

# Quadratic Binary Programming Models in Computational Biology

Richard J. Forrester and Harvey J. Greenberg

Volume 3, Number 2, Fall 2008

URI: [https://id.erudit.org/iderudit/aor3\\_2art02](https://id.erudit.org/iderudit/aor3_2art02)

[See table of contents](#)

Publisher(s)

Preeminent Academic Facets Inc.

ISSN

1718-3235 (digital)

[Explore this journal](#)

Cite this article

Forrester, R. J. & Greenberg, H. J. (2008). Quadratic Binary Programming Models in Computational Biology. *Algorithmic Operations Research*, 3(2), 110–129.

Article abstract

In this paper we formulate four problems in computational molecular biology as 0-1 quadratic programs. These problems are all NP-hard, and the current solution methods used in practice consist of heuristics or approximation algorithms tailored to each problem. Using test problems from scientific databases, we address the question, “Can a general-purpose solver obtain good answers in reasonable time?” In addition, we use the latest heuristics as incumbent solutions to address the question, “Can a general-purpose solver confirm optimality or find an improved solution in reasonable time?” Our computational experiments compare four different reformulation methods: three forms of linearization and one form of quadratic convexification.

# Supplementary Note for *Quadratic Binary Programming Models in Computational Biology*

by Richard J. Forrester and Harvey J. Greenberg

June 11, 2007

## Derivation of the Number of Auxiliary Variables and Equations in RLT of MSA

This supplementary note gives the derivation of the number of auxiliary variables and equations for the RLT linearization of the MSA Problem. These numbers are in the AMPL output from run script `msa-unix.ex`. We ran the MATLAB script, `getdomains.m`, to verify these values.

Notation:

$\ell, \ell'$	sequence indexes
$i, i'$	character position indexes within sequence
$k, k'$	column indexes
$L_\ell$	length of sequence $\ell$
$N_{\max}$	maximum number of columns in MSA
$T_\ell$	$L_\ell (N_{\max} - 1/2(L_\ell - 1))$

Counting the number of auxiliary variables for  $N_{\max}$  columns:

$$\begin{aligned}
 |\text{dom}(w)| &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^m \sum_{i'=1}^{L_{\ell'}} \sum_{k'=i'}^{N_{\max}} \mathbf{1} \\
 &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^m \sum_{i'=1}^{L_{\ell'}} (N_{\max} - i' + 1) \\
 &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^m (L_{\ell'} (N_{\max} + 1) - 1/2 L_{\ell'} (L_{\ell'} + 1)) \\
 &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^m L_{\ell'} (N_{\max} - 1/2 L_{\ell'} (L_{\ell'} - 1)) \\
 &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^m T_{\ell'} \\
 &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} (N_{\max} - i + 1) \sum_{\ell'=\ell+1}^m T_{\ell'} \\
 &= \sum_{\ell=1}^{m-1} T_\ell \sum_{\ell'=\ell+1}^m T_{\ell'}
 \end{aligned}$$

This is equation (8) in the paper.

Counting the number of equations for  $N_{\max}$  columns:

$$\begin{aligned}
\#\text{Eqns} &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{\ell'=\ell+1}^m \sum_{i'=1}^{L_{\ell'}} \sum_{k'=i'}^{N_{\max}} \mathbf{1} \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{\ell'=\ell+1}^m \sum_{i'=1}^{L_{\ell'}} (N_{\max} - i' + 1) \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{\ell'=\ell+1}^m \left( L_{\ell'} (N_{\max} + 1) - \frac{1}{2} L_{\ell'} (L_{\ell'} + 1) \right) \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{\ell'=\ell+1}^m \left( L_{\ell'} (N_{\max} - \frac{1}{2} (L_{\ell'} - 1)) \right) \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{\ell'=\ell+1}^m T_{\ell'} \\
&= \sum_{\ell=1}^{m-1} L_\ell \sum_{\ell'=\ell+1}^m T_{\ell'}
\end{aligned}$$

This is equation (9) in the paper.

Note that  $\#\text{Eqns}$  depends upon the order of the sequence lengths because we restrict the RLT equations to  $\ell' > \ell$ .