Supplementary Material: A Partition Function Algorithm for Nucleic Acid Secondary Structure Including Pseudoknots

ROBERT M. DIRKS[1], **NILES A. PIERCE**[2]

[1] *Department of Chemistry, California Institute of Technology, Pasadena, CA 91125*

[2] *Department of Applied & Computational Mathematics and Department of Bioengineering, California Institute of Technology, Pasadena, CA 91125*

Received 16 December 2002; Accepted 5 March 2003; J Comput Chem 24: 1664–1677, 2003

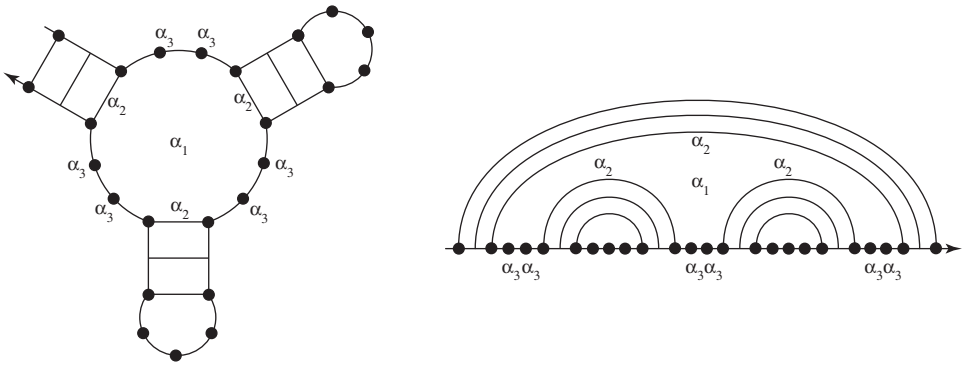

Figure S1. Multiloop energy expression. For this example there are three base pairs defining the multiloop ($B = 3$) and six unpaired bases in the multiloop ($U = 6$) so $G^{multi} = \alpha_1 + 3\alpha_2 + 6\alpha_3$.

Correspondence to: N.A. Pierce; e-mail: niles@caltech.edu

function fastiloops($i, j, l, Q^b, Q^x, Q^{x2}$)
//Q^x **recursion** for $O(N^3)$ internal loop contributions to Q^b
if ($l \geq 15$) //smallest subsequence not added to Q^b as special case
  $L_1 = 4$ //explicitly add in terms for $L_1 = 4$, $L_2 \geq 4$
  $d = i + L_1 + 1$
  for $L_2 = 4, j-d-5$
    $s = L_1 + L_2$
    $e = j - L_2 - 1$
    $G^{\text{partial}} = \gamma_1(s) + \gamma_2(|L_1 - L_2|) + \gamma_3(e, d, e+1, d-1)$
    $Q^x_{i,s}$ += $\exp\{-G^{\text{partial}}/RT\}\, Q^b_{d,e}$
  $L_2 = 4$ //explicitly add in terms for $L_1 \geq 5$, $L_2 = 4$
  $e = j - L_2 - 1$
  for $L_1 = 5, e-i-5$
    $s = L_1 + L_2$
    $d = i + L_1 + 1$
    $G^{\text{partial}} = \gamma_1(s) + \gamma_2(|L_1 - L_2|) + \gamma_3(e, d, e+1, d-1)$
    $Q^x_{i,s}$ += $\exp\{-G^{\text{partial}}/RT\}\, Q^b_{d,e}$
//Next convert Q^x into interior loop energies
for $s = 8, l-7$
  if (sequence permits $i \cdot j$ base pair)
    $Q^b_{i,j}$ += $Q^x_{i,s}\, \exp\{-\gamma_3(i, j, i+1, j-1)/RT\}$
//Extend loops from s to $s+2$ for future calculation
//of $Q^b_{i-1,j+1}$ with subsequence length $l+2$
if ($i \neq 1$ & $j \neq N$)
  for $s = 8, l-7$
    $Q^{x2}_{i-1,s+2} = Q^x_{i,s}\, \exp\{-[\gamma_1(s+2) - \gamma_1(s)]/RT\}$
//Add small inextensible interior loop terms to Q^b as special cases
for $L_1 = 0, 3$
  $d = i + L_1 + 1$
  for $L_2 = 0, \min(3, j-d-5)$
    $e = j - L_2 - 1$
    $Q^b_{i,j}$ += $\exp\{-G^{\text{internal}}_{i,d,e,j}/RT\}\, Q^b_{d,e}$
//Add bulge loops and large asymmetric loops as special cases
for $L_1 = 0, 3$ //Cases $L_1 = 0, 1, 2, 3$, $L_2 \geq 4$
  $d = i + L_1 + 1$
  for $L_2 = 4, j-d-5$
    $e = j - L_2 - 1$
    $Q^b_{i,j}$ += $\exp(-G^{\text{internal}}_{i,d,e,j}/RT)\, Q^b_{d,e}$
for $L_2 = 0, 3$//Cases $L_1 \geq 4$, $L_2 = 0, 1, 2, 3$
  $e = j - L_2 - 1$
  for $L_1 = 4, e-i-5$
    $d = i + L_1 + 1$
    $Q^b_{i,j}$ += $\exp\{-G^{\text{internal}}_{i,d,e,j}/RT\}\, Q^b_{d,e}$

Figure S2. Pseudocode for computing interior loop contributions to Q^b in $O(N^3)$ as an alternative to the $O(N^4)$ interior loop recursion of Figure 8. Here, N is the length of the strand and $l = j-i+1$ is the length of the substrand under consideration at any given point during the recursive process. A schematic representation of "fastiloops" is provided in Supplementary Material Figure S3. The smallest "possible extensible loop" is the case $L_1 = L_2 = 4$ with size $s = 8$. Therefore, the smallest subsequence for which Q^x can be employed is $l = 15$ (adding the four bases for i, d, e, j and a minimum hairpin of three bases between $d \cdot e$). For a given i and j, $Q^x_{i,s}$ already contains the contributions to $Q^b_{i,j}$ for all extensible loops of size s except for the two cases when either $L_1 = 4$ or $L_2 = 4$ (which cannot be obtained by extending smaller loops that use a different energy expression). Enriching $Q^x_{i,s}$ with these two new possible extensible loops, we then convert $Q^x_{i,s}$ into contributions to $Q^b_{i,j}$ by introducing the term for closing these loops with pair $i \cdot j$. $Q^x_{i,s}$ is then extended to provide future values of $Q^x_{i-1,s+2}$. All other interior loop contributions (cases with either $L_1 \leq 3$ or $L_2 \leq 3$) are then added directly to $Q^b_{i,j}$ using the special energy expressions of the standard model implied by $G^{\text{internal}}_{i,d,e,j}$. Note that the subsequence length l is fixed inside each call to the function "fastiloops". Hence, specifying i implies $j = i+l-1$. For subsequences of length l, we use $Q^x_{i,s}$ (j implied) to compute $Q^x_{i-1,s+2}$ ($j+1$ implied) which will later be used to compute contributions to $Q^b_{i-1,j+1}$ for subsequences of length $l+2$. Thus, for a given value of l, the values of $Q^x_{i,s}$ need only be stored for all legal values of i and s until l has been incremented 3 times, at which point it can be discarded. This is accomplished by using $Q^{x1}_{i,s}$ and $Q^{x2}_{i,s}$ to store future contributions for subsequences of length $l+1$ and $l+2$.

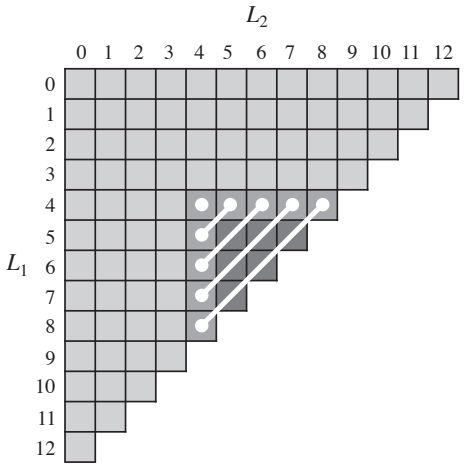

Figure S3. Schematic for interpreting the pseudocode of Supplementary Material Figure S2, which computes the interior loop contributions to the partition function in $O(N^3)$. For a given i (with j implied by the current subsequence length l), the grid illustrates the method of computing contributions to $Q^b_{i,j}$ for all interior loops with sides of length L_1 and L_2. For the depicted case, the maximum interior loop size is $s = L_1 + L_2 = 12$. Adding seven bases to account for the closing bases i, d, e, j and the smallest allowed hairpin of three bases between $d \cdot e$, the subsequence under consideration is therefore of length $l = 19$. The $O(N)$ pale gray cases with either $L_1 \leq 3$ or $L_2 \leq 3$ use special energy functions and are added explicitly to $Q^b_{i,j}$. The medium gray and dark gray cases are the possible extensible loop contributions that are computed using $Q^x_{i,s}$. Each diagonal line spans the terms that will be stored in $Q^x_{i,s}$ for a particular value of s. The medium gray terms are the new possible extensible loops with either $L_1 = 4$ or $L_2 = 4$. The dark gray terms were previously incorporated into to $Q^x_{i,s}$ by extending smaller possible extensible loops.

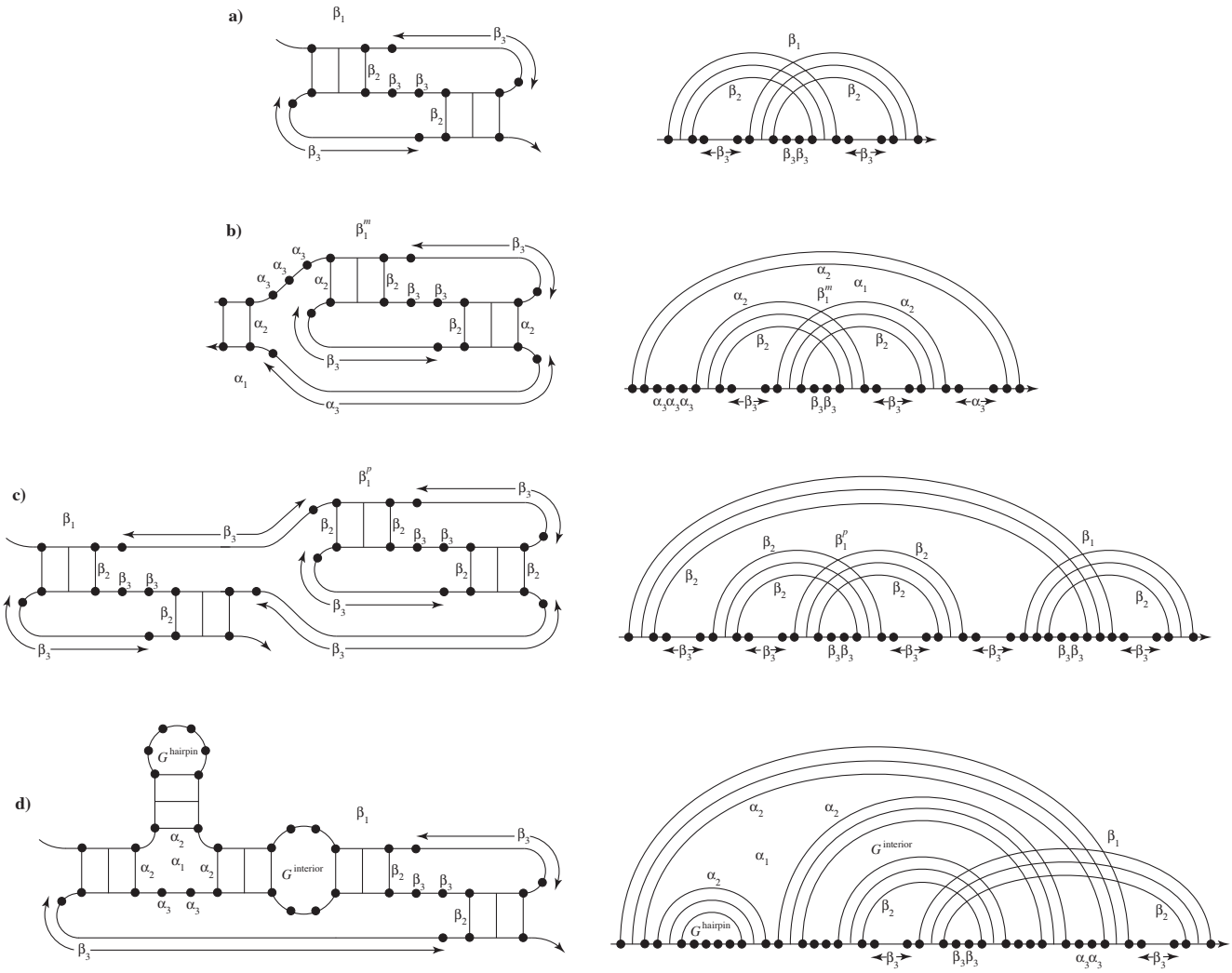

Figure S4. Illustration of the pseudoknot energy expression. a) An external pseudoknot. Bases external to the pseudoknot have no associated energy. The penalty for forming an external pseudoknot is β_1 . Base pairs that border the pseudoknot interior receive penalty β_2 while unpaired bases on the pseudoknot interior receive penalty β_3. The energies associated with the stacked base pairs are described using the standard model. b) A pseudoknot inside a multiloop. The penalty for forming a pseudoknot inside a multiloop is β_1^m. The treatment of β_2 and β_3 inside the pseudoknot remains the same. In addition, the standard penalty for formation of a multiloop is α_1, the two pseudoknot base pairs that border the multiloop are given the standard multiloop penalty α_2, and unpaired bases that are inside the multiloop are given penalty α_3. c) A pseudoknot within a pseudoknot. The energy treatment for the exterior pseudoknot remains the same. The penalty for forming a pseudoknot inside a pseudoknot is β_1^p. Base pairs from the interior pseudoknot that border the exterior pseudoknot receive penalty β_2. Otherwise, the energetic treatment of the interior pseudoknot is the same as for an exterior pseudoknot. The partition function recursions allow an arbitrary number of levels of pseudoknots within pseudoknots. d) Pseudoknot with a hairpin and an interior loop inside a spanning region of the pseudoknot. The multiloop that forms at the base of the hairpin is treated using the standard multiloop potential. In general, the spanning region of a pseudoknot may contain interior loops, hairpins, multiloops or additional pseudoknots.

```
Initialize (Q, Q^b, Q^m, Q^p, Q^z) //O(N^2) space
Initialize (Q^g) //O(N^4) space
Set all values to 0 except Q_{i,i-1} = Q^z_{i,i-1} = 1
for l = 1, N
  for i = 1, N-l+1
    j = i+l-1
    //Q^b recursion
    Q^b_{i,j} = exp(-G^hairpin_{i,j}/RT)
    for d = i+1, j-5 //all possible rightmost pairs d·e
      for e = d+4, j-1
        Q^b_{i,j} += exp(-G^interior_{i,d,e,j}/RT) Q^b_{d,e}
        Q^b_{i,j} += Q^m_{i+1,d-1} Q^b_{d,e} exp{-[α1 + 2α2 + α3 (j-e-1)]/RT}
    for d = i+1, j-9 //all possible rightmost pseudoknots filling [d, e]
      for e = d+8, j-1
        G^recursion = α1 + β1^m + 3α2 + α3 (j-e-1)
        Q^b_{i,j} += exp{-[G^recursion + α3 (d-i-1)]/RT} Q^p_{d,e}
        Q^b_{i,j} += Q^m_{i+1,d-1} Q^p_{d,e} exp{-G^recursion/RT}
    //Q^g recursion
    for d = i + 1, j-5 //set inner pair d·e
      for e = d+4, j - 1
        Q^g_{i,d,e,j} += exp(-G^interior_{i,d,e,j}/RT)
    for d = i + 2, j-6 //set inner pair d·e
      for e = d+4, j - 2
        for c = i + 1, d-1 //recursion on middle pair c·f
          for f = e+1, j - 1
            Q^g_{i,d,e,j} += exp(-G^interior_{i,c,f,j}/RT) Q^g_{c,d,e,f}
    for d = i + 6, j-5 //set inner pair d·e
      for e = d+4, j - 1
        Q^g_{i,d,e,j} += Q^m_{i+1,d-1} exp{-[α1+2α2+α3(j-e-1)]/RT}
    for d = i + 1, j-10 //set inner pair d·e
      for e = d+4, j - 6
        Q^g_{i,d,e,j} += exp{-[α1+2α2+α3(d-i-1)]/RT} Q^m_{e+1,j-1}
    for d = i + 6, j-10 //set inner pair d·e
      for e = d+4, j - 6
        Q^g_{i,d,e,j} += Q^m_{i+1,d-1} exp{-[α1+2α2]/RT} Q^m_{e+1,j-1}
    for d = i + 7, j-6 //set inner pair d·e
      for e = d+4, j - 2
        for c = i + 6, d-1 //recursion on middle pair c·f
          for f = e+1, j - 1
            G^recursion = α1+2α2+α3(j-f-1)
            Q^g_{i,d,e,j} += Q^m_{i+1,c-1} Q^g_{c,d,e,f} exp{-G^recursion/RT}
    for d = i + 2, j-11 //set inner pair d·e
      for e = d+4, j - 7
        for c = i + 1, d-1 //recursion on middle pair c·f
          for f = e+1, j - 6
            G^recursion = α1+2α2+α3(c-i-1)
            Q^g_{i,d,e,j} += exp{-G^recursion/RT} Q^g_{c,d,e,f} Q^m_{f+1,j-1}
    for d = i + 7, j-11 //set inner pair d·e
      for e = d+4, j - 7
        for c = i + 6, d-1 //recursion on middle pair c·f
          for f = e+1, j - 6
            Q^g_{i,d,e,j} += Q^m_{i+1,c-1} exp{-[α1+2α2]/RT} Q^g_{c,d,e,f} Q^m_{f+1,j-1}
    //Q^p recursion
    for a = i + 1, j-7 //place points from left to right
      for b = a+1, j-6
        for c = b + 1, j-5
          for d = max(c+1, a+4), j - 3
            for e = d+1, j - 2
              for f = max(e+1, c+4), j - 1
                Q^p_{i,j} += Q^z_{a+1,b-1} Q^z_{c+1,d-1} Q^z_{e+1,f-1}
                          · Q^g_{i,a,d,e} Q^g_{b,c,f,j} exp{-2β2/RT}
    //Q, Q^m, Q^z recursions
    Q_{i,j} = 1 //empty recursion
    Q^z_{i,j} = exp(-[β3(j-i+1)]/RT)
    for d = i, j-4 //all possible rightmost pairs d·e
      for e = d+4, j
        Q_{i,j} += Q_{i,d-1} Q^b_{d,e}
        Q^m_{i,j} += exp{-[α2 + α3 (d-i) + α3 (j-e)]/RT} Q^b_{d,e}
        Q^m_{i,j} += Q^m_{i,d-1} Q^b_{d,e} exp{-[α2 + α3 (j-e)]/RT}
        Q^z_{i,j} += Q^z_{i,d-1} Q^b_{d,e} exp{-[β2 + β3(j-e)]/RT}
    for d = i, j-8 //all possible rightmost pseudoknots filling [d, e]
      for e = d+8, j
        Q_{i,j} += Q_{i,d-1} Q^p_{d,e} exp{-β1/RT}
        Q^m_{i,j} += exp{-[β1^m + 2α2 + α3 (d-i) + α3 (j-e)]/RT} Q^p_{d,e}
        Q^m_{i,j} += Q^m_{i,d-1} Q^p_{d,e} exp{-[β1^m + 2α2 + α3 (j-e)]/RT}
        Q^z_{i,j} += Q^z_{i,d-1} Q^p_{d,e} exp{-[β1^p + 2β2 + β3(j-e)]/RT}
//Partition function is Q_{1,N}
```

Figure S5. Pseudocode implementation of an $O(N^8)$ dynamic programming partition function algorithm for nucleic acids with pseudoknots. Here, N is the length of the strand and $l = j-i+1$ is the length of the substrand under consideration at any given point during the recursive process. The recursions are described schematically in Figures 12-16. In this pseudocode, care has been taken to define programming loop bounds so as to consider only valid secondary structures. The standard model requires three or more unpaired bases in a hairpin, so a b-curve must satisfy $j-i \geq 4$. Looking at Figure 15, imposing this requirement on all base pairs implies that a p-curve must satisfy $j-i \geq 8$. In the interior of a pseudoknot, the steric constraints on hairpins sometimes lead to two conflicting requirements that are incorporated using a "max" function to define the bounds for d and f. The disallowed structures have infinite energies according to the physical model so it is computationally efficient to exclude them from consideration.

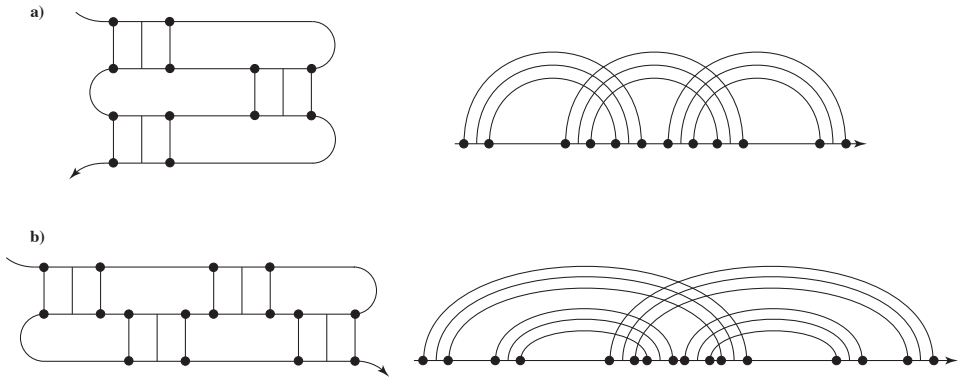

Figure S6. Examples of pseudoknots that are excluded from the partition function recursions. Neither structure can be decomposed into two spanning regions as required by Figure 15. The structure prediction recursions of Rivas and Eddy[25] include both structures while the the structure prediction recursions of Akutsu[24] include the latter.

```
function fastiloops(i, j, l, Q^g, Q^x, Q^x2)
//Q^x recursion for O(N^5) internal loop contributions to Q^g
if (l ≥ 17) //smallest subsequence not added to Q^g as special case
  for d = i+6, j−10
    for e = d+4, j−6
      L1 = 4 //explicitly add in terms for L1 = 4, L2 ≥ 4
      c = i + L1 + 1
      for L2 = 4, j−e−2
        s = L1 + L2
        f = j − L2 − 1
        G^partial = γ1(s) + γ2(|L1 − L2|) + γ3(f, c, f+1, c−1)
        Q^x_{i,d,e,s} += exp{−G^partial/RT} Q^g_{c,d,e,f}
      if ( d ≥ i + 7)
        L2 = 4 //explicitly add in terms for L1 ≥ 5, L2 = 4
        f = j − L2 − 1
        for L1 = 5, d−i−2
          s = L1 + L2
          c = i + L1 + 1
          G^partial = γ1(s) + γ2(|L1 − L2|) + γ3(f, c, f+1, c−1)
          Q^x_{i,d,e,L1+L2} += exp{−G^partial/RT} Q^g_{c,d,e,f}
for d = i+1, j−5
  for e = d+4, j−1
    //Convert Q^x into interior loop energies
    if (l ≥ 17 & sequence permits i·j base pair)
      for size = 8, l−9
        Q^g_{i,d,e,j} += Q^x_{i,d,e,s} exp{−γ3(i, j, i + 1, j − 1)/RT}
    //Extend loops for future use
    if (i ≠ 1 & j ≠ N)
      for s = 8, l−9
        Q^x2_{i−1,d,e,s+2} = Q^x_{i,d,e,s} exp{−[γ1(s+2) − γ1(s)]/RT}
    //Add small inextensible interior loops to Q^g as special cases
    for L1 = 0, min(3, d−i−2)
      c = i + L1 + 1
      for L2 = 0, min(3, j−e−2)
        f = j − L2 − 1
        Q^g_{i,d,e,j} += exp{−G^interior_{i,c,f,j}/RT} Q^g_{c,d,e,f}
    //Add bulge loops and large asymmetric loops as special cases
    for L1 = 0, min(3, d−i −2) //Cases L1 = 0, 1, 2, 3, L2 ≥ 4
      c = i + L1 + 1
      for L2 = 4, j−e−2
        f = j − L2 − 1
        Q^g_{i,d,e,j} += exp{−G^interior_{i,c,f,j}/RT} Q^g_{c,d,e,f}
    for L2 = 0, min(3, j−e−2) //Cases L1 ≥ 4, L2 = 0, 1, 2, 3
      f = j − L2 − 1
      for L1 = 4, d−i−2
        c = i + L1 + 1
        Q^g_{i,d,e,j} += exp{−G^interior_{i,c,f,j}/RT} Q^g_{c,d,e,f}
```

Figure S7. Pseudocode for computing interior loop contributions to Q^g in $O(N^5)$ as an alternative to the $O(N^6)$ interior loop recursion of Figure 19. Here, N is the length of the strand and $l = j-i+1$ is the length of the substrand under consideration at any given point during the recursive process. The smallest "possible extensible loop" is the case $L_1 = L_2 = 4$ with size $s = 8$. Therefore, the smallest subsequence for which Q^x can be employed is $l = 17$ (adding the four closing bases i, c, f, j, the additional spanning pair $d{\cdot}e$, and a minimum hairpin loop of three bases). For given values of i, d and e, $Q^x_{i,d,e,s}$ already contains the contributions to $Q^g_{i,d,e,j}$ for all extensible loops of size s except for the two cases when either $L_1 = 4$ or $L_2 = 4$ (which cannot be obtained by extending smaller loops that use a different energy expression). Enriching $Q^x_{i,d,e,s}$ with these two new possible extensible loops, we then convert $Q^x_{i,d,e,s}$ into contributions to $Q^g_{i,d,e,j}$ by introducing the term for closing these loops with pair $i \cdot j$. $Q^x_{i,d,e,s}$ is then extended to provide future values of $Q^x_{i-1,d,e,s+2}$. All other interior loop contributions (cases with either $L_1 \leq 3$ or $L_2 \leq 3$) are then added directly to $Q^b_{i,j}$ using the special energy expressions of the standard model implied by $G^{\text{internal}}_{i,d,e,j}$. Note that the subsequence length l is fixed inside each call to the function "fastiloops". Hence, specifying i implies $j = i+l-1$. For subsequences of length l, we use $Q^x_{i,d,e,s}$ (j implied) to compute $Q^x_{i-1,d,e,s+2}$ ($j+1$ implied) which will later be used to compute contributions to $Q^g_{i-1,d,e,j+1}$ for subsequences of length $l+2$. Thus, for a given value of l, the values of $Q^x_{i,d,e,s}$ need only be stored for all legal values of i, d, e, and s until l has been incremented 3 times, at which point it can be discarded. This is accomplished by using $Q^{x1}_{i,d,e,s}$ and $Q^{x2}_{i,d,e,s}$ to store future contributions for subsequences of length $l+1$ and $l+2$.