Interaction with DNA – Data treatment aspects

AUTHORS: Members of WORKING GROUP 2 (WG2 - Studies on strong and/or multifunctional ligands, macromolecules, polyelectrolytes) TASK GROUP 2 (TG2 – DNA BINDING)*

1 – INTRODUCTION

Our work on datasets production and analysis (to get **K**, binding constant, and **n**, DNA site size in base pairs) followed the following steps (where EB = Ethidium bromide and CT-DNA = CALF THYMUS DNA).

1) Inter-laboratory exercise where data already present in some of the research labs on EB/CT-DNA titrations are shared and analysed in search for binding parameters.
2) Inter-laboratory exercise where some research labs carried out EB/CT-DNA titrations using the same conditions (**BUFFER = 0.1 M KCl, 0.01M Hepes, pH 7.4, 25.0 °C**), which are shared and analysed in search of binding parameters.
3) Inter-laboratory exercise where some research labs carried out EB/CT-DNA titrations using **NOT ONLY** the same conditions (**BUFFER = 0.1 M KCl, 0.01M Hepes, pH 7.4, 25.0 °C**) but the **SAME PROTOCOL (see the Document "Interaction with DNA - GUIDELINES)**; these titrations are shared and analysed in search for binding parameters.

As for the way the data are analysed, the different options considered are listed below.

I. Excluded site model ("neighbour exclusion") developed by McGhee and von Hippel (J. D. McGhee and P.H. von Hippel, https://doi.org/10.1016/0022-2836(74)90031-X) and Crothers (D. M. Crothers, https://doi.org/10.1002/bip.1968.360060411) and relevant McGhee&vonHippel Equation (see **ANNEX A**) to be used at a single wavelength.
II. Numerical iterative solution of the exact equations describing the simple host-guest model. This approach has been implemented using the MSExcel® solver tool. It is similar to the approach of the free online tool by Thordarson (IV), but the prepared spreasheet, which minimizes the differences between experimental and theoretical binding isotherms, enables it to directly fit over both **K** and **n**.
III. Solution of the system through the general systematic approach to any set of equilibria. This requires that the CT-DNA concentration is divided by the "**n**" constant, according to a procedure where the aim is to find the optimum **n** (and relevant **K**) by inspecting the minimum error parameters of the fit. The software used was HypSpec (http://www.hyperquad.co.uk/HypSpec.htm) and the free online tool K-ev (https://k-ev.org/).
IV. Free online Thordarson software (BINDFIT at http://supramolecular.org), with the same procedure described in III to find **n**. Thordarson's approach applies only to one type of equilibrium.

**Italy: T. Biver, F. Binacchi, University of Pisa; G. Barone, A. Terenzi, University of Palermo. Hungary: E.A. Enyedy, O. Dömötör, University of Szeged. Portugal: N. Basílio, NOVA University of Lisbon; I. Correia, N. Ribeiro, Instituto Superior Tecnico, University of Lisbon; I. Cavaco, University of Algarve. Spain: N. Busto, University of Burgos; E. Garcia-España, J. Gonzalez, University of Valencia. France: J. Hamacek, CNRS Orleans*

2 – n MEANING AND FIT

n is the DNA binding site size. If DNA concentration is provided in BASE PAIRS, **n** is the number of DNA base pairs that (under saturation conditions) are occupied by one single binding molecule. Note that **n** is not supposed to be strictly an integer, due to the inhomogeneity of natural DNA (AT/GC base pairings), mixtures of binding modes, and inherent complexity of systems containing polynucleotides. It is an important parameter not only for correct data treatment but it also has practical implications. For instance, for intercalators the value of **n** may be connected to the helix distortion/unwinding produced (see also https://doi.org/10.1016/S0300-9084(71)80064-0). Note that, while **n** may somehow look like a stoichiometric value, it is not the case. If, for instance, n = 2, we will not be in the presence of two equilibria and two ($K_{1:1}$, $K_{1:2}$) binding constants. The binding will always be **1:1 WITH THE SITE OF DNA.** It means that we will have only one K value to determine. **This model is a simplification**, as more complex theories account for a change of K value upon dye/drug binding (as a function of DNA saturation), for instance defining a "nucleation step" and a "propagation step" (https://doi.org/10.1016/j.jinorgbio.2006.11.009). Also, different classes of DNA sites may be considered for a correct definition of the system (see the book by Cantor and Schimmel "Biophysical Chemistry PART III"). However, these more complicated models are not considered in this work, which focuses on highlighting weak and strong points of simpler and widely used data fitting. Note that the latter, even if simplified, ensures a robust picture of the main aspects of the binding.

Let's go back to 1:1 binding with the site (**n** base pairs) of DNA. If one class of sites is present, the concentration of DNA IN SITES will be C_{DNA}/**n**, where C_{DNA} is the total molar concentration of the nucleic acid in base pairs. The approach to evaluate **n** (and use the data set from the experiment) is different for the different I-IV approaches cited before.

I. Uses only the binding isotherm at one selected wavelength. The fit returns the **K** and **n** as outputs. The DNA concentration is not simply C_{DNA}/**n** but a more complex function (f(r), see ANNEX A) that accounts for some statistical factors for dye/drug rearrangement over the DNA base pairs.
II. Uses only the binding isotherm at one selected wavelength. The fit returns the **K** and **n** outputs. The DNA concentration is defined as C_{DNA}/**n**.
III. Uses all the wavelengths uploaded and simultaneously fits over all of them. The fit returns only the **K** value according to a 1:1 binding model. **n** is calculated separately: the user will upload different C_{DNA}/**n** values and search for the **n** value which minimizes the residuals of the fit.
IV. Same as III.

This document will focus on the K values obtained by the different approaches. We will not discuss here the **n** values in detail as this additional parameter, unfortunately encounters very high bias and it was decided to focus on it in the second step of our NECTAR work.

3 – RESULTS

It was concluded that dataset (1) – inter-laboratory dataset collected previously under different conditions - contains such different conditions (buffer, temperature, concentration ranges) that the numbers for **K** and **n** evaluated cannot be compared. Note that the conditions are not really dramatically different, as the buffer will be one of those commonly used to mimic physiological conditions (TRIS, Hepes, phosphate or

cacodylate) with a narrow pH range (7.0-7.4); also the temperature will be around r.t. (20-25 °C) and the concentrations will not differ by magnitude orders, being in line with the common fluorescence instrumental needs. However, this still produces a very high variability of results. It's worse reminding that, even in the case of small differences in the buffers, this reflects on differences in the ionic strength of the medium and ionic strength has a dramatic effect on the EB-DNA binding affinity (https://doi.org/10.1016/0022-2836(67)90353-1). Figure 1A shows the distribution of K values obtained from the different groups which tried to analyse the data according to some of the I-IV approaches. According to these first results, it was concluded that dataset (1) is too inhomogeneous to be further considered. As for dataset (2) - inter-laboratory dataset using the same conditions - it was for sure more homogeneous but produced the results shown in Figure 1B, which were still considered to be not sufficiently consistent.

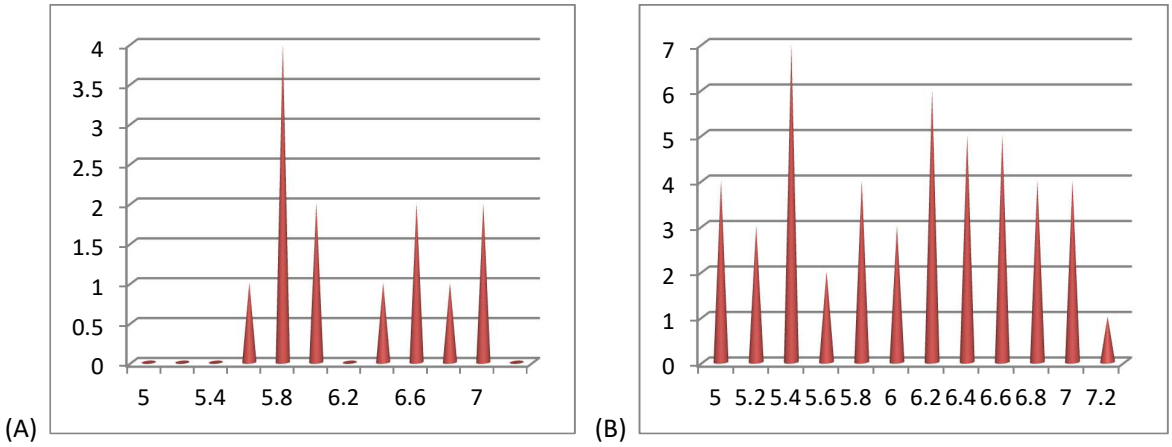

Figure 1 – Histogram of the results from (A) dataset (1) or (B) dataset (2); x-axis = log K; y-axis = counts.

Dataset (3) - inter-laboratory dataset using the same conditions and protocol - is supposed to be the more reliable. Figure 2 shows that, after some normalization, the binding isotherms of the different groups turn out to be very reproducible.

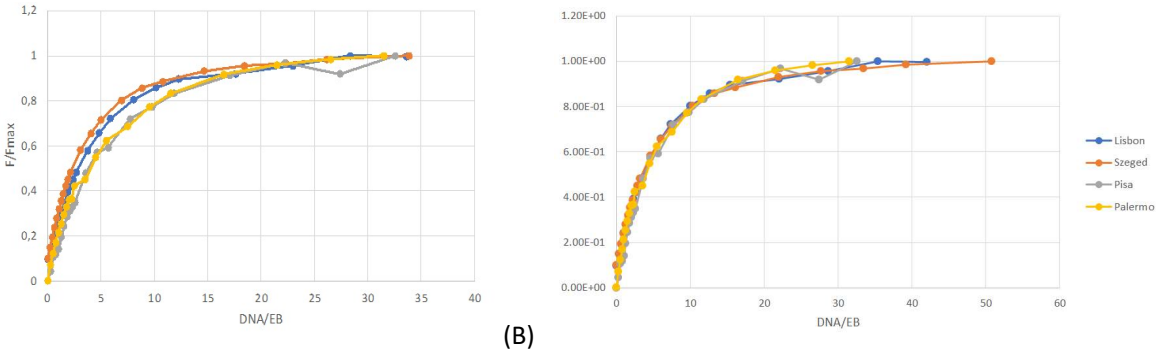

Figure 2 – Binding isotherms at fixed emission wavelength for EB/DNA titrations from different research labs: (A) normalised to maximum intensity (F); (B) normalised to maximum intensity (F) and adjusted by a factor of 1.25 (Lisbon) and 1.5 (Szeged) to optimize overlap.

Figure 3 (A) and (B) show the results for dataset (3). Figure 3A collects the usual data according to I-IV approaches, Figure 3B focuses on the data where the EXCEL solver approach is used but where the f(r) function is used to correct for DNA concentration according to statistical factors and base pairs' saturation degree. Clearly, in Fig. 3B more reproducible values are obtained.

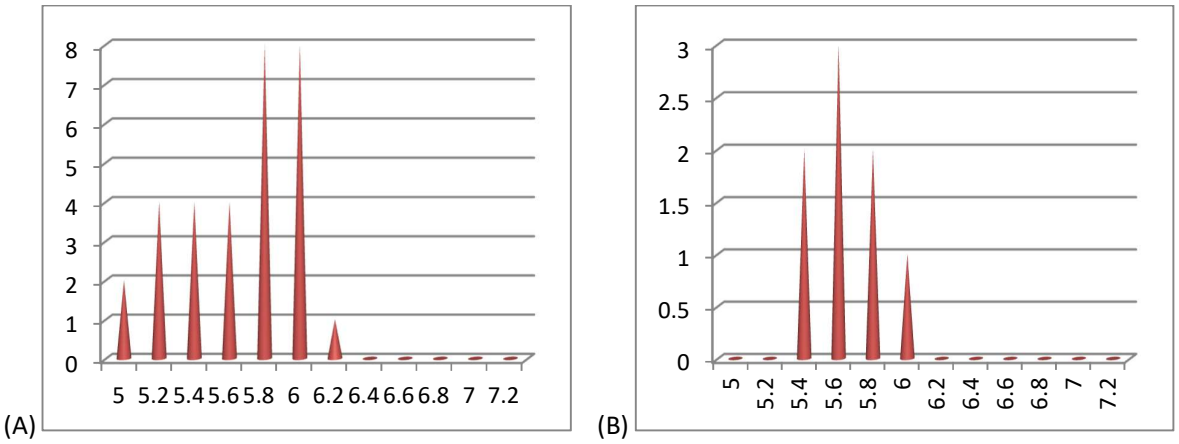

Figure 3 - Distributions of log K values (x-axis = log K; y-axis = counts) according to the numbers obtained by different calculation methods by WG-TG2 participants on dataset (3). (A) Calculation methods I-IV; (B) calculation method is IV with DNA concentration corrected for the f(r) function.

A comparison of Figure 2 and Figure 3 demonstrates that the main problem of lack of reproducibility for K values is now not the Experimental procedure, but the choice for the Data Treatment procedure.

4 – STRENGTHS AND WEAK POINTS OF THE CALCULATIONS

4.I Scatchard and McGhee&vonHippel models

- **PROS** The model, that takes into account occupancy of multiple binding sites, is called excluded site model ("neighbour exclusion") and an extended form of this model introduces a cooperativity factor (ω) to account for the interaction between bound drug molecules, with $\omega > 1$ corresponding to the cooperative and $\omega < 1$ to the anti-cooperative binding. The neighbour exclusion model contains the correction factor f(r) which accounts for statistical effects related to possible bound molecule rearrangements, whose extent is a function of DNA saturation degree (r). Overall, this model better fits the complexity of the EB/DNA system.
- **CONS** In Scatchard plots, the problem of point linearization exists, as well as that of the choice of the points to be used. This approach uses the data at a single wavelength (even if the creation of an Excel solver routine to use more than one wavelength would not be difficult to do). These equations, based on DNA saturation degree, can be applied only to titrations where DNA is added to EB and calculations need the plateau to be perfectly measured for an evaluation of the optical parameters related to bound EB only. The choice of points to to be fitted with linear regressionis crucial and not easily reproduced.

4.II Excel Solver

- **PROS** Very widely used, not exactly freeware but everyone has it. Easily handled, customized and improved. The titrations can be done in both EB/DNA or DNA/EB ways, affording a more extended range of reactant ratios to be considered. If it does not consider f(r) corrections, these may be added, more easily than in a software.
- **CONS** This approach uses data at a single wavelength (even if the creation of an Excel solver that uses more than one wavelength would not be so difficult to afford). Some tests done to add f(r) correction changed logK values in a non-unique way (decreasing or increasing it) and lowering **n** so that they become farther away from the 2.5 reference value: doubts arise on the efficacy of the correction.

4.III HypSpec Software and K-ev

- **PROS** HypSpec is a widely diffused software in our community. It uses an iterative procedure where many wavelengths and different titrations can be fitted at the same time. The titrations can be done in both EB/DNA or DNA/EB ways, affording a more extended range of ratios to be considered, with possible higher accuracy of **n** determination. Both apply a systematic approach to solving equilibrium which, in theory, can be applied to any system.
- **CONS** HypSpec is not freeware and exists only for the MSWindows operative system. There is uncertainty about the future purchase of this software and its upgrades to meet new hardware standards. K-ev is freely available online for now, but based on Russia. It can be downloaded as an R package but this is not a user friendly solution. The software does not enable itself correction for the f(r) function (which contains **n**). The procedure to divide DNA concentration by **n** and search for the minimum of the experimental-calculated fit parameters is sometimes troubled by the absence of sharp minima.

4.IV Other Software (Supramol/Thordarson)

- **PROS** Freeware, online. Both use an iterative procedure where many wavelengths can be simultaneously fitted. The titrations can be done in both EB/DNA or DNA/EB ways, affording a more extended range of reactant ratios to be considered.
- **CONS** No correction for site rearrangement statistical factors. They seem sometimes like black boxes where the exact way data are is analysed is not clear. In particular, Supramol/Thordarson website considers options to be ticked whose effect is not clear.

5 – STATISTICAL ASPECTS

5.1 Sources of uncertainty Sources of uncertainty will affect the final results differently depending on the calculation method. There are two "families" of calculation approaches which can be considered:

1) linearization methods: McGee and von Hippel. This consists in transforming the measured signal and total concentrations into variables which fit a linear equation. The values of K and binding site size can be calculated from the slope and intercept of the best line fitting the data;

2) iterative methods: Excel solver, Hypspec, PSEQUAD, Thordarson, K-ev. These consist in solving a non-linear system of equations (mass balances and equilibrium constants) using an iterative method.

From the experimental point of view, the sources of uncertainty may be those collected in Table 1.

Table 1. Comments on some experimental sources of uncertainty.

Source	Value	Estimated uncertainty
EB purity	HPLC grade: >95.0% (sigma-aldrich) Molecular biology grade: 98% (Alfa Aesar)	u_r=0.0144 (type B) [2]
EB stock concentration	~1 mM	Depends on mass weight and volume. (Estimated as ~10%)
EB concentration		Depends on measured volumes and pipette. (estimated as ~14%)
CT-DNA purity	The highest level of purity contains 0.7% protein.[1]	The presence of protein is probably not relevant, but molecular weight may affect DNA coiling and the average binding site size
CT-DNA stock concentration	Determined from ε(260 nm)= 13200 $M^{-1}cm^{-1}$ (bp) ?	In bibliography, $\mathcal{E}$ varies from 6412 to 6900 (in nucl $M^{-1}cm^{-1}$). A range of 976 (bp $M^{-1}cm^{-1}$) could be translated to a sd of 474 (d_2=2.059 for n=4), or a u_r of 7.2%
CT-DNA concentration (M)	3.88 x10^{-7} - 3.35 x10^{-4}	Depends on measured volumes and pipette. (estimated as ~14%)
Signal (Fluorescence intensity) noise		Estimated as s_y or s_b from linear regression of EB calibration
Temperature	25ºC	
Ionic strength	0.1 M KCl	
pH	7.4	Typically +/- 0.1 units

[1] Welsh RS, Vyska K. Relationship between the purity and molecular weight of calf thymus DNA. Hoppe Seylers Z Physiol Chem. 1981 Jul;362(7):969-81. DOI: 10.1515/bchm2.1981.362.2.969. PMID: 7275016.

[2] Type B approach, estimating 95% purity as a rectangular distribution with an amplitude of 5%, $u = \frac{a}{2\sqrt{3}}$

We can estimate the effect of the uncertainty in the concentrations of EB and CT-DNA by measuring how logK varies when the values of C_{EB} and C_{DNA} are replaced with their values at 95% confidence limits. Based on the data from Szeged (Table 2), **this can easily justify a range of 0.4 in logK.**

Table 2. Values of logK obtained varying C_{EB} and C_{DNA} within their 95% confidence limits. Calculations with a systematic approach method (k-EV) using data for the direct titration from Szeged. C_{EB}= (1.00±0.28)x10^{-4} M and C(DNA)=(1.10±0.29)x10^{-4} M.

			C_{EB} (M)		
			min	**mean**	**max**
		LogK (sd)	***7.17E-05***	***1.00E-04***	***1.28E-04***
C_{DNA} (M)	***max***	***1.39E-04***	5.078 (0.0077)	5.050 (0.0086)	5.013 (0.013)
	mean	***1.10E-04***	5.196 (0.0078)	5.174 (0.0086)	5.143 (0.013)
	min	***8.14E-05***	5.355 (0.0080)	5.345 (0.0086)	5.327 (0.013)

5.2 Dataset (2) Pisa, Szeged, Burgos, Lisbon and Valencia performed direct and inverse titrations of EB with CT-DNA, using the same experimental conditions, previously agreed: T = 25.0 ºC, Buffer: 10 mM HEPES, 0.1 M KCl, pH 7.4, C_{EB} from 1 to 10 µM. The experimental data from the five groups were analysed using different calculation methods already cited. There are significant differences in results obtained by applying the same method to data from different labs, and also from applying different methods to data from the same lab. Two-factor ANOVA confirms that **for dataset (2) variability is much higher between labs than between calculation methods**. **This result further strengthens the need for a PROTOCOL** (see the Document "Interaction with DNA - GUIDELINES), so to use for instance the same concentrations, the same procedural details, and the same number of points, to obtain datasets which are strongly robust and appropriate **for an inter-laboratory exercise**.

5.3 Dataset (3) The second round of comparison of EB/CT-DNA titrations was done on experiments of dataset (3). Data was collected by Pisa, Szeged, Lisbon and Palermo following a detailed protocol (see the Document "Interaction with DNA - GUIDELINES). Results from the direct titration (EB in the cuvette, CT-DNA added) are summarized in Table 3 and some histograms emphasizing them are shown in Figures 4 and 5. It is clear that, now, for dataset (3) the highest variability arises from the calculation methods. Figure 6 shows that the variability decreased significantly in dataset (3) when compared to dataset (2).

Table 3 – Values of logK obtained in the 2nd round of inter-comparison. Calculations from the direct titration (EB with DNA) only.

	FITTED THE DATA →													
DID EXP ↓	PALERMO Thordarson	Mg Ghee VH	Thordarson	Supramol	SZEGED HypSpec	KEV	Supramol	Supramol for n = 1	Univ. NOVA Lisboa EXCEL SOLVER	PISA MGH	PISA HypSpec all points	PISA HypSpec - points*	PISA Supramol	PISA k-ev
LISBON	5.2	6.8	4.6	4.7	5.8	5.8	6.0	5.2	5.6	5.2	6.0	6.1	5.4	5.9
PALERMO	5.1	5.4	5.1	5.1	5.7	5.7	5.8	5.2	5.4	5.4	5.8		5.3	5.9
PISA	5.4	4.9	4.8	4.7	6.0	5.8	5.4	4.8	5.7	5.0	5.9	5.8	5.2	5.6
SZEGED	5.8	5.3	4.9	5.0	5.8	5.8	5.7	5.1	5.9	5.0	5.7	5.9	5.5	5.8

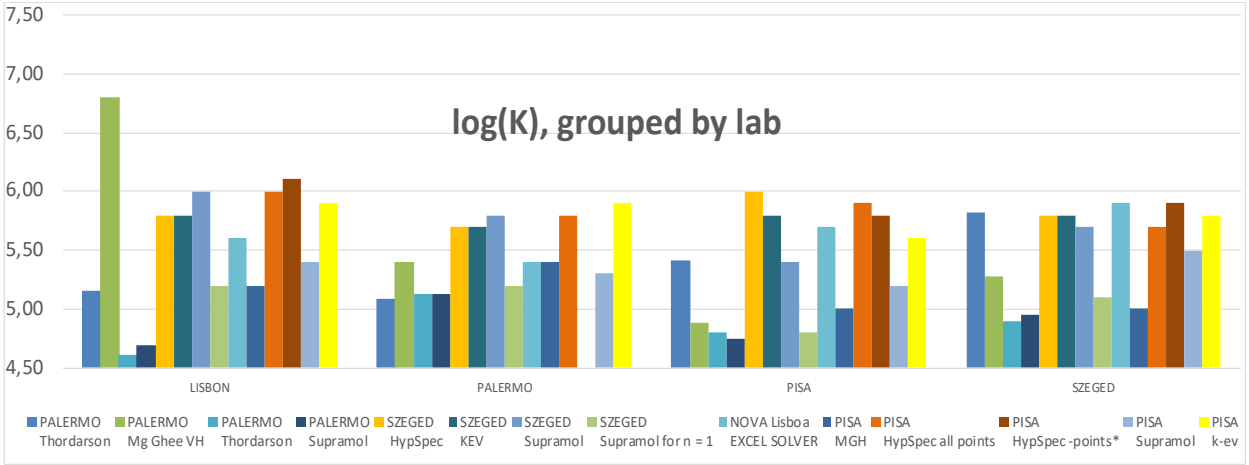

Figure 4 – Values of logK obtained in the 2nd round of inter-comparison (dataset (3)) grouped by laboratory

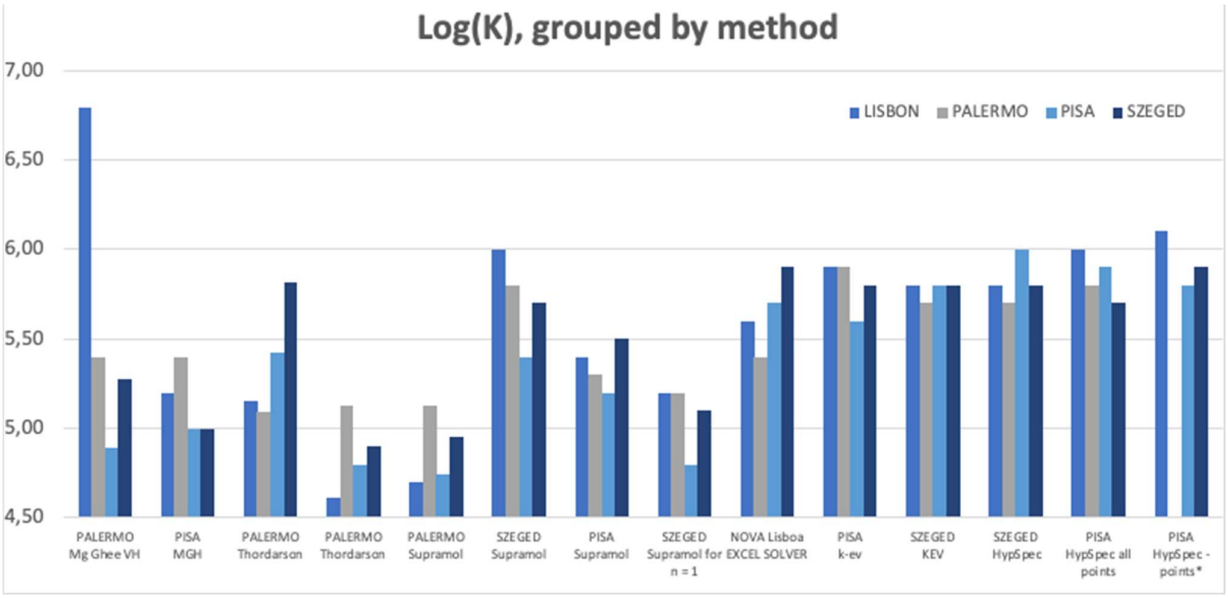

Figure 5 – Values of log(K) obtained in the 2nd round of inter-comparison (dataset (3)) grouped by the method.

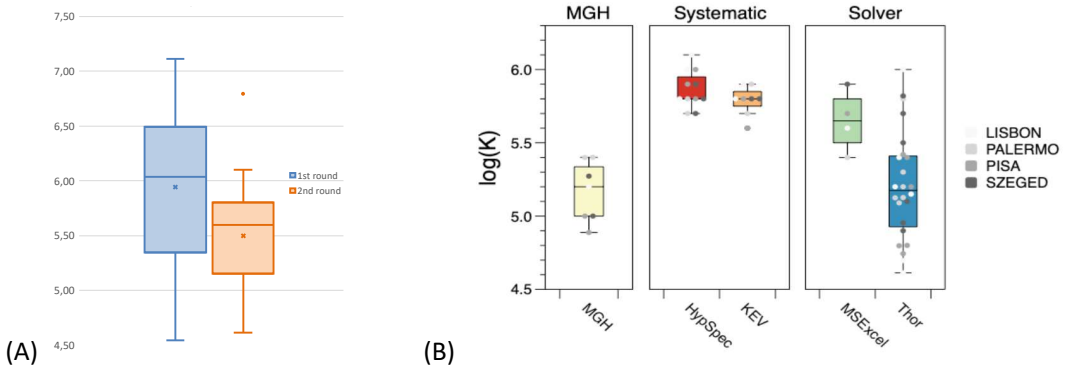

Figure 6 – Box-and-whisker plots of (A) logK values obtained in the 1st round (dataset 2) and 2nd round (dataset 3) of inter-comparison; (B) data of 2nd round (dataset 3) of inter-comparison over different methods.

One-factor Anova using all points from dataset (3) also confirms that there are no significant differences when comparing results from different labs. Comparing different calculation methods, on the other hand, still gives different results. P-values show no significant differences between laboratory datasets ($p = 0.58$) but significant between calculation methods ($p = 2x10^{-6}$). The contribution to overall variability (measured as the standard deviation of log units) from experimental data sets is 0.11, from calculation methods is 0.53 and the residual error is 0.09.

6 – CONCLUSIONS

It can be concluded that the variability between labs dropped from 22% to 6% after the adoption of a common procedure. On the other hand, the choice of calculation method significantly affects both bias and precision. Systematic approaches (HypSpec/K-ev) seem, at this stage, to be better for reproducibility and narrower data distribution. On the other hand, the accuracy of McGhee and von Hippel/Excel solver/Thordarson approaches may be higher. In fact, according to the conditions: BUFFER = 0.1 M KCl, 0.01M Hepes, pH 7.4, 25.0°C, literature data collected (ANNEX A of Document "Interaction with DNA - GUIDELINES) suggests that logK is close to 5.4 and DNA bp/EB ratio is 2.5.

WG2-TG2 is currently working on two other (hopefully simpler) systems (calixarene+fluorophore 1:1 binding, bovine serum albumin+ibuprofen 1:1 binding) to verify what has been suggested here by the analysis of the EB/CT-DNA system and better enlighten the strengths-weaknesses of the different data fit procedures.

ANNEX A

I. Determination of the equilibrium constant for complex formation when the ligand is a polynucleotide

The reaction between a polymer site, S, and a dye, D, to form the DS complex can be expressed by the relationship:

$$D + S \leftrightarrows DS \tag{I.1}$$

whose equilibrium constant is:

$$K = \frac{[DS]}{[D][S]} \tag{I.2}$$

where [DS], [D] and [S] represent the equilibrium concentrations of DS, D and S respectively, with the site concentration, [S], expressed in *base pairs*.

The following mass balance also applies:

$$C_D = [D] + [DS] \tag{I.3}$$

where C_D is the total concentration of dye.

If the Lambert & Beer law applies, for a wavelength where only the free and bound dye absorb and for a 1 cm path length, the overall absorbance is given by the equation:

$$Abs = \varepsilon_D [D] + \varepsilon_{DS} [DS] \tag{I.4}$$

Substitution of equation (I.3) into (I.2) yields:

$$\frac{1}{K} = \frac{(C_D - [DS])[S]}{[DS]} \tag{I.5}$$

that can be rewritten as:

$$\frac{C_D}{[DS]} = 1 + \frac{1}{K[S]} \tag{I.6}$$

Substitution of (I.3) into (I.4) yields:

$$Abs - \varepsilon_D C_D = (\varepsilon_{DS} - \varepsilon_D) [DS] \tag{I.7}$$

If we now define:

$$\Delta Abs = Abs - \varepsilon_D C_D \tag{I.8}$$

$$\Delta\varepsilon = \varepsilon_{DS} - \varepsilon_D \tag{I.9}$$

then, (I.7) becomes:

$$[DS] = \frac{\Delta Abs}{\Delta\varepsilon} \tag{I.10}$$

Upon introducing (I.10) into (I.6) and rearranging, one obtains:

$$\frac{C_D}{\Delta Abs} = \frac{1}{\Delta\varepsilon} + \frac{1}{\Delta\varepsilon K}\frac{1}{[S]} \tag{I.11}$$

For $C_P >> C_D > [DS]$ it turns out that $[S] \cong C_P$. Then, eq. I.11 becomes:

$$\frac{C_D}{\Delta Abs} = \frac{1}{\Delta\varepsilon} + \frac{1}{\Delta\varepsilon K}\frac{1}{C_P} \tag{I.12}$$

known as the Hildebrand and Benesi equation.

A plot of $C_D/\Delta Abs$ *vs.* $1/C_P$ is a straight line whose slope and intercept are equal to $1/\Delta\varepsilon K$ and $1/\Delta\varepsilon$ respectively. Therefore, K is obtained as the intercept/slope ratio, whereas $\Delta\varepsilon$ is the intercept reciprocal.

Under conditions of polymer excess ($C_P > 10C_D$), the mass conservation for the polymer sites holds ordinarily, i.e.:

$$C_P = [S] + [DS] \tag{I.13}$$

where C_P is the total polymer concentration expressed in the molarity of base pairs.

If eq. I.13 holds, then an alternative equation could be derived as follows: from (I.2), (I.3) and (I.13) one obtains:

$$K = \frac{[DS]}{(C_P - [DS])(C_D - [DS])} \tag{I.14}$$

that, taking into account relationship (I.10) can be rewritten as:

$$\left(\frac{C_P C_D}{\Delta Abs} + \frac{\Delta Abs}{\Delta\varepsilon^2}\right) = \frac{1}{K\Delta\varepsilon} + \frac{C_P + C_D}{\Delta\varepsilon} \tag{I.15}$$

Such an equation enables K and $\Delta\varepsilon$ to be obtained iteratively. That is, disregarding the $\Delta Abs/\Delta\varepsilon^2$ term on a first approximation, $\Delta\varepsilon$ can be calculated from the reciprocal of the slope of the straight line fitting the experimental $C_P C_D/\Delta Abs$ *vs.* $(C_P + C_D)$. This $\Delta\varepsilon$ value will be used to re-evaluate the $C_P C_D/\Delta Abs + \Delta Abs/\Delta\varepsilon^2$ term and so on until convergence is reached.

The Hildebrand and Benesi treatment leading to eqs. (I.12) and (I.15) can be successfully applied to ordinary equilibria, for instance, complex formation between a ligand and a metal ion.

For more complex cases, concerning macromolecules such as nucleic acids, the free site concentration is a function of the saturation degree and therefore eq. (I.13) no longer applies.

Therefore, if eqs. (I.12) and/or (I.15) enable a first rough estimate of the K and $\Delta\varepsilon$ parameters to be obtained, eq. (I.11) only has to be used for correct data treatment.

Hence, a relationship between the free site concentration [S] and the experimental data is needed. Such an equation was found by Mc Ghee and Von Hippel (J. D. McGhee and P. H. von Hippel, 1974) that, based on a statistical approach, defined the f(*r*) function (eq. I.16). This function takes into account both the saturation degree of the polynucleotide $r = [DS]/C_P$ and the site size n, at complete saturation of the polynucleotide:

$$f(r) = \frac{[1 - nr]^n}{[1 - (n-1)r]^{n-1}} = \frac{[S]}{C_P} \tag{I.16}$$

The value of *n* can be obtained both by low ionic strength titrations, where the complex formation is quantitative or by using the Scatchard analysis of the data (see II). The variable *r* is directly evaluated from the experimental data as:

$$r = \frac{[DS]}{C_P} = \frac{\Delta Abs}{(\Delta\varepsilon Cp)} \tag{I.17}$$

Form (I.16) it turns out that

$$[S] = C_P f(r) \tag{I.18}$$

and therefore eq. (I.11) becomes

$$\frac{C_D}{\Delta Abs} = \frac{1}{\Delta\varepsilon} + \frac{1}{\Delta\varepsilon K}\frac{1}{C_P f(r)} \tag{I.19}$$

As already said, a first rough estimate of $\Delta\varepsilon$ can be obtained using eqs. (I.12) or (I.15), or from the amplitude of the titration curve. This first estimate will enable us to obtain approximate values of *r* and f(*r*). These values will be used for fitting the data according to eq. (I.19), which will provide a better value for $\Delta\varepsilon$, that will be used to re-estimate *r* and f(*r*). The procedure is repeated until convergence is reached (three iterations are usually sufficient). The equilibrium constant K is equal to the ratio intercept/slope of the trendline obtained at the end of the treatment.

For fluorescence data the same procedure can be applied, just replacing ΔAbs and $\Delta\varepsilon$ by the analogous $\Delta F = F - F_0$ and $\Delta\varphi = \varphi_{DS} - \varphi_D$ parameters, respectively.

II. Determination of the equilibrium constant for complex formation and of the site size via the Scatchard equation

Let the interaction between a dye molecule D and a free site on the polymer be expressed by the equation

$$D + S \leftrightarrows DS \tag{II.1}$$

whose equilibrium constant is:

$$K_{SC} = \frac{[DS]}{[D][S]}$$

(II.2)

If C_P is the polynucleotide concentration in base pairs and B is the number of binding sites for every base pair, the total site concentration is:

$$[S]_0 = BC_P \quad \text{(II.3)}$$

Following the Scatchard hypothesis, the sites, independent to each other, are saturated in such a way that the mass conservation equation applies in its classical form, that is:

$$[S]_0 = [DS] + [S] \quad \text{(II.4)}$$

From (II.3) and (II.4) we get:

$$[DS] + [S] = BC_P \quad \text{(II.5)}$$

Introducing now the r parameter, already defined in (I)

$$r = \frac{[DS]}{Cp} \quad \text{(II.6)}$$

eq. (II.5) becomes:

$$rC_P + [S] = BC_P \quad \text{(II.7)}$$

and, therefore:

$$[S] = C_P (B-r) \quad \text{(II.8)}$$

Finally, from (II.6), (II.8) and (II.2) we get:

$$\frac{r}{[D]} = K_{SC}B - K_{SC}r \quad \text{(II.9)}$$

known as the Scatchard equation (G. Scatchard, 1949).

The plot of *r*/[D] *vs. r*, when linear has the slope equal to $-K_{SC}$ and intercept on the X-axis equal to *B*. According to the Scatchard argument, the reciprocal of *B* yields the site size defined in eq. (II.3).

It can be demonstrated (introduction of eqs. (I.3), (I.13) and (I.17) in eq. (I.2) and of eqs. (I.3), (II.5) and (II.6) in eq. (II.2)) that, for isolate binding of the dye, i.e. r®0, $K_{sc}B = K$, with K the equilibrium constant of the binding defined in Appendix I.

However, the linearity supposed by eq. (II.9) is rarely fully obeyed. This might happen only when a single class of independent sites is present on the polymer that, moreover, has to be saturated in an ordered way, without any space between an occupied site and the following, i.e. when *B* = 1.

When the sites are not independent and cooperativity effects are present, gaps that cannot be occupied are formed and the Scatchard equation can no longer afford a correct model for the equilibrium. The K_{SC} value is no longer a constant, but a function of polynucleotide saturation. More precisely, K_{SC} increases for positive and decreases for negative cooperativity, producing a curved Scatchard plot with opposite concavities (C. R. Cantor and P. R. Schimmel, 1980).

This behaviour was rationalised by McGhee and Von Hippel through rigorous mathematical models that introduce correcting factors into the Scatchard equation, based on cooperativity and probabilistic models (J. D. Mc Ghee and P. H. von Hippel, 1974).

These authors demonstrated that the Scatchard plot should display a positive deviation from linearity at the end of the titration curve, i.e. for high values of *r*. Owing to this phenomenon, due to site overlapping, the intercept on the X-axis is larger than *B* (Fig. II.1). Its value, 1/*n*, is related to *B* through the relationship *n* = (1/*B*+1)/2 (J. D. Mc Ghee and P. H. von Hippel, 1974).

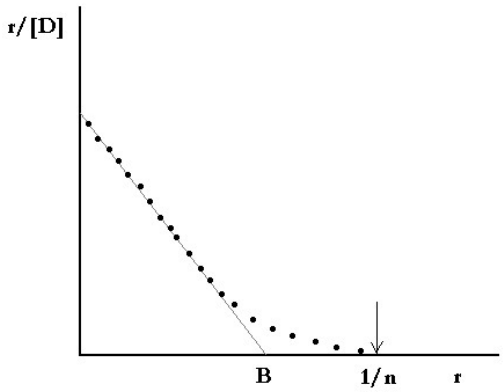

Fig. II.1 Scatchard plot displaying a positive deviation from linearity for high values of *r*.

III. Comments on how to plot data according to the Scatchard plot and then fit them according to the McGhee and von Hippel equation

- **LINEARITY CHECK** In all the equations the assumption made is that fluorescence F is proportional to molar concentration and can be expressed as the sum of the molar concentration of any fluorophore multiplied by its optical parameter. This is not straightforward and needs to be checked.

- **DEFINITION OF r PARAMETER** r is the FRACTION OF SATURATED (bound) polynucleotide in base pairs (or phosphate groups), thus, it is a number between 0 and 1. Accordingly, r x 100 will be the percentage of bound DNA base pairs. This means that r = [bound]/C_{DNA} where C_{DNA} is the total analytical DNA concentration and [bound] is the molar concentration of bound DNA base pairs. This means that no alternative expression of r is possible and this will not depend on the way the titration is done (DNA added or EB added). However, by inspecting the additional comments below, it becomes clear that this data analysis is possible only in case the DNA is added to EB.

- **HOW TO CALCULATE r - PART I** As shown in Appendix I Eq. (I.10), it can be demonstrated that, starting from the analogous fluorescence (instead of absorbance):

  $F = \varphi_{EB}$ [EB] + $\varphi_{EB\text{-}DNA}$ [EB-DNA] (where [EB-DNA] = [bound] cited above)

  It turns out that:

  [EB-DNA] = $\Delta F/\Delta\varphi$

  where $\Delta F = F - \varphi_{EB}C_{EB}$ and $\Delta\varphi = \varphi_{EB\text{-}DNA} - \varphi_{EB}$

  If we know [EB-DNA], we may calculate r = [EB-DNA]/C_{DNA} = $\Delta F/(\Delta\varphi \times C_{DNA})$

  In the above equations, C_{EB} is the EB total analytical concentration at any point of the titration (it changes as it contains dilution effects) and φ_{EB} is known by using the first point of the titration (at zero DNA addition, [EB-DNA] = 0, $F° = \varphi_{EB}C°_{EB}$ and thus obviously $\varphi_{EB} = F°/C°_{EB}$).

  So, we can evaluate ΔF at each point of the titration. The problem is that we do not know $\varphi_{EB\text{-}DNA}$ and thus $\Delta\varphi$ is also not known.

- **HOW TO CALCULATE r - PART II: how to evaluate Δφ** If we plot a binding isotherm as $\Delta F/C_{EB}$ *vs.* C_{DNA} the plateau at "infinite" DNA addition will correspond to $\Delta\varphi$. In fact, under these conditions:
$F = \varphi_{EB}[EB] + \varphi_{EB\text{-}DNA}[EB\text{-}DNA] = \varphi_{EB\text{-}DNA}C_{EB}$

(as there will be no free EB, since all EB will be bound).

Thus, $\Delta F = F - \varphi_{EB}C_{EB} = \varphi_{EB\text{-}DNA}C_{EB} - \varphi_{EB}C_{EB} = (\varphi_{EB\text{-}DNA} - \varphi_{EB})C_{EB}$ and $\Delta F/C_{EB} = \Delta\varphi$

Note that, in this view, [EB-DNA] and r (as a "fraction of bound species") will be related to the ratio between the signal change at some point of the titration (ΔF) and the overall maximum change corresponding to 100% binding ($\Delta\varphi$).

These parameters cannot be obtained if the titration is done the opposite way (EB added), because the titration plot would not correspond to the change from 0% bound to 100% bound form of the fluorophore, of which we are following the signal (EB).

- **COMMENT ON Δφ EVALUATION** This number is critical for any further evaluation/plot. Therefore, either the titration is very good in the sense that the plateau is reached and its value is robustly evaluated from the last values of the experiment OR we need some way to extrapolate the limiting value in the $\Delta F/C_{EB}$ *vs.* C_{DNA} plot. In principle, this limiting value is not a function of site size/stoichiometry, it refers to the signal response of EB under 100% bound conditions, irrespective of site size. Therefore, any simple model considering a 1:1 binding may be used for extrapolation. For instance, the simple Hildebrand and Benesi equation may be useful (Eq. I.15, reported below):

$$\left(\frac{C_P C_D}{\Delta Abs} + \frac{\Delta Abs}{\Delta\varepsilon^2}\right) = \frac{1}{K\Delta\varepsilon} + \frac{C_P + C_D}{\Delta\varepsilon}$$

Here, $C_P = C_{DNA}$, $C_D = C_{EB}$, $\Delta Abs = \Delta F$, $\Delta\varepsilon = \Delta\varphi$. $\Delta\varphi$ is the reciprocal of the straight line obtained by plotting the overall left term *vs.* ($C_{DNA} + C_{EB}$). This is an iterative procedure where the left term is first plotted according to a raw first $\Delta\varphi$ estimate, then we obtain a new $\Delta\varphi'$ from the reciprocal slope, re-use it to get new " y' " values and so on, until convergence is reached. Other equations are possible to be used with the same aim.

- **SCATCHARD PLOT** Given that we have now everything needed, we can prepare the Scatchard plot as r/[EB] *vs.* r where [EB] is the free/unbound EB content. Fortunately, for EB we do not need to know the site size to weight for different contributions and the simple equation below will hold (total = unbound + bound): $C_{EB} = [EB] + [EB\text{-}DNA]$. Thus, the needed [EB] value is $[EB] = C_{EB} - [EB\text{-}DNA] = C_{EB} - \Delta F/\Delta\varphi$

- **SCATCHARD ANALYSIS** The Scatchard equation is given below.

$$\frac{r}{[D]} = K_{SC}B - K_{SC}r$$

Based on the original paper by Scatchard [D], the free fluorophore concentration, is $[EB] = C_{EB} - [EB\text{-}DNA] = C_{EB} - \Delta F/\Delta\varphi$). As explained in II, K_{sc} is the binding constant under the Scatchard model, $K_{sc}B$ should correspond to the K obtained by a simple 1:1 binding, B = 1/n with n the site size, even if a statistical correction of Mc Ghee & von Hippel says that the correct value is n = (1/B+1)/2 (J. D. Mc Ghee and P. H. von Hippel, 1974). However, linearization may not be a good option and, as better explained in (II), this model has many limitations and does not account for EB's possible rearrangement over different DNA sites during titration to enable the maximum saturation possible.

- **McGhee and von Hippel analysis** This analysis starts exactly from the same Scatchard plot. Just, the points are fitted according to an equation which was obtained by McGhee and von Hippel by re-discussing the Scatchard model so that it becomes more suited for DNA polynucleotide (polyelectrolyte) and accounts for re-arrangement (re-distribution) processes which may become crucial on a linear polymer. The crucial point of this model is the f(r) function defined as:

$$\mathrm{f(r)} = \frac{(1-nr)^n}{[1-(n-1)r]^{n-1}}$$

Under this model, the Scatchard equation becomes:

$$\frac{r}{[EB]} = K_{GH} \times \frac{(1-nr)^n}{[1-(n-1)r]^{n-1}}$$

Therefore, we just need to fit the Scatchard plot according to the equation above, to obtain K_{GH} and n (site size).

Note that, even if the equation above is a step forward from the one established by Scatchard, it is again a non-comprehensive model as possible cooperative effects are not taken into account. These cooperative effects will become crucial at the beginning of the titration (high fluorophore excess "fighting" over DNA strand), these are the points at the right part of the Scatchard plots (high r). Also, points at low r will also have problems as the signal change is low at high DNA excess (end of titration). Therefore, the problem of the need to "arbitrarily" exclude some outliers still holds…even though this is less dramatic than in the case of Scatchard analysis (because a straight line is a very limiting trend and the model is even more simplified and unrealistic). Expression for the McGhee and von Hippel equation considering cooperativity are included in doi:10.1016/j.jinorgbio.2006.11.009 and Schwarz's theory for cooperativity can be found in doi:10.1016/j.abb.2006.06.021.

IV. Determination of the equilibrium constant and number of sites for complex formation when the ligand is a polynucleotide

The reaction between a polymer site, S, and a dye, D, to form the DS complex can be expressed by the relationship

$$D + S \leftrightarrows DS \qquad \text{(IV.1)}$$

Considering a polymer (P) with n non-specific sites where D can bind to form a DS, the concentration of sites ([S]) is given by [S] = n[P] and the fraction of occupied sites (θ_b) occupied with dye molecules is given by:

$$\theta_b = \frac{[DS]}{[S]_0} = \frac{[DS]}{n[P]_0} \qquad \text{(IV.2)}$$

Following this reasoning, the fraction of free sites is:

$$\theta_f = 1 - \theta_b = 1 - \frac{[DS]}{n[P]_0} \qquad \text{(IV.3)}$$

and the mass balance for the dye is:

$$[D]_0 = [D] + [DS] \quad \text{(IV.4)}$$

Now, the apparent binding constant for the complexation of dye molecules to the macromolecule binding sites can be defined as:

$$K = \frac{\theta_b}{\theta_f [D]} \quad \text{(IV.5)}$$

Combinining quantions IV.2-IV.4 with IV.5, leads to:

$$K = \frac{\frac{[DS]}{n[P]_0}}{\left(1-\frac{[DS]}{n[P]_0}\right)([D]_0-[DS])} \quad \text{(IV.6)}$$

This equation can be further rearranged to a second order polynomial form:

$$[DS]^2 - \left(n[P]_0 + [D]_0 + \frac{1}{K}\right)[DS] + n[P]_0[D]_0 = 0 \quad \text{(IV.7)}$$

Which can be solved using the quadratic formula:

$$[DS] = \frac{1}{2}\left(n[P]_0 + [D]_0 + \frac{1}{K}\right) - \sqrt{\left(n[P]_0 + [D]_0 + \frac{1}{K}\right)^2 - 4n[P]_0[D]_0} \quad \text{(IV.8)}$$

As previously stated, the absorvance of the dye in the presence of macromolecule is given by:

$$\text{Abs} = \varepsilon_D[D] + \varepsilon_{DS}[DS] \quad \text{(IV.9)}$$

$$\text{Abs} = \varepsilon_D\,[D]_0 + (\varepsilon_{DS} - \varepsilon_D)\frac{1}{2}\left(n[P]_0 + [D]_0 + \frac{1}{K}\right) - \sqrt{\left(n[P]_0 + [D]_0 + \frac{1}{K}\right)^2 - 4n[P]_0[D]_0} \quad \text{(IV.10)}$$

Equation IV.10 can be fitted to the experimental Abs data to estimate K, n, e_D and e_{DS} . Please note that when the concentration of macromolecule is given in number of base-pairs, n ≤ 1 being the average number of base-pairs per site 1/n.