Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary Information

On the delocalization length in RNA single strands of cytosine: How many bases see the light?

*Nykola C. Jones, Steen Brøndsted Nielsen and Søren Vrønning Hoffmann**

Department of Physics and Astronomy, Aarhus University
DK-8000 Aarhus C, Denmark
*Email: vronning@phys.au.dk

SM1: Strand length effect on the spectral shape of the CD spectra

The model used in this paper, where the CD signal, CD(X_n), of an n-long strand of nucleotide X is broken down into nearest neighbour coupling, a_1, next nearest neighbour, a_2 and so on, results in Eqns. 1 and 2 for describing the CD signal at a single wavelength. However, this model does not result in wavelength independent fitting parameters, and is thus not a representation of the evolution of the entire CD spectrum with strand length. In other words, the CD spectra are not directly scalable to each other. This is shown in Fig. SM1-1, where the CD spectra of Fig. 1 are scaled to have the same value at 278 nm. Both the overall shape and the magnitude of the CD spectra features are not scalable and the zero crossings of the spectra change with strand length. In general, the part of the CD spectrum originating from nearest neighbour, next nearest neighbour, and so on, interactions should not just be a scalable to each other as they involve different types of geometries and thus couplings.

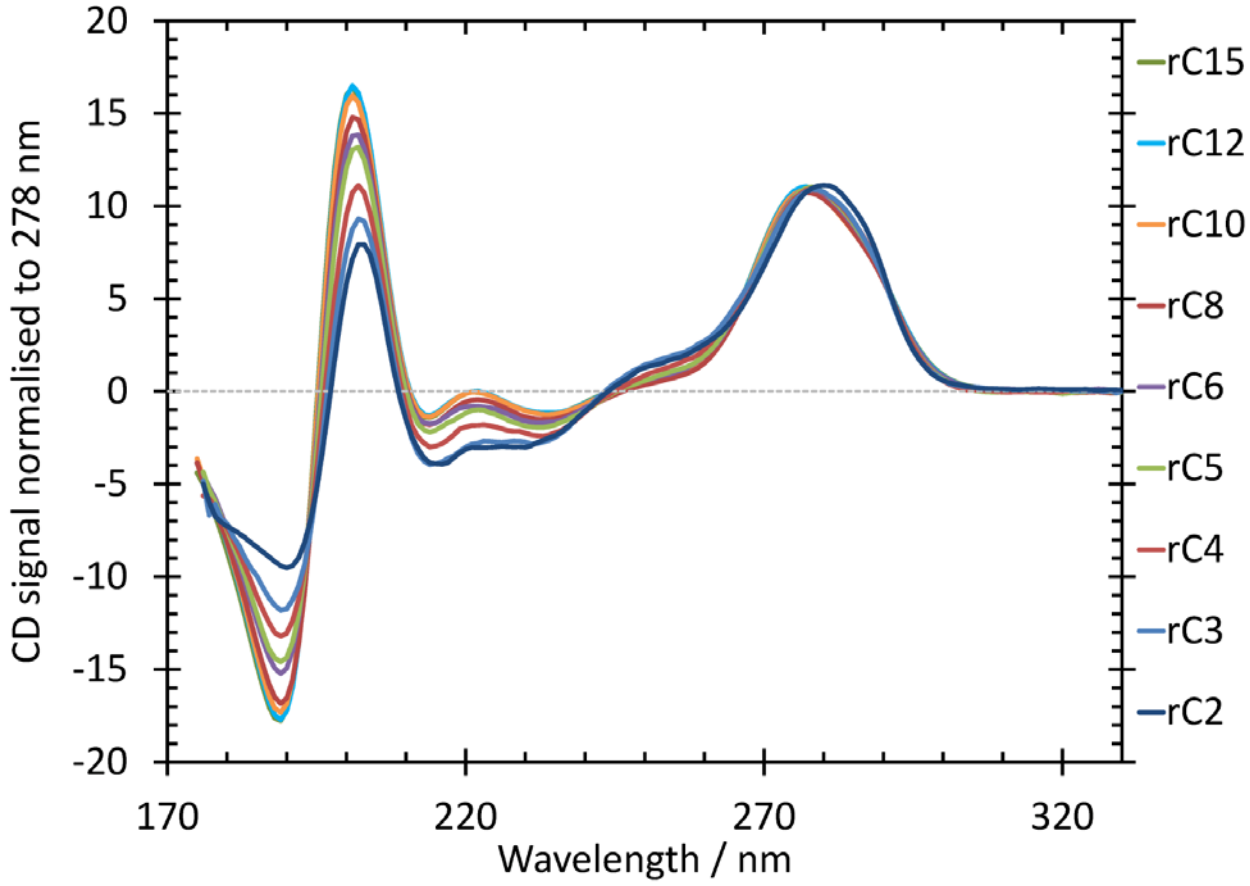

Figure SM1-1: The CD spectra of Fig. 1 all scaled to have the same signal at 278 nm.

SM2: Effect due to concentration errors

Concentrations play an important role in the analysis of the data in this paper, and are based on photo-absorption measurements at 260 nm together with a model for the extinction coefficients for varying lengths of the $(rC)_n$ strands. The extent to which inaccurate concentration determinations influence the conclusions is analysed in this section. Two approaches are used: 1) errors in the nearest neighbour model used in this paper and 2) and overall uncertainty in the concentrations.

Errors in the nearest neighbour model

Concentrations used in this paper are determined via a nearest neighbour model. In this model, the major contribution to the extinctions comes from the interaction between nearest neighbours, in this case from CpC pairs. As an example, a $(rC)_4$ strand of four bases has three nearest neighbour CpC pairs, so a first approximation to the extinctions coefficient is $3 \cdot \varepsilon_{CpC}(260\ nm)$. However, this counts two pC bases twice, which is corrected by subtraction of a corresponding extinction coefficient for the individual bases $2 \cdot \varepsilon_{pC}(260\ nm)$. In total this gives

$$\varepsilon_{(rC)_4}(260\ nm) = 3 \cdot \varepsilon_{CpC}(260\ nm) - 2 \cdot \varepsilon_{pC}(260\ nm)$$

In general, the extinction coefficient for a strand length *n* is calculated using

$$\varepsilon_{(rC)_n}(260\ nm) = (n-1) \cdot \varepsilon_{CpC}(260\ nm) - (n-2) \cdot \varepsilon_{pC}(260\ nm)$$

Using the nearest-neighbour model is considered to be the most accurate method for determining the extinction coefficients at 260 nm as long as the pH is near neutral. The method has been claimed to have an average error of about 2-5% [Cavaluzzi2004] if the correct extinction coefficients are used. The nearest neighbour model parameters used in the calculations of the extinction coefficient in Table 1 is 7400 M^{-1} cm^{-1} for pC and 14200 M^{-1} cm^{-1} for CpC, very close to the parameters used by Tataurov and Owczarzy [Tataurov2008] of 7200 M^{-1} cm^{-1} for pC. Historically the parameter for pC has been as high as 7600 M^{-1} cm^{-1}, which differs by 7% from the values claimed by Cavaluzzi *et al.* [Cavaluzzi2004] of 7060 M^{-1} cm^{-1}, using high precision concentration determinations with NMR.

The effect on the conclusions in this paper could potentially be significant, in the most extreme case it might be possible that non-linear behaviour at low strand lengths in Fig. 2 becomes linear. Shown in Fig. SM2-1 are the 188 nm data points vs. strand length for the original concentrations used in Fig. 2 as well as for concentrations calculated for the extreme $\varepsilon_{pC}(260\ nm)$ extinctions coefficients found in the literature of 7060 M^{-1} cm^{-1} and 7600 M^{-1} cm^{-1}. The same data for 201 nm and 278 nm are found in Figs. SM2-2 and SM2-3. Note that the concentration corrections for $\varepsilon_{pC}(260\ nm)$ = 7060 M^{-1} cm^{-1} compared to the concentrations used in this paper changes from 0% to 4.1% for *n* = 2 to *n* = 15 whereas for $\varepsilon_{pC}(260\ nm)$ = 7600 M^{-1} cm^{-1} corrections change from 0% to -2.6%. As is evident in figures SM2-1 to SM2-3, the data points for low *n* values are still far from linear, and the linear fit to the high *n* values ($n \geq 6$) crosses zero in almost the same point. The zero crossing point is 3.10±0.07 for 188 nm and 201 nm and 2.45±0.03 for 278 nm.

It is clear that uncertainties in the parameters of the nearest neighbour model do not translate into any significant change in the zero crossing values. In addition, the non-linear behaviour of the low *n* CD data is unaltered.

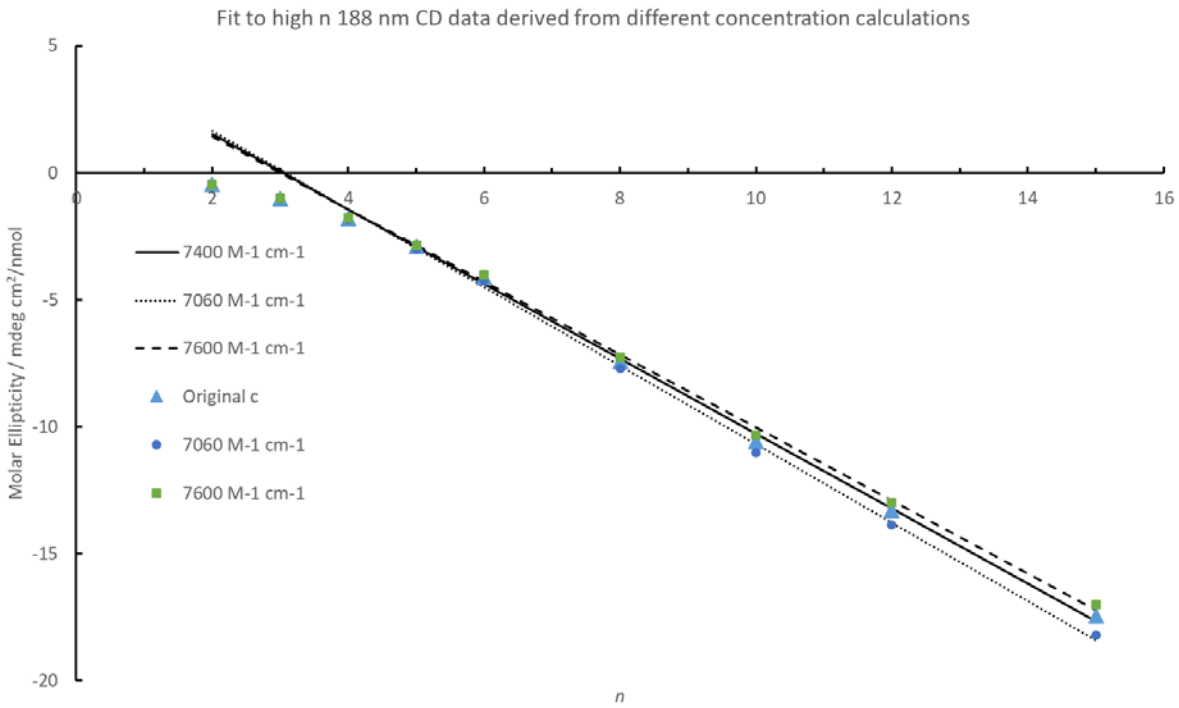

Figure SM2-1: The 188 nm CD signal vs. strand length (*n*) for the original concentrations in the main paper based on $\varepsilon_{pC}(260\ nm)$ = 7400 M^{-1} cm^{-1} (triangles) as well as for $\varepsilon_{pC}(260\ nm)$ = 7060 M^{-1} cm^{-1} (circles) and for $\varepsilon_{pC}(260\ nm)$ = 7600 M^{-1} cm^{-1} (squares). The lines are linear fits to the high *n* values ($n \geq 6$).

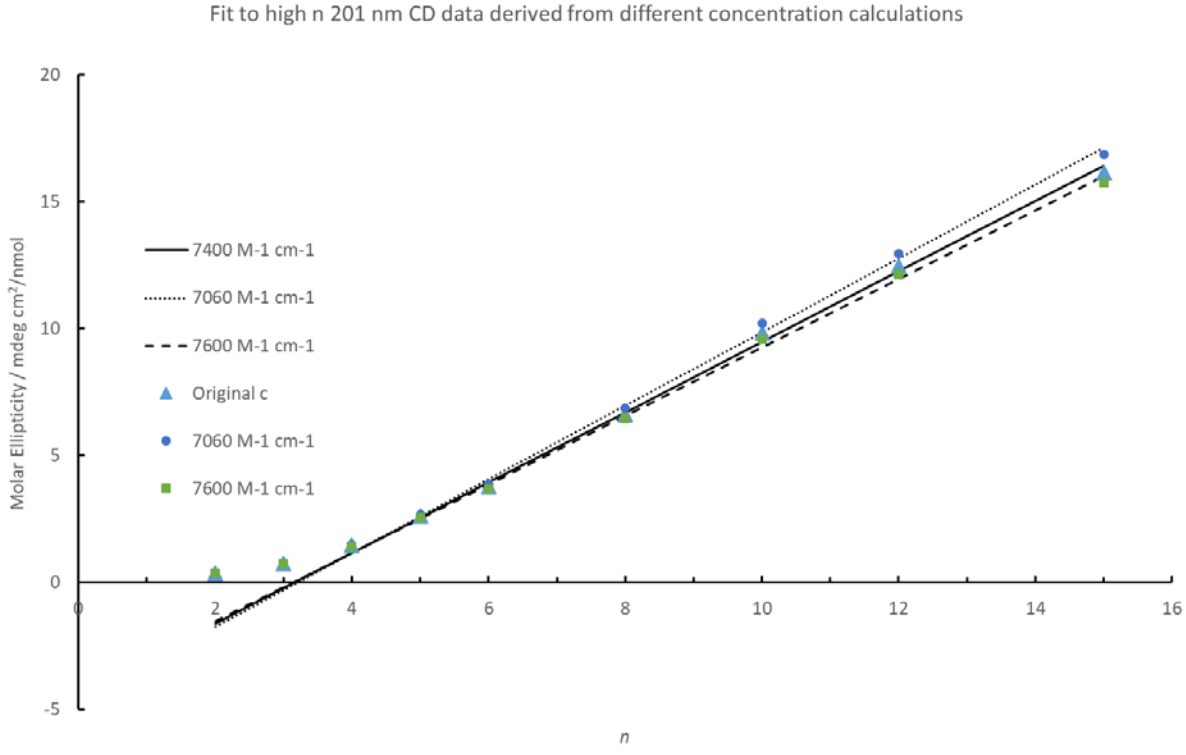

Figure SM2-2: The 201 nm CD signal vs. strand length (*n*) for the original concentrations in the main paper based on $\varepsilon_{pC}(260\ nm)$ = 7400 M^{-1} cm^{-1} (triangles) as well as for $\varepsilon_{pC}(260\ nm)$ = 7060 M^{-1} cm^{-1} (circles) and for $\varepsilon_{pC}(260\ nm)$ = 7600 M^{-1} cm^{-1} (squares). The lines are linear fits to the high *n* values ($n \geq 6$).

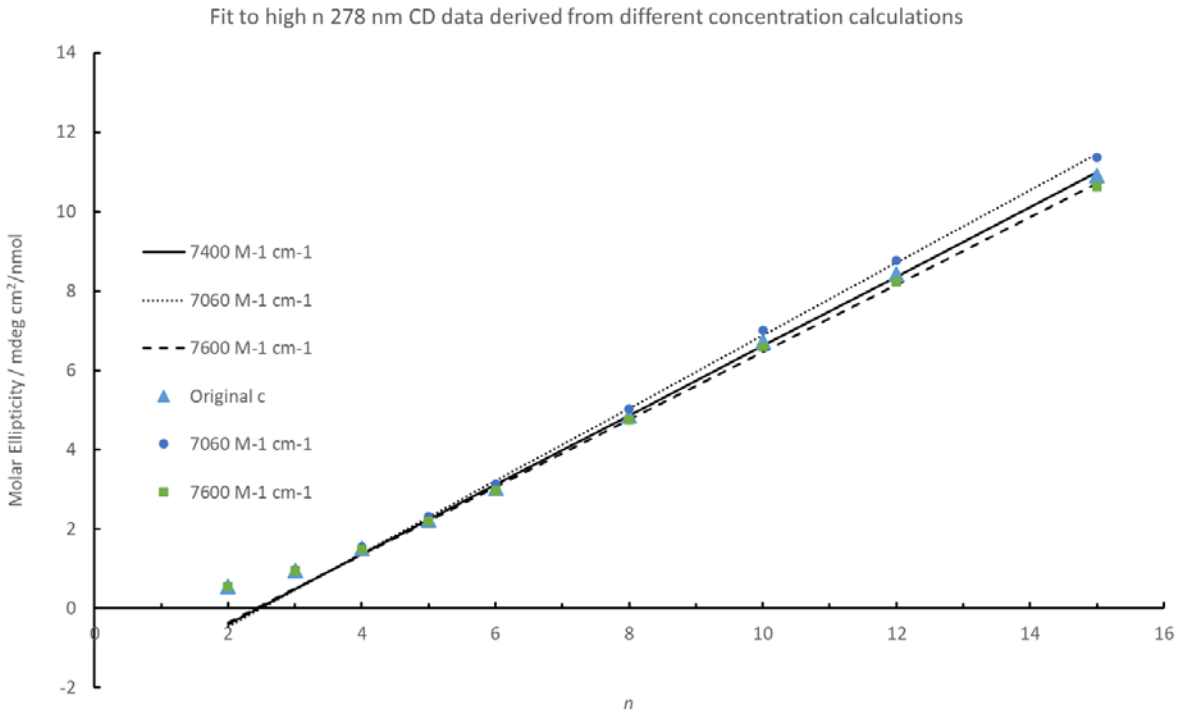

Figure SM2-3: The 278 nm CD signal vs. strand length (*n*) for the original concentrations in the main paper based on $\varepsilon_{pC}(260\ nm)$ = 7400 M^{-1} cm^{-1} (triangles) as well as for $\varepsilon_{pC}(260\ nm)$ = 7060 M^{-1} cm^{-1} (circles) and for $\varepsilon_{pC}(260\ nm)$ = 7600 M^{-1} cm^{-1} (squares). The lines are linear fits to the high *n* values ($n \geq 6$).

Overall uncertainty in the concentrations

Further analysis of how concentration determination errors, which may influence the conclusions of the paper, is made via a large (±10%) uniform scaling of all the concentrations, i.e. a non-strand length dependent signal change. The results of this analysis are shown in Figs. SM2-4, SM2-5 and SM2-6 for the 188 nm, 201 nm and 278 nm CD data, respectively. Just as for changes in the parameters in the nearest neighbour model, which resulted in a strand length dependent concentrations change, a uniform scaling of 10% does not alter the zero crossing in any significant way with a value of 3.10±0.06 for 188 nm and 201 nm and 2.45±0.00 for 278 nm.

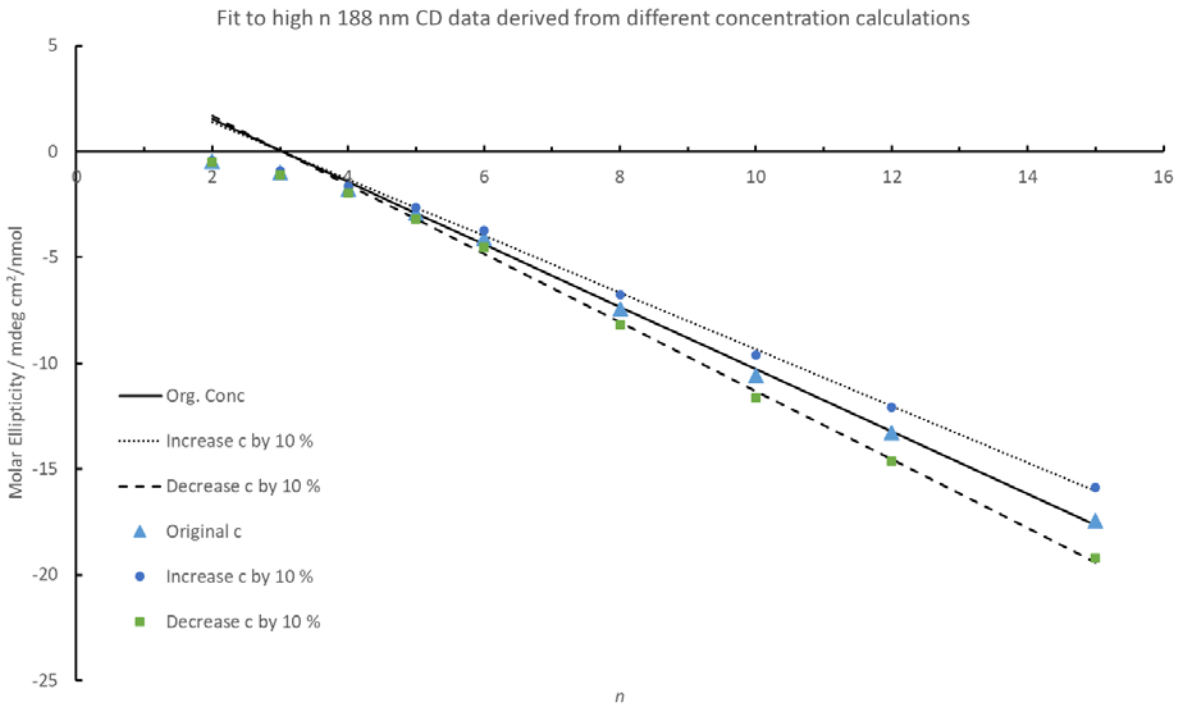

Figure SM2-4: The 188 nm CD signal vs. strand length (*n*) for the original concentrations in the main paper as well as a concentration scaling of +10% (circles) and -10% (squares). The lines are linear fits to the high *n* values ($n \geq 6$).

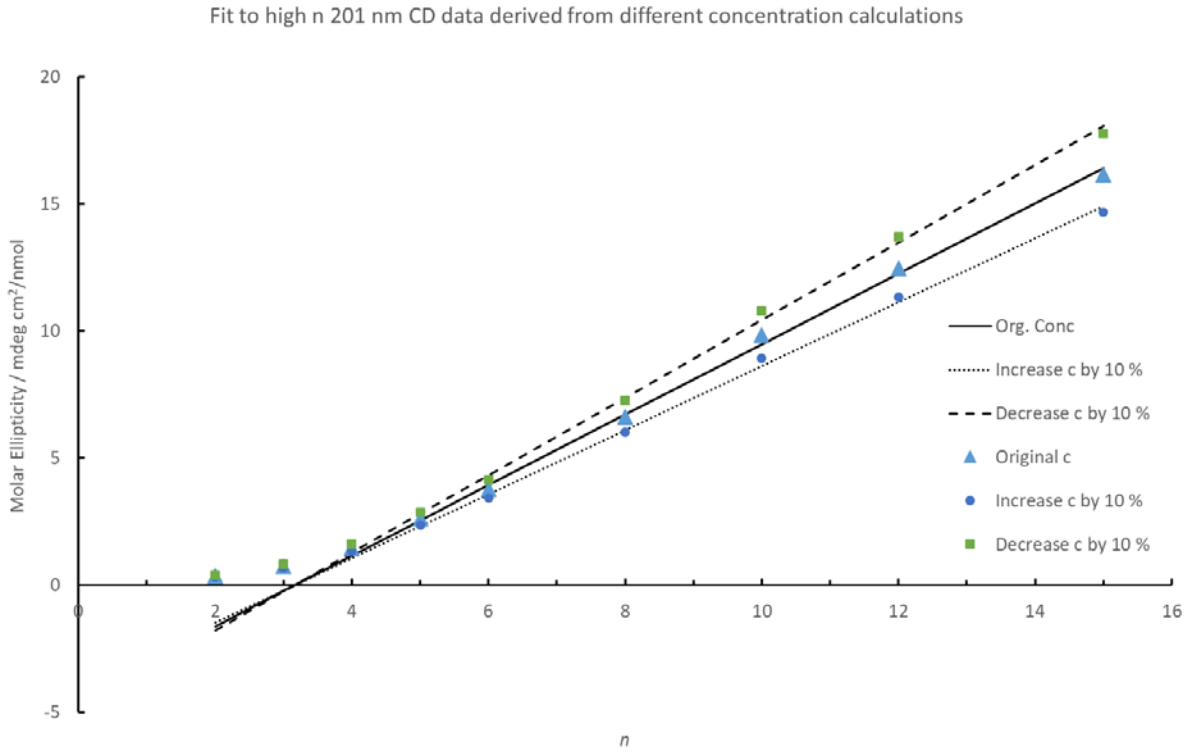

Figure SM2-5: The 201 nm CD signal vs. strand length (*n*) for the original concentrations in the main paper as well as a concentration scaling of +10% (circles) and -10% (squares). The lines are linear fits to the high *n* values ($n \geq 6$).

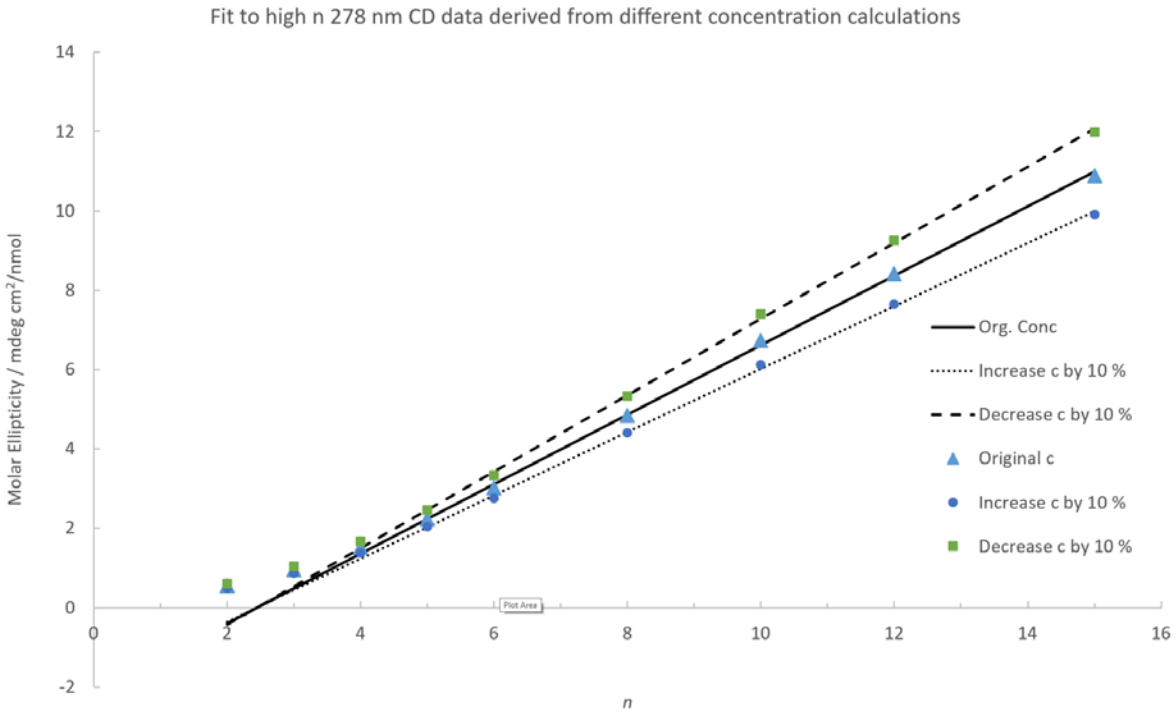

Figure SM2-6: The 278 nm CD signal vs. strand length (*n*) for the original concentrations in the main paper as well as a concentration scaling of +10% (circles) and -10% (squares). The lines are linear fits to the high *n* values ($n \geq 6$).

Overall, the influence of concentration errors on the conclusions drawn in the paper is not significant. In this connection, it is interesting to draw parallels to the study of strands of the DNA form of adenine $(dA)_n$ [Kadhane2008]. Here the behaviour of the CD signal with strand length was very different for wavelengths below and above 200 nm. In the latter case, the CD signal scales linearly with strand length and crossed at n=1, whereas the non-linear model of Eqns. 1 and 2 was needed below 200 nm. If there were large concentration errors, which were non-constant with strand length, this could have led to a non-linear CD change with strand length for the CD data above 200 nm. The fact that the data were linearly scalable above 200 nm and not scalable below 200 nm, demonstrates that the influence of errors in the nearest neighbour model for extinction coefficients were not significant in that study. The above analysis confirms this for the present data on RNA strands of cytosine.

SM3: The CD signals vs strand length

$(rC)_n$ Strand length	Molar Ellipticity (mdeg cm^2/nmol) 188 nm	201 nm	278 nm
2	-0.46	0.35	0.55
3	-1.01	0.75	0.95
4	-1.79	1.46	1.52
5	-2.91	2.61	2.24
6	-4.12	3.75	3.03
8	-7.43	6.60	4.85
10	-10.59	9.81	6.74
12	-13.30	12.44	8.42
15	-17.46	16.14	10.89

References

Cavaluzzi2004

Michael J. Cavaluzzi and Philip N. Borer, *Nucleic Acids Research*, 2004, **32**, e13.

Tataurov2008

A. V. Tataurov, Y. You and R. Owczarzy, *Biophys. Chem.*, 2008, **133**, 66-70.

Kadhane2008

U. Kadhane, A. I. S. Holm, S. V. Hoffmann and S. Brøndsted Nielsen, *Phys. Rev. E*, 2008, **77**, 021901