \leq j$, such that $\{w_i, w_{k0}\}$ is a WCP and $M[i,j] = M[k_0,i] + M[k_0+1,j] = \min_{i+4 \leq k \leq j}\{M[k,i] + M[k+1,j]\}$. This search is of complexity $O(|w|)$ in computing time.

(ii) Then, we make a call to Algorithm 3. From Proposition 6, this algorithm is of complexity $O(|w|^2 * \log_3(|w|))$ in computing time.

Therefore, each call to Algorithm 4 is of complexity $O(|w|^2 * \log_3(|w|))$ in computing time. Hence, Algorithm 4 is of complexity $O(|w|^3 * \log_3(|w|))$ in computing time.

■

Its easy to verify that the complexities of our algorithms remain the same for $m \neq 3$, i. e., $m = 2$ or $m = 4$, where m is the number of the considered branches in a multiloop.

5. Experimental results

We have executed the program corresponding to our algorithm on strings coding ^{t}RNA, 5S rRNA, 12S rRNA, 16S rRNA and 23S rRNA macromolecules. We have been provided with these data by the *European Molecular Biology Laboratory* (Heidelberg, Germany).

The program is written in C and implemented on a SUN SPARCstation computer. *Tab. 1* shows the processed data sizes and the corresponding results, where θ is the success rate, it is defined by:

$$\theta = \frac{n_{\text{pred}}}{n_{\text{total}}} * 100. \tag{9}$$

With n_{pred} is the number of the predicted structures that are *close* to the true ones and n_{total} is the total number of the structures. We have measured the *closeness* of the predicted structures to the true ones by using the following rate:

$$\beta = \frac{n'_{\text{pred}}}{n'_{\text{total}}} * 100. \tag{10}$$

With n'_{pred} is the number of the *Watson-Crick Pairs* (WCP) in the true structure that were predicted and n'_{total} is the total number of the WCPs in the true structure.

Table 1

Experimental results

Macromolecule	Approximate String Length	Number of Strings Processed	m	$\theta, \%$
^{t}RNA	80	200	3	95.06
5S rRNA	120	150	2	94.68
12S rRNA	950	100	3	95.24
16S rRNA	1600	100	3	94.62
23S rRNA	2900	50	4	94.50

6. Conclusion

In this paper, we have tackled the problem of the *prediction by energy computation* of the stable secondary structures of RNA macromolecules. The algorithms that we have presented deal with this problem under the *Hypothesis of Loops Dependent Energy* (HLDE): we compute the free energies of the stable secondary structures by using a new approach called *m-Multiloop Approach* (m-MA), where $m > 1$. This computation is made in a time proportional to n^4 and using a memory space proportional to n^2. The prediction of the stable secondary structures is made within a time proportional to $n^3 * \log_3(n)$. Compared to other approaches, the m-MA enables to improve the estimation of the minimal energetic contributions of the multiloops. And hence, it enables to improve the estimation of the free energies of the stable secondary structures. The other approaches, either ignore the energetic contributions of the multiloops, or compute these contributions under the HLE.

Our prediction algorithm, under the HLDE, has predicted secondary structures *close* to the true ones with a success rate of the order of 95 %. This result is very interesting, when we know that the other algorithms, either do not reach this rate, or reach it with an exponential complexity [16, 31, 29, 32, 15].

References

[1] Aho A. V., Hopcroft J. E., Ullman J. D. The Design and Analysis of Computer Algorithms. Addison-Wesley Publishing Company, 1974. P. 60–65.

[2] Auron B. E., Rindowe W. P., Vary C. P. H. et. al. Computer aided prediction of RNA secondary structures // Nucleic Acids Research. 1982. Vol. 10, No. 1. P. 403–419.

[3] Bellman R. E. Dynamic Programming. New Jersey: Princeton Univ. Press, 1957.

[4] Bellman R. E., Dreyfus S. E. Applied Dynamic Programming. New Jersey: Princeton Univ. Press, 1962.

[5] Berge C. Graphes et Hypergraphes. Paris: Dunod Editeur, 1970.

[6] Dumas J. P., Ninio J. Efficient algorithms for folding and comparing nucleic acid sequences // Nucleic Acids Research. 1982. Vol. 10, No. 1. P. 197–206.

[7] Elloumi M. Analysis of Strings Coding Biological Macromolecules. Science Doctorate Dissertation. The Univ. of Aix-Marseilles III, France, June 1994.

[8] Elloumi M. New Algorithms to Predict Secondary Structures of RNA Macromolecules // 11^{th} Intern. Conf. on Industrial and Eng. Appl. of Artificial Intelligence and Expert Systems (Benicassim, Castellon, Spain), Springer Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, June 1998, P. 864–875.

[9] Gesteland R. F., Atkins J. F. (Eds.) The RNA World. N. Y.: Cold Spring Harbor Lab. Press, 1993.

[10] Gralla J., Steitz J. A., Crothers D. M. Direct physical evidence for secondary structure in an isolated fragment of R17 bacteriophage mRNA. 1974. No. 248. P. 204–208.

[11] Gralla J., Crothers D. M. Free energy of imperfect nucleic acid helices II. Small hairpin loops // J. Mol. Biol. 1973. No. 73. P. 497–511.

[12] Gralla J., Crothers D. M. Free energy of imperfect nucleic acid helices III. Small internal loops resulting from mismatches // J. Mol. Biol. 1973. No. 78. P. 301–319.

[13] Larmore L. L., Schieber B. On-line dynamic programming with applications to the prediction of RNA secondary structure // J. of Algorithms. 1991. No. 12. P. 490–515.

[14] Martinez H. A New Algorithm for Calculating RNA Secondary Structure. Manuscript, 1980.

[15] Meidanis J., Setubal J. C. Introduction to Computational Molecular Biology. Boston: PWS Publ. Company, 1997.

[16] Ninio J. Prediction of pairing schemes in RNA molecules-loop contributions and energy of wobble and non-wobble pairs // Biochimie. 1979. No. 61. P. 1133–1150.

[17] Nussinov R., Jacobson A. Fast algorithm for predicting the secondary structure of single-stranded RNA // Proc. Nat. Acad. Sci. USA. Nov. 1980. Vol. 77, No. 11. P. 6309–6313.

[18] Oppenheimer N. J., James T. L. Nuclear magnetic resonance: Part A, Spectral techniques and dynamics // Methods in Enzymology. 1989. No. 176.

[19] Oppenheimer N. J., James T. L. Nuclear magnetic resonance: Part B, Structure and mechanism // Ibid. No. 177.

[20] Pipas J. M., Mc Mahon J. E. Method for prediction RNA secondary structure // Proc. Nat. Acad. Sci. USA. June 1975. Vol.72, No. 6. P. 2017–2021.

[21] Salser W. Globin messenger-RNA sequences: analysis of base-pairing and evolutionary implications // Cold Spring Harbor Symp. Quant. Biol. 1977. No. 42. P. 985–1002.

[22] Sankoff D., Morin A. M., Cedergren R. J. The evolution of 5sRNA secondary structure // J. Canadien de Biochimie. 1978. No. 56. P. 440–443.

[23] Sankoff D., Kruskal J. B., Mainville S., Cedergren R. Fast algorithms to determine RNA secondary structures containing multipleloops. Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley Publ., Massachusetts. 1983. P. 93–120.

[24] Simons R. W., Grunberg-Manago M. (Eds.) RNA Structure and Function. N. Y.: Cold Spring Harbor Lab. Press, 1998.

[25] Studnicka G. M., Rahn G. M., Cummings I. W., Salser W. A. Computer method for predicting the secondary structure of single-stranded RNA // Nucleic Acids Research. 1978. Vol. 5, No. 9. P. 3356–3387.

[26] Tinoco I. Jr, Borer Ph. N., Dengler B. et. al. Improved estimation of secondary structure in ribonucleic acids // Nature New Biology. 1973. No. 246. P. 40–41.

[27] Uhlenbeck O. C., Borer Ph. N., Dengler B., Tinoco I. Stability of RNA hairpin loops: A_6-C_m-U_6 // J. Mol. Biol. 1973. No. 73. P. 483–496.

[28] Waterman M. S. Secondary structure of single-stranded nucleic acids // Studies in Foundations and Combinatorics, Advances in Mathematics Supplementary Studies. 1978. No. 1. P. 167–212.

[29] Waterman M. S. Introduction to computational biology / J. Wiley (Eds.), 1995.

[30] Waterman M. S., Smith T. F. RNA Secondary structure: A complete mathematical analysis // Mathematical Biosciences, 1978. No. 42. P. 257–266.

[31] Zuker M. The use of dynamic programming algorithms in RNA secondary structure prediction // Mathematical Methods for DNA Sequences, CRC Press Inc., Boca Raton, Florida, 1989. P. 168–170.

[32] Zuker M. Prediction of RNA Secondary Structure by Energy Minimization. Washington Univ. Press, St. Louis, Mo., 1996.

[33] Zuker M., Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information // Nucleic Acids Research. 1981. Vol. 9, No. 1. P. 133–148.

Received for publication March 27, 2001