# RNA Folding Algorithms with G-Quadruplexes -Supplemental Material-

Ronny Lorenz[1], Stephan H. Bernhart[2], Fabian Externbrink[2], Jing Qin[3], Christian Höner zu Siederdissen[1], Fabian Amman[1], Ivo L. Hofacker[1,4], and Peter F. Stadler[2,1,3,4,5,6]

[1]Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; [2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; [3]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany; [4]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark; [5]Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany; [6]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

## A Combinatorics of RNA Structures with Quadruplexes

Here we describe a more detailed combinatorial model of secondary structures with G-quadruplexes than the simplified version outlined in the main text.

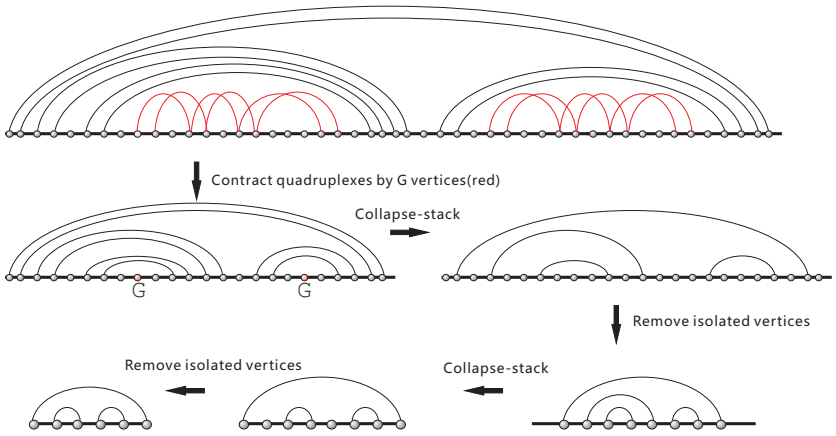
### A1 Model

A secondary structure of length $n$ is a noncrossing partial matching (matching with isolated vertices) such that each base pair is of length at least 3. For simplicity, we consider an arbitrary number of $L \geq 2$ stacked G-quartets and three linkers of length $l_1$, $l_2$ and $l_3 \geq 1$. We allow quadruplexes in any context with the following exception: If $(i, j)$ is a base pair that encloses a single quadruplexes, then at least one of three conditions is satisfied: (1) $i + 1$ and $j - 1$ are both unpaired; (2) $i + 1$, $i + 2$, $i + 3$ are unpairedm or (3) $j - 3$, $j - 2$, $j - 1$ are unpaired. We call such structures G-structures in the following.

A stack of length $\tau$ consists of exactly $\tau$ "parallel" arcs $((i, j), (i + 1, j - 1), \ldots, (i + (\tau - 1), j - (\tau - 1)))$. We say that a G-structure is $\tau$-canonical if all stacks consist of at least $\tau$ arcs.

The enumeration is based on the notion of shapes, that is, matchings in which each stack consists of exactly one arc. The shape of an arbitrary G-structure $s$ is obtained by (1) contracting each G-quadruplex to a single vertex labelled 'G', and (2) iteratively collapsing each stack to a single arc and then removing any

isolated vertices from the resulting diagram as in the following example:




### A2 Generating functions

Let $\mathbf{s}_{n,t}$ denote the number of all noncrossing shapes over $2n$ vertices with $t$ arcs of length 1 (1-arcs) and its corresponding generating function $\mathbf{S}(u,e) = \sum_n \sum_t \mathbf{s}_{n,t} u^n e^t$. Denote by $\mathbf{m}_t$ the number of noncrossing matchings with $t$ arcs. Note that $\mathbf{m}_{2t}$ is the well-known $t$-th Catalan number. Using the generating function $\mathbf{M}(u) = \frac{1-\sqrt{1-4u}}{2u}$ for the matchings we have

$$\mathbf{S}(u,e) = \frac{1+u}{1+2u-ue}\mathbf{M}\left(\frac{u(1+u)}{(1+2u-ue)^2}\right).$$

In the following we will make use of several auxiliary functions:

$$\mathbf{Q}(x) = \frac{x^{11}}{(1-x^4)(1-x)^3}, \qquad \mathbf{P}_0(x) = \frac{(1-x)\,\mathbf{Q}(x)}{1-x-x\mathbf{Q}(x)}, \qquad \mathbf{R}(x) = \frac{x^2}{(1-x)^2},$$

$$\mathbf{T}(x) = \frac{1}{(1-x)^2}, \qquad \mathbf{L}(x) = \frac{2x^3}{1-x}, \qquad \mathbf{P}_1(x) = \frac{x}{1-x} + \mathbf{T}(x)\mathbf{P}_0(x),$$

$$\mathbf{P}_2(x) = \frac{x^3}{1-x} + (\mathbf{R}(x)+\mathbf{L}(x))\mathbf{P}_0(x), \qquad \mathbf{P}_3(x) = \frac{1}{1-x} + \mathbf{T}(x)\mathbf{P}_0(x).$$

Our main result is

**Theorem 1.** *The generating function of $\tau$-canonical G-structures is*

$$\boldsymbol{G}^{\tau}(x) = \sum_n \boldsymbol{g}_n^{\tau} x^n = \boldsymbol{P}_3(x)\boldsymbol{S}\left(\frac{x^{2\tau}\cdot \boldsymbol{P}_3^2(x)}{(1-x^2)-x^{2\tau}(2\boldsymbol{P}_1(x)+\boldsymbol{P}_1^2(x))}, \frac{\boldsymbol{P}_2(x)}{\boldsymbol{P}_3(x)}\right).$$

*Proof.* We utilize the following combinatorial classes: $\mathcal{E}$ (neutral class, consisting of a single element of size 0), $\mathcal{Z}$ (vertices, with size 1), $\mathcal{U}$ (arcs, comprising two vertices thus having size 2), and $\mathcal{W}$ (quadruple arcs taking 4 vertices).

*Claim 1.* The generating function of the numbers $\mathbf{q}_n$ of quadruplexes on length $n$ is

$$\mathbf{Q}(x) = \frac{x^{11}}{(1-x^4)(1-x)^3}.$$

Let $\mathcal{Q}$ denote the combinatorial class of G-quadruplexes. By construction, each quadruplex consists of $L \geq 2$ stacked G-quartets and three linkers of length at least 1. Thus we have $\mathcal{Q} = \mathcal{W}^2 \times \mathbf{SEQ}(\mathcal{W}) \times (\mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}))^3$. This implies the Claim 1.

Denote by $\mathbf{p}_n$ the number of G-structures of length $n$ without base pairs outside quadruplexes with two additional restrictions: (1) its first and last vertices are part of a quadruplex and (2) if there exist two consecutive G-quartets, then there exists at least one isolated vertex between them.
*Claim 2.* The generating function of $\mathbf{p}_n$ is $\mathbf{P}_0(x) = \frac{(1-x)\,\mathbf{Q}(x)}{1-x-x\mathbf{Q}(x)}$.

We proceed by induction on the number of G-quadruplexes. Let $\mathbf{p}_n^k$ denote the number of single stranded secondary structures with $k$ G-quadruplexes of length $n$, then we have its corresponding generating function $\mathbf{P}_0^k(x) = \mathbf{Q}(x)^k \cdot (\frac{x}{1-x})^{k-1}$. The claim follows by summing all $k \geq 1$.
*Claim 3.* Let $\lambda$ be a fixed noncrossing shape with $s \geq 1$ arcs and $m \geq 0$ 1-arcs (arcs of length 1). Then the generating function of $\tau$-canonical G-structures containing arc length at least 3 that have shape $\lambda$ is given by

$$\mathbf{Q}_\tau^\lambda(x) = \mathbf{P}_3(x) \cdot \left(\frac{x^{2\tau} \cdot \mathbf{P}_3^2(x)}{(1-x^2) - x^{2\tau}(2\mathbf{P}_1(x) + \mathbf{P}_1^2(x))}\right)^s \left(\frac{\mathbf{P}_2(x)}{\mathbf{P}_3(x)}\right)^m$$

In order to prove the claim, we use the additional notations $\mathcal{Z}_{r/b/g}$ for red/blue/green vertices. We can construct an arbitrary $\tau$-canonical G-structure with arc-length at least 3 and shape $\lambda$ in the following way: Starting from the shape $\lambda$, insert at most one red isolated vertex into the $(2s+1)$ intervals except the interval $[i, i+1]$ for which that $(i, i+1)$ is an 1-arc in $\lambda$. The corresponding combinatorial class is $\mathcal{M}_1 = \mathcal{U}^s \times (\mathcal{E} + \mathcal{Z}_r)^{2s-m+1}$. Next insert exactly one green isolated vertex after each vertex $j$ such that $(j, j+1)$ forms an 1-arc in $\lambda$. This yields the class $\mathcal{M}_2 = \mathcal{M}_1 \times \mathcal{Z}_g^m$.

Next, we inflate each arc into a stack of size $t \geq 0$. In case of $t \geq 1$, between the arcs of the obtained stack we insert a blue isolated vertex to the left or the right, or on both sides in order to separate the arcs and for each such insertion exactly one blue isolated vertex is used. This results in the combinatorial class $\mathcal{M}_3$ from $\mathcal{M}_2$ by the substitution

$$\mathcal{U} \to \sum_{t \geq 1} \mathcal{U}^t \times (2\mathcal{Z}_b + \mathcal{Z}_b^2)^{t-1}.$$

Now we inflate each arc in the resulting structure into a stack of size at least $\tau$. The combinatorial class $\mathcal{M}_4$ results from $\mathcal{M}_3$ via the substitution $\mathcal{U} \to \frac{\mathcal{U}^\tau}{1-\mathcal{U}}$.

Next we inflate each red isolated vertex into either a sequence of isolated vertices of length at least one or a $\mathcal{P}_0$-structure $\vartheta_1$ in addition with two sequences of isolated vertices (at least 1) at both ends of $\vartheta_2$ or a sequence of isolated vertices

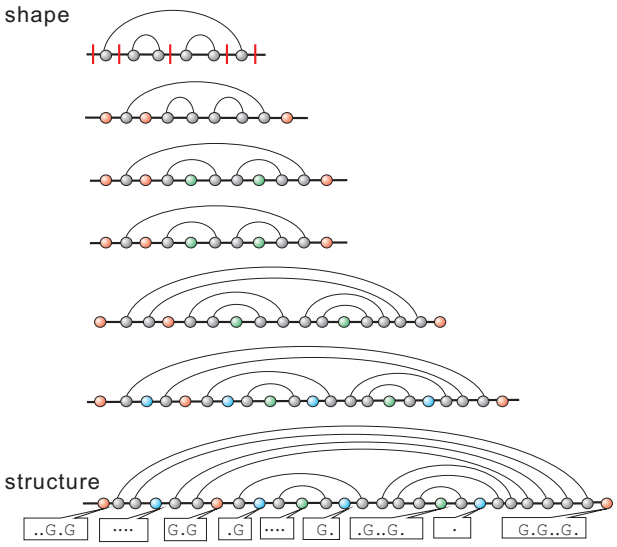



**Fig. 1.** From a shape to a secondary structure with G-quadruplexes. Each G-quadruplex is shown as a vertex labelled by G.

(at least 3) at one of the ends of $\vartheta_2$. The corresponding class $\mathcal{M}_6$ is symbolically obtained from $\mathcal{M}_5$ by the substitution $\mathcal{Z}_g \to \mathcal{Z}^3 \times \mathbf{SEQ}(\mathcal{Z}) + (\mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}))^2 \times \mathcal{P}_0 + 2\mathcal{Z}^3 \times \mathbf{SEQ}(\mathcal{Z}) \times \mathcal{P}_0$.

We then inflate each green isolated vertex into either a sequence of isolated vertices of length at least three or a $\mathcal{P}_0$-structure $\vartheta_2$ in addition with two sequences of isolated vertices (at least 1) at both ends of $\vartheta_2$. The corresponding class $\mathcal{M}_6$ is symbolically obtained from $\mathcal{M}_5$ by the substitution $\mathcal{Z}_g \to \mathcal{Z}^3 \times \mathbf{SEQ}(\mathcal{Z}) + (\mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}))^2 \times \mathcal{P}_0$.

We finally inflate each blue isolated vertex into either a sequence of isolated vertices of length at least one or a $\mathcal{P}_0$-structure $\vartheta_3$ in addition with two sequences of isolated vertices at both ends of $\vartheta_3$. The corresponding combinatorial class $\mathcal{M}_7$ is symbolically obtained from $\mathcal{M}_6$ by the substitution $\mathcal{Z}_b \to \mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}) + (\mathbf{SEQ}(\mathcal{Z}))^2 \times \mathcal{P}_0$.

Combine the steps together, the claim follows. The procedure is illustrated in Fig. 1.

In particular, $\mathbf{Q}_\tau^\lambda(x)$ depends only upon the number of arcs and 1-arcs in $\lambda$. Then by definition of the generating function $\mathbf{S}(u, e)$, we obtain $\mathbf{G}^\tau(x)$ by summing over all the possible shapes and the theorem follows. □

### A3 Asymptotics

Let us briefly recall some facts concerning the singularity analysis of functional composition [1]. Suppose $f(x)$ and $g(x)$, with $g(0) = 0$, have non-negative coeffients and are analytic at the origin. We consider the composition $h = f(g(x))$. Let $\rho_f$, $\rho_g$, and $\rho_h$ be the corresponding radii of the convergence, and let $\tau_g = g(\rho_g)$. The asymptotic behavior of $h$ then depends on the comparison of $\tau_g$ and $\rho_f$:

1. $\tau_g > \rho_f$ (supercritical case) the singularity type is that of the external function $f$;
2. $\tau_g < \rho_f$ (subcritical case) the singularity of $f(g)$ is driven by that of the inside function $g$;
3. $\tau_g = \rho_f$ (critical case) the singularity type is a mix of the types of the internal function and the external function and needs special attention.

**Theorem 2.** *Let $\boldsymbol{g}_n^\tau$ denote the number of $\tau$-canonical G-structures on length $n$. Then we have for $\tau = 1, 2$*

$$\boldsymbol{g}_n^\tau \sim k_\tau \, n^{-3/2} \, (\rho_\tau^{-1})^n.$$

*Here, $\rho_1^{-1} \approx 2.2903$, $\rho_2^{-1} \approx 1.8643$, and $k_1$, $k_2$ are positive constants.*

*Proof.* Combining the expressions for $\mathbf{S}(u, e)$ and $G^\tau(x)$ we arrive at

$$\mathbf{G}^\tau(x) = \frac{A_1(x)}{A_2(x)} \cdot \mathbf{M}\left(\frac{B_1(x)}{B_2(x)}\right),$$

where $A_1(x)$, $A_2(x)$, $A_3(x)$, and $A_4(x)$ are fixed polynomials. Clearly $\mathbf{Q}^\tau(x)$ is algebraic. Furthermore, since the composition scheme is supercritical [1] for the cases $\tau = 1$ and $\tau = 2$, the singularity type is that of the external function, i.e., $\mathbf{M}(x)$. In particular, we have $\rho_1^{-1} \approx 2.2903$ and $\rho_2^{-1} \approx 1.8643$. □

For the corresponding structures without any G-quadruplex, we obtain the results immediately by setting $\mathbf{Q}(x) = 0$ in the above derivation. Thus, we obtain $\hat{\rho}_1^{-1} \approx 2.2887$ and $\hat{\rho}_2^{-1} \approx 1.8489$. Numerical values were obtained with `Maple`, version 11.

## B RNA Structure Prediction with and without Quadruplexes

### Human Telomerase RNA component hTERC

It has been previously shown that a G-quadruplex in the $5'-$ UTR of human telomerase RNA hTERC is likely to hinder the formation of the P1 helix of this ncRNA [2]. Since the P1 helix seems essential to serve as a template boundary the quadruplex impairs the telomerase activity. Using the human telomerase RNA *(Acc. No.: AF221907)* we investigated the difference in secondary structure prediction of the newly implemented G-quadruplex aware `RNAfold 2.0g` and `RNAfold` in its regular form. Our prediction results clearly confirm the formation of the G-quadruplex in the $5'-$ UTR that is in conflict with the P1 helix (see Fig. 2). Furthermore, despite the template region beeing predicted within an unpaired loop region by both methods, the probability of the template nucleotides to be unpaired is much lower when taking the G-quadruplex formation into account. This suggests that not only the template boundary function of the P1 helix is disrupted but the template itself might be inaccessible for binding (see Fig. 3 and Fig. 4).

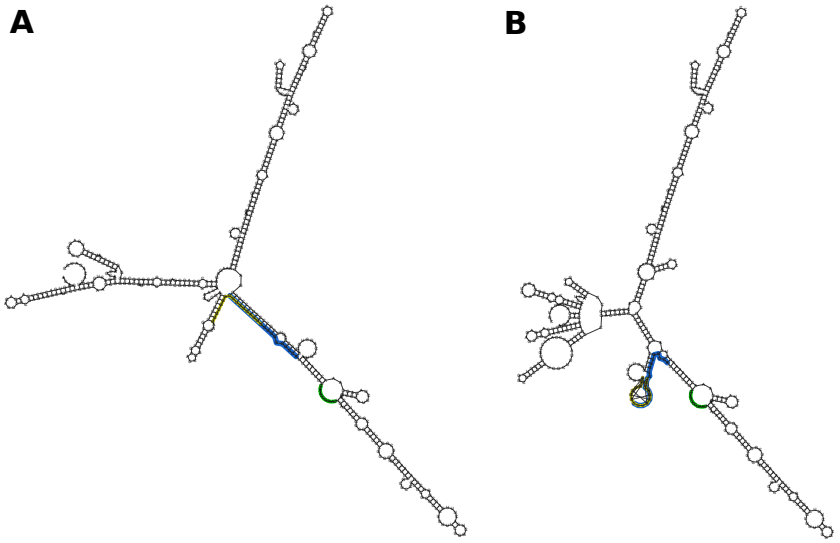



**Fig. 2.** Secondary structure predictions for RNA component of human telomerase (hTERC). The highlighted stretches of the RNA sequence represent the following regions of interest: "template" (green), "$5'-$part of P1 helix" (blue) and "G-quadruplex forming sequence" (yellow). **(A)** Predicted MFE structure without using G-quadruplex capabilities of `RNAfold 2.0`. **(B)** Predicted MFE structure with G-quadruplex aware `RNAfold 2.0g`.

## References

1. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press, New York (2009)
2. Gros, J., Guédin, A., Mergny, J.L., Lacroix, L.: G-Quadruplex formation interferes with P1 helix formation in the RNA component of telomerase hTERC. ChemBioChem **9**(13) (2008) 2075–2079

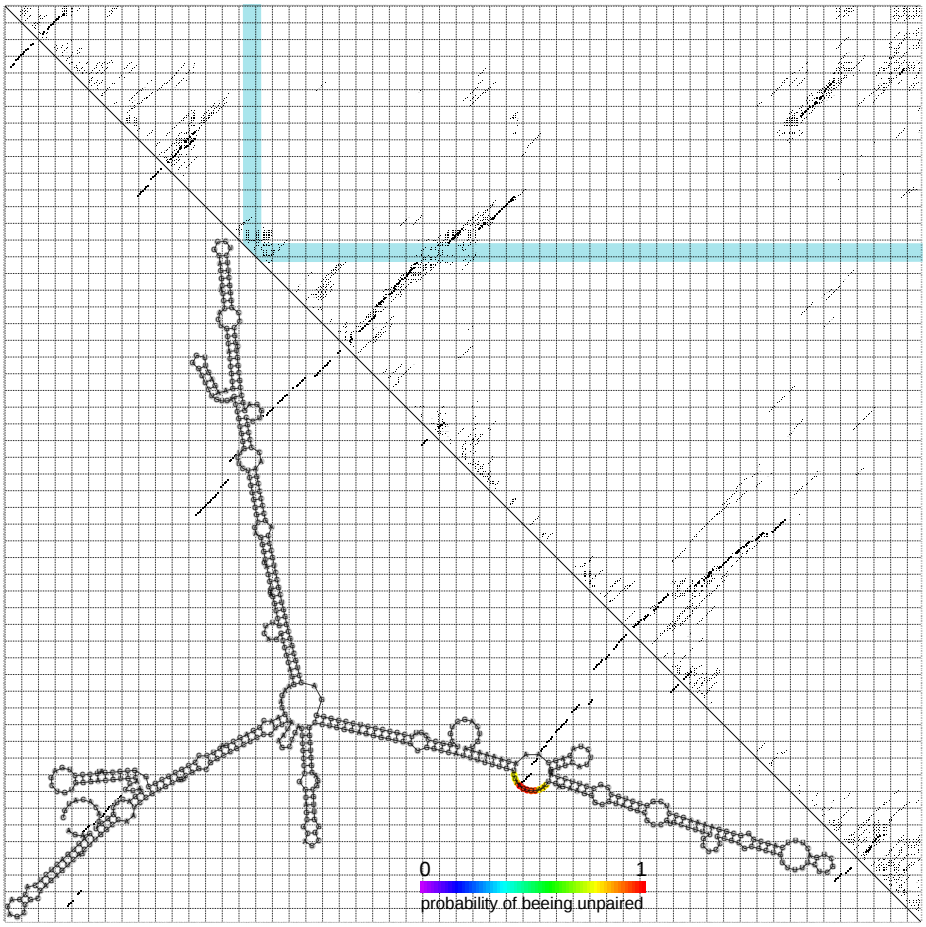



**Fig. 3.** Dot plot of hTERC (AF221907) predicted by `RNAfold`. Upper triangle shows base pair probabilities. Highlighted in light blue are all interaction probabilities that involve nucleotides from the "template" region of the telomerase RNA. The lower part contains the secondary structure plot of the MFE structure as reported by `RNAfold`. Here, the probability for beeing unpaired within the "template" region is colored by a gradient from magenta to red.

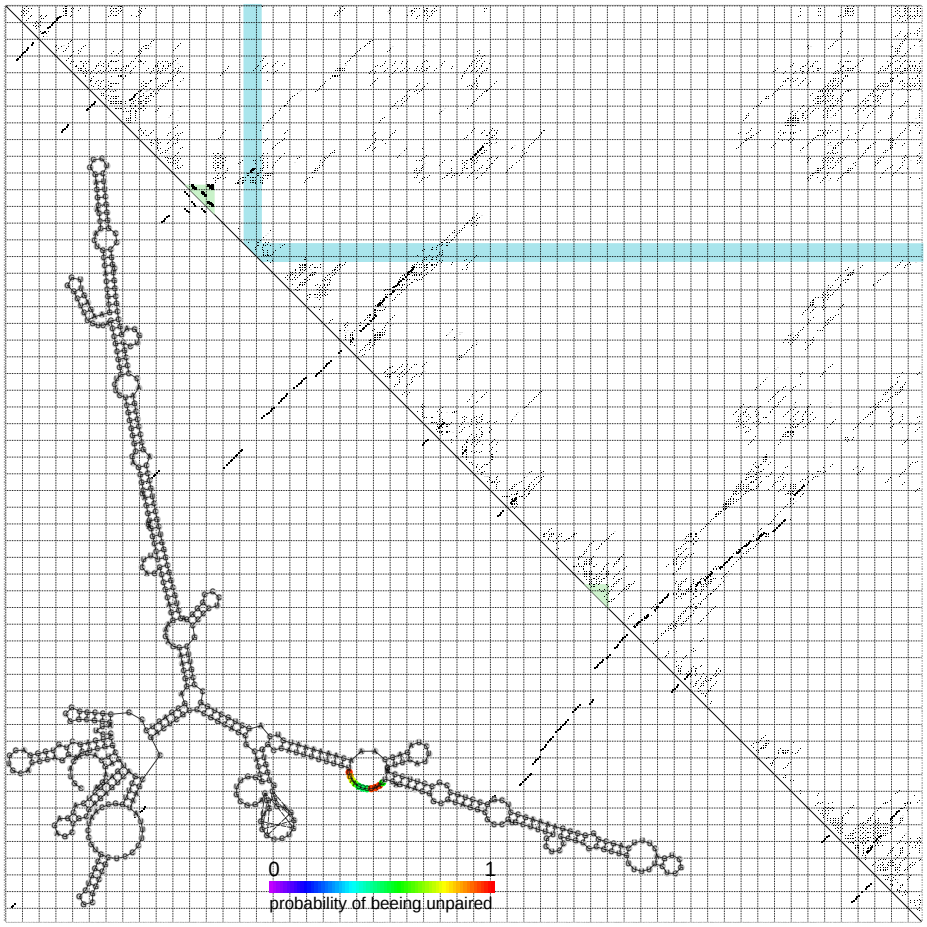



**Fig. 4.** Dot plot of hTERC (AF221907) predicted by the new Q-quadruplex aware `RNAfold 2.0g`. In the upper triangle base pair probabilities are shown. Highlighted in light blue are all interaction probabilities that involve nucleotides from the "template" region of the telomerase RNA. The lower part contains the secondary structure plot of the MFE structure as reported by `RNAfold 2.0g`. Here, the probability for beeing unpaired within the "template" region is colored by a gradient from magenta to red.