**Approach A (G4 Hunter) :**

At it stands, this is the worst predictor among the three methods, which is somehow good for us as we want to convey the message that it is not optimal. However, the C positive / T negative approach is not the most appropriate model.

Can you do the following changes/tests:

* *Traditional* G4-Hunter: give +1, +2, +3 or +4 to each cytosine in a run of 1, 2, 3 or 4+ consecutive C and a score of 0 to all T. (not negative)
* *Extended* G4-Hunter: give +1, +2, +3, +4, +5, +6to each cytosine in a run of 1, 2, 3, 4, 5 or 6+ consecutive C and a score of 0 to all T. (not negative)
* *Optimized* G4-hunter: give +w, +x, +y or +z to each cytosine in a run of 1, 2, 3 or 4 (with w, x, y, z positive decimal numbers and w<x<y<z) consecutive C and a score of 0 to all T. (not negative)
* *Optimized*+Extended G4-hunter: give +u +v +w, +x, +y or +z to each cytosine in a run of 1, 2, 3, 4, 5, 6+ (with u, v, w, x, y, z positive decimal numbers and u<v<w<x<y<z) consecutive C and a score of 0 to all T. (not negative)

No base is given a negative score – we don’t have any G in the sequence space tested.

**Approach B (Gradient boosting machines)**

You mention that Features are not extremely correlated (slide 4) but total length is actually fairly correlated with C (which makes sense, given that Cytosines constitute the major component of the sequence).

My advice would be to drop length completely – L = (Cx4+T1+T2+T3); it’s not an independent variable.

The predicted vs observed correlation is superb even without Length (r2=0.98 for Tm; 0.95 for pHT). I would keep the graph on the right for the paper

**Approach C:** **Non-Linear Analytical Equation**

I like it very much too.

Slide 13: I tried to extrapolate this formula to other sequences motifs, to find sequences with pHT close to 7.0... you can approach this value with C=11, T2 long, and T1 and T3 short. It makes complete sense !

In that case I am fine with keeping “length” which makes the equation simplier than (Cx4+T1+T2+T3)...