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:

ℓ, ℓ'	sequence indexes
i, i'	character position indexes within sequence
k, k'	column indexes
L_ℓ	length of sequence ℓ
$N_{\max}$	maximum number of columns in MSA
T_ℓ	$L_\ell \left(N_{\max} - {}^1\!/_2(L_\ell - 1)\right)$

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

$$
\begin{aligned}
|\mathrm{dom}(w)| &= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^{m} \sum_{i'=1}^{L_{\ell'}} \sum_{k'=i'}^{N_{\max}} \mathbf{1} \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^{m} \sum_{i'=1}^{L_{\ell'}} (N_{\max} - i' + 1) \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^{m} \left(L_{\ell'}\,(N_{\max}+1) - {}^1\!/_2 L_{\ell'}(L_{\ell'}+1)\right) \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^{m} L_{\ell'} \left(N_{\max} - {}^1\!/_2 L_{\ell'}(L_{\ell'}-1)\right) \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} \sum_{k=i}^{N_{\max}} \sum_{\ell'=\ell+1}^{m} T_{\ell'} \\
&= \sum_{\ell=1}^{m-1} \sum_{i=1}^{L_\ell} (N_{\max} - i + 1) \sum_{\ell'=\ell+1}^{m} T_{\ell'} \\
&= \sum_{\ell=1}^{m-1} T_\ell \sum_{\ell'=\ell+1}^{m} T_{\ell'}
\end{aligned}
$$

This is equation (8) in the paper.

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

$$\begin{aligned}
\#\text{Eqns} &= \sum_{\ell=1}^{m-1}\sum_{i=1}^{L_\ell}\sum_{\ell'=\ell+1}^{m}\sum_{i'=1}^{L_{\ell'}}\sum_{k'=i'}^{N_{\max}}\mathbf{1} \\
&= \sum_{\ell=1}^{m-1}\sum_{i=1}^{L_\ell}\sum_{\ell'=\ell+1}^{m}\sum_{i'=1}^{L_{\ell'}}(N_{\max}-i'+1) \\
&= \sum_{\ell=1}^{m-1}\sum_{i=1}^{L_\ell}\sum_{\ell'=\ell+1}^{m}\Big(L_{\ell'}(N_{\max}+1)-{}^1\!/_2 L_{\ell'}(L_{\ell'}+1)\Big) \\
&= \sum_{\ell=1}^{m-1}\sum_{i=1}^{L_\ell}\sum_{\ell'=\ell+1}^{m}\Big(L_{\ell'}\big(N_{\max}-{}^1\!/_2(L_{\ell'}-1)\big)\Big) \\
&= \sum_{\ell=1}^{m-1}\sum_{i=1}^{L_\ell}\sum_{\ell'=\ell+1}^{m}T_{\ell'} \\
&= \sum_{\ell=1}^{m-1}L_\ell\sum_{\ell'=\ell+1}^{m}T_{\ell'}
\end{aligned}$$

This is equation (9) in the paper.

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