# DISTRIBUTION OF GRAPH-DISTANCES IN BOLTZMANN ENSEMBLES OF RNA SECONDARY STRUCTURES

Rolf Backofen[1−2], Markus Fricke[3], Manja Marz[3],
Jing Qin[4], and Peter F. Stadler[4−8]

[1] Department of Computer Science,Chair for Bioinformatics, University of Freiburg, Georges-Koehler-Allee 106, D-79110 Freiburg,
[2] Center for Biological Signaling Studies (BIOSS), Albert-Ludwigs-Universität, Freiburg, Germany
[3] Bioinformatics/High Throughput Analysis Faculty of Mathematics und Computer Science Friedrich-Schiller-University Jena Leutragraben 1, 07743 Jena
[4] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
[5] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, 04107 Leipzig, Germany
[6] Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1,04103 Leipzig, Germany
[7] Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria
[8] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM87501, USA.

## Appendix A: Proof of the $E[d_G(v,w)] = \sum_d d \times \frac{Z^{v,w}[d]}{Z}$

*Proof of* $E[d_G(v,w)] = \sum_d d \times \frac{Z^{v,w}[d]}{Z}$.

$$E[d_G(v,w)] = \sum_G d_G(v,w) \times Pr[G|\xi] \;=\; \sum_d \sum_{G \text{ with } d_G(v,w)=d} d \times \frac{e^{-f(G)/RT}}{Z}$$
$$= \sum_d d \times \frac{\sum_{G \text{ with } d_G(v,w)=d} e^{-f(G)/RT}}{Z} \;= \sum_d d \times \frac{Z^{v,w}[d]}{Z}$$

## Appendix B: Proof of Lemma 1

**Lemma 1.** *The expected distance* $E[d^G_{i,j}]$ *can be calculated as:*

$$E[d^G_{i,j}] = (a + E[d^G_{i+1,j}]) \cdot \frac{1 \cdot Q_{i+1,j}}{Q_{i,j}} + \sum_{i<k\leq j} (b + E[d^G_{k+1,j}]) \cdot \frac{Q^b_{i,k} \cdot Q_{k+1,j}}{Q_{i,j}} \quad (1)$$

*Proof of Lemma 1.*

Let $G$ be a structure. For simplicity of notation, we write $G = \bullet G'$ if the first position is unpaired, and $G = (\ldots)_j G'$ is the first base is paired to some position $j$, and $G'$ is the substructure of $G$ starting from position $j+1$. Alternatively, we may use the notation $(i,j) \in G$ for the case where the position $i$ and $j$ are base paired in $G$.

The expected length $E[d_G(i,j)]$ can be calculated as: follows:

$$\begin{aligned} E[d_G(i,j)] &= \sum_{G \text{ struct. of } \xi[i\ldots j]} d_G(i,j) Pr[G|\xi[i\ldots j]] \\ &= \sum_{G=\bullet G'} (a + d_{G'}(i,j)) Pr[G|\xi[i\ldots j]] \quad + \sum_{i<k\leq j} \sum_{G=(\ldots)_k G'} (b + d_{G'}(i,j)) Pr[G|\xi[i\ldots j]] \\ &\overset{def.}{=} \qquad\qquad EL_{sg} \qquad\qquad\qquad + \sum_{i<k<j} \qquad\qquad EL_{bp(k)} \end{aligned}$$

Now $EL_{sg}$ can be simplified as follows:

$$\begin{aligned} EL_{sg} &= \sum_{G=\bullet G'} (a + d_{G'}(i,j)) Pr[G|\xi[i\ldots j]] \\ &= \left( \sum_{G=\bullet G'} a \cdot Pr[G|\xi[i\ldots j]] \right) + \left( \sum_{G=\bullet G'} d_{G'}(i,j) \cdot Pr[G|\xi[i\ldots j]] \right) \\ &= a \cdot Pr[G = \bullet G'|\xi[i\ldots j]] + \left( \sum_{G=\bullet G'} d_{G'}(i,j) \cdot Pr[G|\xi[i\ldots j]] \right), \end{aligned}$$

where $Pr[G = \bullet G'|\xi[i\ldots j]]$ can be calculated as the probability of the first position to be single-stranded in the sequence $\xi[i\ldots j]$, i.e.,

$$Pr[G = \bullet G'|\xi[i\ldots j]] = \frac{1 \cdot Q_{i+1,j}}{Q_{i,j}}$$

We are also able to push the second term since

$$\sum_{G=\bullet G'} d_{G'}(i,j) \cdot Pr[G|\xi[i\ldots j]] = \sum_{G'} d_{G'}(i,j) \cdot Pr[\bullet G'|\xi[i\ldots j]]$$

Now we know that for every $G'$ we have that the Boltzmann weighted energy of $G'$ is part of the partition function of $Q_{i+1,j}$. Thus we get

$$\begin{aligned}
&= \sum_{G'} d_{G'}(i,j) \cdot \frac{\exp(-E(\bullet G')/kT)}{Q_{i,j}} \\
&= \sum_{G'} d_{G'}(i,j) \cdot \frac{\exp(-E(G')/kT)}{Q_{i+1,j}} \frac{Q_{i+1,j}}{Q_{i,j}} \\
&= \frac{Q_{i+1,j}}{Q_{i,j}} \sum_{G'} d_{G'}(i,j) \cdot \frac{\exp(-E(G')/kT)}{Q_{i+1,j}} \\
&= Pr[G = \bullet G' | \xi[i \dots j]] \sum_{G'} d_{G'}(i,j) \cdot Pr[G' | \xi[i+1 \dots j]] \\
&= Pr[G = \bullet G' | \xi[i \dots j]] \cdot E[d_G(i+1,j)]
\end{aligned}$$

Overall we get

$$EL_{sg} = (a + E[d_G(i+1,j)]) \cdot Pr[G = \bullet G' | \xi[i \dots j]]$$

For the term $EL_{bp(k)}$, we have a similar reduction:

$$\begin{aligned}
EL_{bp(k)} &= \sum_{G=(\dots)_k G'} (b + d_{G'}(i,j)) Pr[G|\xi[i \dots j]] \\
&= \left( \sum_{G=(\dots)_k G'} b \cdot Pr[G|\xi[i \dots j]] \right) + \left( \sum_{G=(\dots)_k G'} d_{G'}(i,j) Pr[G|\xi[i \dots j]] \right) \\
&= (b \cdot Pr[G = (\dots)_k G' | \xi[i \dots j]]) + \left( \sum_{G=(\dots)_k G'} d_{G'}(i,j) Pr[G|\xi[i \dots j]] \right),
\end{aligned}$$

where $Pr[G = (\dots)_k G' | \xi[i \dots j]] = \frac{Q^b_{ik} \cdot Q_{k+1,j}}{Q_{i,j}}$.

Now

$$\begin{aligned}
\sum_{G=(\dots)_k G'} d_{G'}(i,j) Pr[G|\xi[i \dots j]] &= \sum_{G'} \sum_{G''=(G''')_k} d_{G'}(i,j) Pr[G''G'|\xi[i \dots j]] \\
&= \sum_{G'} \sum_{G''=(G''')_k} d_{G'}(i,j) \frac{\exp(E(G'')/kT) \exp(G'/kT)}{Q_{ij}} \\
&= \sum_{G'} \sum_{G''=(G''')_k} d_{G'}(i,j) \frac{\exp(E(G'')/kT) \exp(G'/kT)}{Q_{ij}} \\
&= \sum_{G'} d_{G'}(i,j) \frac{\left( \sum_{G''=(G''')_k} \exp(E(G'')/kT) \right) \exp(G'/kT)}{Q_{ij}} \\
&= \sum_{G'} d_{G'}(i,j) \frac{Q^b_{i,k} \exp(G'/kT)}{Q_{ij}}
\end{aligned}$$

Now we can again simply extend by $Q_{k+1,j}$, getting

$$
\begin{aligned}
&= \sum_{G'} d_{G'}(i,j) \frac{Q^b_{i,k} \cdot Q_{k+1,j} \cdot \exp(G'/kT)}{Q_{ij} \cdot Q_{k+1,j}} \\
&= \sum_{G'} d_{G'}(i,j) \frac{Q^b_{i,k} \cdot Q_{k+1,j}}{Q_{ij}} \cdot \frac{\exp(G'/kT)}{Q_{k+1,j}} \\
&= Pr[G = (\ldots)_k G' | \xi[i \ldots j]] \sum_{G'} d_{G'}(i,j) Pr[G' | \xi[k+1 \ldots j]] \\
&= Pr[G = (\ldots)_k G' | \xi[i \ldots j]] \cdot E[d_G(k+1,j)]
\end{aligned}
$$

Overall we get

$$EL_{bp(k)} = (b + E[d_G(k+1,j)]) \cdot Pr[G = (\ldots)_k G' | \xi[i \ldots j]]$$

and thus the second summand.

## Appendix C: Recurrsions of different cases for $Z^{B,v}_{i,j}$.

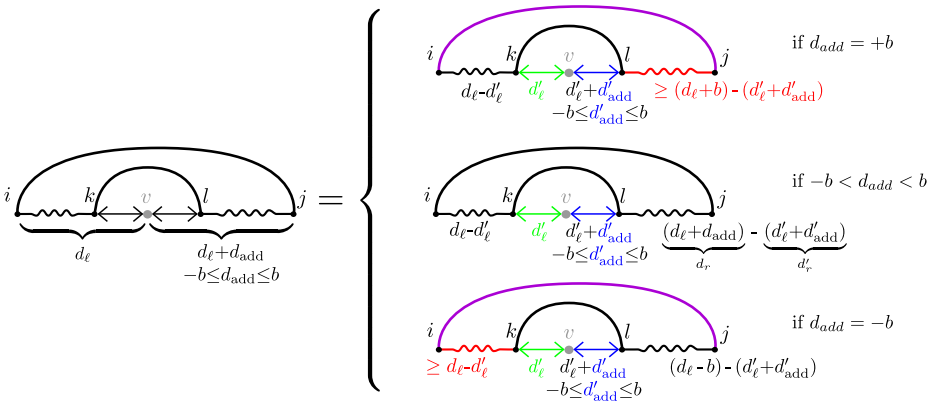



**Fig. 1.** Different cases for $Z^{B,v}_{i,j}[d_\ell, d_\ell + d_{\text{add}}]$. The values that are choose to split $d_\ell$ and $d_{\text{add}}$ are indicated in green and blue. When the arc $\{i,j\}$ is colored violet, then there is a shortest path that does not use the distance marked in red but uses the other direction together with the arc $(i,j)$.

Now there are three sub-cases (see Figure 1). If $-b < d_{add} < +b$, then we know that neither a shortest path $v \overset{p}{\rightsquigarrow} i$ nor $v \overset{p}{\rightsquigarrow} j$ uses the arc $\{i,j\}$. The left distance is thus given by $d_\ell - d'_\ell$. Using the shortcuts $d_r = d_\ell + d_{\text{add}}$ and $d'_r = d'_\ell + d'_{\text{add}}$, then the distance between $l$ and $j$ must be $d_r - d'_r = (d_\ell + d_{\text{add}}) - (d'_\ell + d'_{\text{add}})$. If, on the other hand, $d_{\text{add}} = +b$, then we know that there

is at least one shortest path that can be composed by using a shortest path $v \rightsquigarrow i$, followed by the arc $\{i, j\}$. This of course implies that the shortest path $v \overset{p}{\rightsquigarrow} j$ is has exactly the length $d_\ell + b$, or is larger. For a sub-path $l+1 \overset{p'}{\rightsquigarrow} j$ this implies that the length is greater or equal $d = d_r - d'_r = (d_\ell + b) - (d'_\ell + d'_{\text{add}})$. Thus, we just have to add all partition functions $Z^{I'}_{k,j}[d']$ with $d' > d$. This can be done efficiently by using a precalculated matrix $Z^{I'\geq}_{i,j}[d]$, which is defined as $\sum_{d' \geq d} Z^{I'}_{i,j}[d']$. Note that $Z^{I'\geq}_{i,j}[d]$ can also be defined if we restrict in all recursion the distance $d$ to a threshold $\theta_d$, since $Z^{I'\geq}_{i,j}[d] = \sum_{d' \geq d} Z^{I'}_{i,j}[d'] = Q'_{i,j} - \sum_{d'<d} Z^{I'}_{i,j}[d']$. In which, where $Q'_{i,j}$ is $Q_{i+1,j-1}$ if $j > i+1$, 1 if $j = i+1$ and 0 otherwise. Note, furthermore, that all $Z^{I'}_{i,j}[d']$ for $d' < d \leq \theta_d$ are calculated when we restrict the distance to $\theta_d$

Finally, if $d_{\text{add}} = -b$, then the shortest path $l \overset{p}{\rightsquigarrow} j$ has distance $(d_\ell - b) - (d'_\ell + d'_{\text{add}})$. For the shortest path $k \overset{p}{\rightsquigarrow} i$, we know that it has length $d_\ell - d'_\ell$ or greater, which can be resolved by again using $Z^{I'\geq}_{i,k-1}[d_\ell - d'_\ell]$. Overall, we get the following optimized recursion for $Z^{B,v}_{i,j}[d_\ell, d_\ell + d_{\text{add}}]$ with $d_\ell \neq 0$ and $d_\ell + d_{\text{add}} \neq 0$:

$$
\begin{aligned}
&Z^{B,v}_{i,j}[d_\ell, d_\ell + d_{\text{add}}] = \\
&\widehat{Q}^b_{i,j} \cdot \begin{cases}
\displaystyle\sum_{\substack{k \neq l \\ i<k\leq v \\ v \leq l < j}} \sum_{d'_\ell \leq d_\ell} \sum_{\substack{d'_{\text{add}} \\ -b \leq d'_{\text{add}} \leq b}} \begin{pmatrix} Z^{I'}_{i,k}[d_\ell - d'_\ell] \cdot Z^{B,v}_{k,l}[d'_\ell, d'_\ell + d'_{\text{add}}] \\ \cdot Z^{I'}_{l,j}[(d_\ell + d_{\text{add}}) - (d'_\ell + d'_{\text{add}})] \end{pmatrix} & \text{if } -b < d_{\text{add}} < b \\
\displaystyle\sum_{\substack{k \neq l \\ i<k\leq v \\ v \leq l < j}} \sum_{d'_\ell \leq d_\ell} \sum_{\substack{d'_{\text{add}} \\ -b \leq d'_{\text{add}} \leq b}} \begin{pmatrix} Z^{I'}_{i,k}[d_\ell - d'_\ell] \cdot Z^{B,v}_{k,l}[d'_\ell, d'_\ell + d'_{\text{add}}] \\ \cdot Z^{I'\geq}_{l,j}[(d_\ell + b) - (d'_\ell + d'_{\text{add}})] \end{pmatrix} & \text{if } d_{\text{add}} = b \\
\displaystyle\sum_{\substack{k \neq l \\ i<k\leq v \\ v \leq l < j}} \sum_{d'_\ell \leq d_\ell} \sum_{\substack{d'_{\text{add}} \\ -b \leq d'_{\text{add}} \leq b}} \begin{pmatrix} Z^{I'\geq}_{i,k}[d_\ell - d'_\ell] \cdot Z^{B,v}_{k,l}[d'_\ell, d'_\ell + d'_{\text{add}}] \\ \cdot Z^{I'}_{l,j}[(d_\ell - b) - (d'_\ell + d'_{\text{add}})] \end{pmatrix} & \text{if } d_{\text{add}} = -b
\end{cases}
\end{aligned}
$$

## Appendix D: Detailed Data Related to Fig. 3

**Table 1.** Graph distance of the genomic RNA of *human coronavirus 229E* computed from a concatenation of position 1-576 and 25188-25688. The graph distance is computed from the most 5' end to the most 3' end of the sequence (yellow). The RNA secondary bring the leader transcription regulation site (L-TRS) in close spatial proximity with the body transcription regulation site (B-TRS). The structures (a), (b) and (c) are the most stable structures considering the sub-ensembles which are the sets of structures of graph distance 14, 5 and 35, respectively. In which, the graph distances 7, 6 and 14 are the 1st, 6th and 8th most favorate graph distances considering Boltzmann facor.

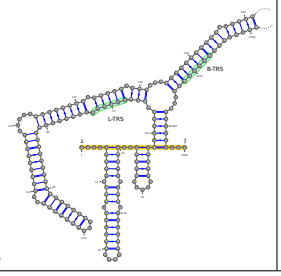
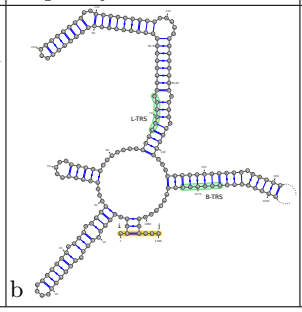
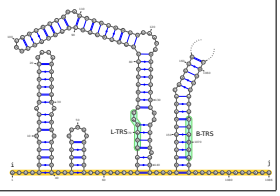
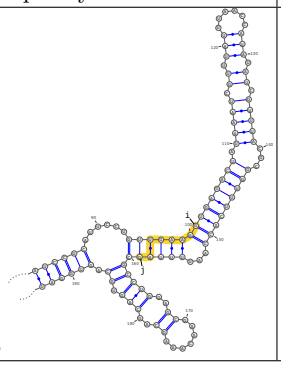
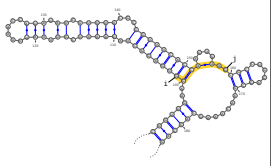
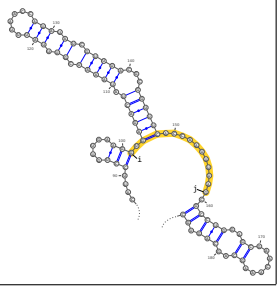
| 1st | 6th | 8th |
|---|---|---|
| distance=14 | distance=5 | distance=35 |
| frequency=5132 | frequency=204 | frequency=92 |
|  a |  b |  c |

**Table 2.** Graph distance of intron CG16979-RA_intron_0_0_chr3L_15569803 from *drosophila melanogaster* (dm3). The intron is extended at the 5' and 3' end with 100 bases. The graph distance is computed between $i = 101(G)$ and $j = 159(G)$ (yellow). The structures (a), (b) and (c) are the most stable structures considering the sub-ensembles which are the sets of structures of graph distance 7, 6 and 14, respectively. In which, the graph distances 7, 6 and 14 are the 1st, 6th and 10th most favorate graph distances considering Boltzmann facor.

| 1st | 6th | 10th |
|---|---|---|
| distance=7 | distance=6 | distance=14 |
| frequency=5593 | frequency=13 | frequency=1 |
|  a |  b |  c |