11-1-2017

# Using competition assays to quantitatively model cooperative binding by transcription factors and other ligands.

Jacob Peacock
*Thomas Jefferson University*

James B Jaynes
*Thomas Jefferson University*

## Recommended Citation

Jacob Peacock, James B. Jaynes, Using competition assays to quantitatively model cooperative binding by transcription factors and other ligands, In Biochimica et Biophysica Acta (BBA) - General Subjects, 2017, ISSN 0304-4165, https://doi.org/10.1016/j.bbagen.2017.07.024. (http://www.sciencedirect.com/science/article/pii/S0304416517302416)

# Using competition assays to quantitatively model cooperative binding by transcription factors and other ligands

**Jacob Peacock** and **James B. Jaynes**
Dept. of Biochemistry and Molecular Biology, Thomas Jefferson University, Philadelphia PA 19107 United States of America

## Abstract

**BACKGROUND—**The affinities of DNA binding proteins for target sites can be used to model the regulation of gene expression. These proteins can bind to DNA cooperatively, strongly impacting their affinity and specificity. However, current methods for measuring cooperativity do not provide the means to accurately predict binding behavior over a wide range of concentrations.

**METHODS—**We use standard computational and mathematical methods, and develop novel methods as described in Results.

**RESULTS—**We explore some complexities of cooperative binding, and develop an improved method for relating *in vitro* measurements to *in vivo* function, based on ternary complex formation. We derive expressions for the equilibria among the various complexes, and explore the limitations of binding experiments that model the system using a single parameter. We describe how to use single-ligand binding and ternary complex formation in tandem to determine parameters that have thermodynamic relevance. We develop an improved method for finding both single-ligand dissociation constants and concentrations simultaneously. We show how the cooperativity factor can be found when only one of the single-protein dissociation constants can be measured.

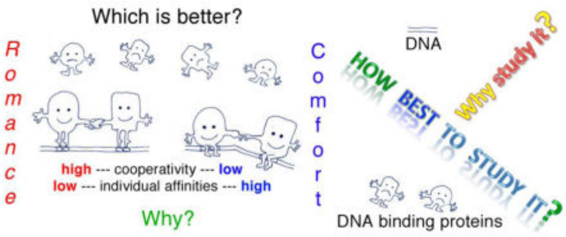
**CONCLUSIONS—**The methods that we develop constitute an optimized approach to accurately model cooperative binding.

**GENERAL SIGNIFICANCE—**The expressions and methods we develop for modeling and analyzing DNA binding and cooperativity are applicable to most cases where multiple ligands bind to distinct sites on a common substrate. The parameters determined using these methods can be fed into models of higher-order cooperativity to increase their predictive power.

## Graphical Abstract

contact: james.jaynes@jefferson.edu (JBJ).

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



**Keywords**

cooperative DNA binding; competition EMSA; modeling ligand-substrate interactions; Hill plots; Engrailed; Extradenticle-Homothorax; curve fitting; finding dissociation constants; quantifying cooperativity

## 1. INTRODUCTION

Cooperative binding by multiple ligands to a substrate is ubiquitous in biological systems. Methods of detecting and analyzing cooperative binding have been well developed over time at a theoretical level. Cooperative binding occurs when the binding of a first ligand to a substrate increases (or decreases) the complex's affinity for subsequent ligands. The phenomenon was first observed and modeled in the oxygen and hemoglobin system, where the binding of one oxygen to unsaturated hemoglobin increases the affinity for the next oxygen [1]. Hill proposed a one-step cooperative binding model,

$$A+ia \rightarrow Aa_i$$

were *A* is the binding substrate, in this case hemoglobin, *a* is the ligand, oxygen, and *i* is the total number of oxygens that bind cooperatively. Using the equilibrium (association) constant for the reaction, $K_a$, gives rise to the Hill equation for the fractional occupancy, $\theta$:

$$\theta=\frac{[Aa_i]}{[A]+[Aa_i]}=\frac{[a]^i}{K_a+[a]^i}$$

Applying the logarithm gives a linearized form, and allows determination of the Hill number, *i*, from measured [*a*] and $\theta$. While computationally and experimentally accessible, the Hill model has numerous pitfalls, and a Hill plot can obscure important cooperative properties of a system (e.g., see Fig. 1 and Fig. 1C,D in Peacock and Jaynes [2]).

The shortcomings of the Hill model in describing the hemoglobin and other systems motivated a slew of subsequent models [3], including Adair-Klotz, KNF, and MWC models [4,5]. Generally, any binding model for a given system may be expressed as a binding polynomial or partition function [3,6,7]. Any specific binding model can be related to the terms and parameters of the binding polynomial. Some authors have introduced theoretically and pedagogically useful formalism into the binding polynomial, which aid in relating the general parameters to specific binding models [8–11].

In recent decades, the importance of cooperativity in eukaryotic transcription regulation has been revealed [12–21]. Cooperativity among transcription factors and cofactors is crucial for achieving nucleic acid binding specificity *in vivo*, allowing a relatively small group of transcription factors to combinatorially regulate large genomes (see e.g. [12]). Using *in vitro* approaches such as gel shifts (EMSA, or electromobility shift analysis) of labeled oligonucleotides (oligos) by purified binding proteins and surface plasmon resonance (SPR), cooperativity has been revealed and quantified in a few cases, serving as conceptual models [22–29]. While SPR can provide more detailed interaction information than EMSAs, it is limited to quantitative analysis of cooperativity for multiple binding sites of a single protein (although it can provide qualitative information on heteromultimeric binding) [30–32]. However, these practical approaches have not been integrated fully into the theoretical framework used for other substrate-ligand interactions. There are many challenges to elucidating cooperative interactions among transcription factors. Conditions in nuclei are difficult to measure and reproduce. All the protein domains that affect function must be included, sometimes making protein purification difficult. Nucleotide sequence flanking the core binding motifs that are sufficient to capture any effects that they may have on binding site shape should also be included [33–35]. Additional complexities are involved when relating *in vitro* binding data to transcriptional readouts. For example, cooperative interactions involving chromatin templates, which may occur through cooperative displacement and modification of nucleosomes, are typically not measured in such studies. Despite the inherent limitations with *in vitro* systems, computational and bioinformatic approaches have had some success in describing cooperativity (see e.g. [24, 36]) using high-throughput methods. Such approaches often continue to rely on previously analyzed, cooperating proteins to entrain the system and provide meaningful quantitative output [37].

Models oriented toward allosteric binding such as MWC can quantify the various conformers of the complex based on the free (unbound) concentrations of each ligand [4,5]. This is useful when measuring free ligand is relatively straightforward, as for the partial pressure of oxygen in the hemoglobin system or the concentration of $Ca^{2+}$ in the calmodulin system. However, the analysis of transcriptional regulation may involve the measurement of one or more cooperative complexes as a function of total added ligand. For this context, where free ligand can be very difficult to measure, quantitative, experimentally accessible models have not been fully developed. Further, saturation binding may require high ligand concentrations, which are attainable for ligands like $O_2$ and $Ca^{2+}$, but can result in aggregation and precipitation of protein ligands. When the binding substrate is DNA or RNA, high ligand concentration may also produce significant off-target (i.e., non-specific) binding. Finally, saturation binding measurements can require the production of large quantities of purified protein, which can be a technical limitation.

Here, we revisit the use of Hill plots to quantify cooperativity, and illustrate key shortcomings as a motivation to develop a more practical and descriptive approach. We develop this approach within a broad theoretical framework applicable to any system involving multiple ligands binding to two or more distinct sites on a substrate, while focusing on the practical application of cooperative binding to DNA by proteins. In that context, we develop methods for quantifying cooperativity by two ligands binding to distinct sites, first measuring individual binding constants and then a cooperativity factor (previously referred to as the cooperativity parameter in, e.g., [22]). This provides the means to predict the binding behavior of cooperating proteins over the full range of concentrations. Motivated by recent identification and analysis of cooperative binding sites for Engrailed with its partner complex Extradenticle/Homothorax (Exd/Hth) [38], we devise a novel method for determining binding constants for individual ligands using competition assays [39], which we show has significant advantages over saturation binding assays. We go on to devise a method to find the cooperativity factor and the second equilibrium constant when only one of the two equilibrium constants can be accurately measured using single-ligand binding. Finally, because the cooperativity factor is thermodynamically meaningful, once it has been determined for pairwise interactions, we show how it can be employed to model more complex systems involving multiple components.

## 2. MATERIALS & METHODS

All computational methods used here broadly follow the same three steps, with different parameters and inputs. First, a binding system is defined with equilibrium constants and total concentrations of each component. Second, the equations relating these equilibrium constants and total concentrations to concentrations of individual species are specified. Third, these equations are manipulated and combined, either manually or using a computer algebra system (Wolfram Mathematica 10.4.0), to relate the desired quantities. Lastly, numerical values are substituted for parameters and the desired outputs computed and/or graphed. The graphs for Figs. 1–3 were produced using the Mathematica notebooks included as supplements. Microsoft Excel and Pacific Tech Graphing Calculator 4.0 were used to generate graphs for Fig. 4. Inkscape and Adobe Photoshop were used to compose figures.

In Fig. 1, the Hill plots are derived for the system shown in Fig. 1A. Using the equilibrium equations (Fig. 1A, lower) and conservation of mass equations, an expression for the total fractional occupancy as a function of free ligand concentration ($[a]$) and the parameters $K_A$, $K_B$ and $n$ was derived (manually in Fig. 1A of Peacock and Jaynes [2]; using Mathematica in Fig_1.nb within Supplemental_Mathematica_notebooks). By definition, a Hill plot shows $\theta/(1 - \theta)$ as a function of $[a]$ on a log-log plot, producing Fig. 1B. The maximum slope was calculated by taking the derivative at the point of 1/2 occupancy (always appropriate for 2 sites, see below).

In Fig. 2, a more complex system is described with two distinct ligands, $a$ and $b$. Again using the relevant equilibrium and conservation equations, an equation describing the concentration of ternary complex, $[AaBb]$, as a function of total ligand concentration $[a]_T$ and parameters was derived (by hand in Fig. 2A of Peacock and Jaynes [2]; using Mathematica in Fig_2.nb within Supplemental_Mathematica_notebooks). Illustrative parameter values were selected to demonstrate the relevant concepts, and each curve calculated and plotted. Points of half-maximum occupancy were found by calculating the limit of $[AaBb]$ as $[a]_T$ approaches positive infinity (see Fig. 2B in Peacock and Jaynes [2]).

In Fig. 3A,B, a competition system is treated, consisting of the Fig. 2 system plus unlabeled competitor, $U_{AB}$, with its own cooperativity and equilibrium parameters. This system is used to describe 3 experiments [38], each with the same labeled DNA oligo (B1a), competing with an unlabeled oligo, either A2a, B1b, or B1a itself. Each experiment uses the same set of equations, but with different parameter values, chosen to approximate a situation encountered experimentally [38]. For Fig. 3A, the equations were solved to find $[AaBb]$ as a function of added competitor $[U_{AB}]_{\mathrm{T}}$. For Fig. 3B, the concentrations of complexes formed on the unlabeled competitor are solved for as a function of $[U_{AB}]_{\mathrm{T}}$. In Fig. 3C, the three DNAs are considered without competitor, with fixed concentrations of both DNA and ligand $b$, and with increasing $[a]_{\mathrm{T}}$. This system is identical to that of Fig. 2, here solved for $[AaB]$, $[ABb]$ and $[AaBb]$ in terms of $[a]_{\mathrm{T}}$. For complete methods, see Fig_3.nb within Supplemental_Mathematica_notebooks.

In Fig. 4A, a simple competition system is considered, consisting of a substrate DNA oligo ($A$) binding a ligand ($a$), in competition with an identical unlabeled DNA, $U_A$. Solving the governing equations of this system, $[Aa]$ is plotted as a function of $[U_A]_{\mathrm{T}}$ for a given set of parameters (see also Fig. 4A in Peacock and Jaynes [2]).

In Fig. 4B, the results of competition and saturation binding experiments were computationally simulated under 7 sets of parameter values ($[A]_{\mathrm{T}}$, $K_A$, and $[a]_{\mathrm{T}}$). To simulate a competition experiment, the system of Fig. 4A was used, and a series of 15 points ($[U_A]_{\mathrm{T}}$, $[Aa]$) were calculated, with simulated experimental noise added to $[Aa]$. For the saturation binding experiment, the simple system of ligand $a$ binding substrate $A$ to form complex $Aa$ was simulated to produce points ($[a]_{\mathrm{T}}$, $[Aa]$). For comparison with competition binding, it was assumed that $[a]_{\mathrm{T}}$ was actually a dilution series from an unknown stock $[a]_{\mathrm{T0}}$. The points generated in each simulated experiment were then fit with least squares regression [40, 41] to the corresponding equations used to generate them, but now with $[a]_{\mathrm{T0}}$ and $K_A$ unknown, and experimental noise in $[Aa]$. This simulation was repeated 100 times to produce a Monte Carlo estimate of the 95th percentile of the absolute percent error of the parameter estimates, which is shown for each of the 7 conditions. For details, see Fig. 4B in Peacock and Jaynes [2], which shows the same system as Fig. 4B, again solved for ternary complex $[AaBb]$ as a function of $[a]_{\mathrm{T}}$, plotted at a variety of parameter values, and showing the 50th percentile as well as the 95th.

In Fig. 4C, the system of Fig. 2 is used to draw graphs for different values of the cooperativity factor $n$. This is done to demonstrate how the expression derived for $[AaBb]$ as a function of total ligand $[a]_{\mathrm{T}}$ can be used to determine $n$, via regression analysis, once $K_A$ and $K_B$ have been determined (using, e.g., the method illustrated in Fig. 4A).

## 3. RESULTS AND DISCUSSION

### 3.1 One ligand binding cooperatively to two sites

In order to develop a practical, systematic approach to the problem of cooperative binding, we begin with the simplest case. A straightforward model to describe the binding of a single ligand to two sites is illustrated in Fig. 1A. A species with two distinct ligand binding sites ($AB$) can form a ternary complex ($AaBa$) with free ligand ($a$) by first forming either of two single-ligand intermediates, *AaB* or *ABa*. Adding a second ligand then converts either of these to the ternary complex. Assigning standard equilibrium binding constants (dissociation constants, or Kd's) to each of these reactions gives the relationships shown in Fig. 1A below the flow chart. While there are 4 distinct Kd's, one of them is not independent, but is completely determined if the other 3 are fixed. Reflecting this, we use only 3 variables to describe these Kd's, noting that the two Kd's governing the occupancy of site *A* are related to each other by the same factor that relates the two Kd's that govern the occupancy of site *B*. We call this factor *n*, the cooperativity factor. It represents the fold decrease in Kd that results from prior occupancy of the other site, or, equivalently, the fold increase in affinity for a second ligand caused by binding of the first ligand. Thermodynamically, it is related to the change in free energy of association (or dissociation) of one ligand that results from binding by the other. A cooperativity factor greater than 1 indicates positive cooperativity, where bound ligand increases the affinity for additional ligands, while a cooperativity factor less than one means that bound ligand reduces the affinity for additional ligands. Using these relationships, for any specified Kd's $K_A$ and $K_B$, cooperativity factor *n*, and total concentration of *AB*, we can calculate the occupancy of each site, and the overall fractional occupancy, $\theta$, at any concentration of ligand. The fractional occupancy as a function of free ligand concentration (both on a log scale) can then be displayed as a Hill plot [1, 42, 43], to see how such plots vary for different sets of constants.

Fig. 1B shows several interesting cases that illustrate some of the difficulties with using such plots for measuring cooperativity experimentally. The blue curves show two cases that conform to the situation of equivalent sites ($K_A = K_B$, or $K_A/K_B = 1$), for which the Hill formalism was originally derived, while the solid blue curve shows the classical case of positive cooperativity. The maximum slope (the purple line is tangent to the curve at the point of maximum slope) is greater than that without cooperativity, where the plot is a straight line of slope 1 (not shown). As is well known, this maximum slope approaches the number of cooperating sites (in this case 2) as the cooperativity becomes very large. When the cooperativity is negative ($n < 1$; $n = 0.04$ for the dashed blue curve), the plot shows the opposite shape, having a slope < 1 at its minimum point. (This maximum or minimum slope always occurs at 50% occupancy for two sites, discussed further below.) These effects depend strongly on the assumption of equivalent sites, a condition that is not often achieved with biological molecules, particularly for protein binding sites on DNA. When the two sites have different Kd's and no cooperativity ($n = 1$), the Hill plot shows the same behavior as with equivalent sites and negative cooperativity, illustrated for the case where the Kd's differ by a factor of 100 (Fig. 1B, dashed red curve).

Strikingly, positive cooperativity can be completely masked for non-equivalent sites [7], resulting in a straight line (solid red), which is indistinguishable from equivalent sites with no cooperativity (not shown). This, of course, occurs only for particular values of relative Kd and cooperativity, but in all cases of non-equivalent sites, apparent cooperativity is reduced. If the ratio of Kd's is far enough from 1, it will look like negative cooperativity rather than positive cooperativity (not shown). The precise conditions for these effects are given in the legend of Fig. 1B, and derived in Fig. 1A of Peacock and Jaynes [2]. Clearly, then, in cases where binding sites may not be equivalent, Hill plots are highly unreliable indicators of cooperativity [44, 45].

### 3.2 One ligand binding cooperatively to two or more equivalent sites

The Hill equation assumes an implausibly high-order reaction mechanism, equivalent to the simultaneous binding of multiple ligands. However, in the special case of equivalent sites and very high cooperativity, the Hill formalism can serve as a good approximation, especially if the cooperativity occurs in a single step. Fig. 1C,D in Peacock and Jaynes [2] provides an illustration of Hill plots for this case, and for the related case of progressive cooperativity, where each additional bound ligand changes the affinity for subsequent ligands in equal increments.

To summarize the lessons that can be gleaned from these examples (see Fig. 1 in Peacock and Jaynes [2] for a full description), for a Hill plot to reveal either the number of cooperating sites or the degree of cooperativity, accurate binding data must be obtained for a wide concentration range in order to determine the point of maximum slope. This is because the maximum slope may not occur at 50% occupancy, and slopes at either extreme of concentration approach 1. Perhaps most unrealistically, these approaches are only effective for equivalent ligand binding sites, which is rarely expected for natural DNA binding sites and ligands.

### 3.3 Two proteins binding cooperatively to two sites: an alternative to measuring co-complex formation as a function of one protein concentration

If the Hill formalism is not appropriate for most realistic situations involving cooperative binding, what is the solution? The one clear advantage of the Hill equation is its simplicity, and this is of course also the source of its limitations. To better model the complexities of real life, it is useful to first study a relatively simple case in some detail, and then use the results to build up to more complex situations. We therefore consider the case of two ligands that cooperate on two distinct binding sites (here and in Sections 3.4 and 3.5). We will use the example of DNA binding proteins cooperating on nearby DNA sites as our model system, but the methods we describe are general. These methods apply to any case where two different ligands bind two distinct sites on a receptor or substrate, with cooperative interactions affecting the relevant binding constants.

This situation has been studied in some detail in a variety of contexts, and it is useful to consider commonly used methods and their limitations. One approach models the system as a single ligand binding to its site, by measuring an apparent Kd of one protein binding in the presence of the other (e.g. [46]). Conceptually, the single "site" is represented by the combination of binding sites and the ligand with constant concentration. Experimentally, we choose a fixed concentration of double-stranded DNA oligonucleotide (oligo) containing the cooperating pair of binding sites, along with a fixed concentration of one of the proteins, and measure the amounts of ternary complex that form as the concentration of the second protein is varied. Typically, the fixed concentrations are chosen based on preliminary experiments that reveal cooperative binding. When a chosen concentration of each protein alone gives little or no detectable binding to the oligo, while mixing all three components results in a clearly detectable ternary complex, positive cooperativity is indicated. It is often of interest in such cases to compare the affinities of related pairs of proteins for the same DNA sites, or of similar (e.g., mutated) sites for the same pair of proteins. We consider here and in Section

3.4 the methods typically used for such comparisons, along with their limitations. We use an example from the literature in some detail, to illustrate how a more complete analysis of the binding parameters can reveal important aspects of the underlying mechanism of binding.

First, we consider the option of measuring an apparent Kd as described above, and using it to compare two related sets of proteins or sites. A binding scheme to describe the situation is shown in Fig. 2A. It is very similar to that considered in Fig. 1A for a single ligand binding cooperatively to two different sites, but here there are two ligands, each binding to only one of the sites. As before, we assume that the binding can occur to each site independently (binding is not ordered), which implies that the ternary complex can dissociate in either of two ways. Also as before, the binding can be described fully using three parameters: *n*, along with the two binary complex Kd's, $K_A$ and $K_B$. In principle, $K_A$ and $K_B$ can be measured independently and directly, assuming that non-specific binding (including binding of each protein to the other site) occurs only with orders of magnitude lower affinity, which is often found to be the case. As we demonstrate later, once the individual Kd's are known, *n* can be readily determined, at least in principle. In practice, there are interesting cases where one of the individual Kd's is so high (i.e., the affinity is so low) as to be difficult to measure directly. In Section 5 of Peacock and Jaynes [2], it is shown how to leverage an individual measurement of only one of the Kd's, along with the results of cooperative binding, to determine all three parameters.

Strikingly, in the situation described above where total concentration of protein *a* ($[a]_T$) is varied, and total concentrations of both oligo ($[AB]_T$) and protein *b* ($[b]_T$) are fixed, a wide variety of binding behaviors can result. When comparing the concentration of ternary complex ($[AaBb]$) formed when using either of two sets of parameters, we find that $[AaBb]$ for one set of parameters may be lower than for the other set at low $[\mathrm{a}]_T$, but higher at high $[\mathrm{a}]_T$. In other words, the curves of $[AaBb]$ vs. $[\mathrm{a}]_T$ may cross (Fig. 2B). This can occur in either of two types of comparisons: comparing two different pairs of sites bound by the same two cooperating proteins, or comparing two sets of proteins binding to the same pair of binding sites. Importantly, in either case, if data from such an experiment are used to determine an apparent Kd for each binding curve, the expectation would be that where the apparent Kd is lower, $[AaBb]$ would be higher at all values of $[a]_T$. While this is, of course, true for a single ligand binding to a single site, this expectation fails with multiple ligands.

This is illustrated in Fig. 2B for the special case where $K_B$ is the same for all the curves. This would apply, for example, to comparisons of cooperative binding to an oligo by different members of a DNA-binding protein family, in combination with a common DNA-binding protein partner. The green curve gives the lowest apparent Kd, but it has the lowest $[AaBb]$ at high values of $[a]_T$. The red curve gives the highest apparent Kd, but it has the highest $[AaBb]$ at high values of $[a]_T$. And the blue curve gives a lower apparent Kd than does the red curve, yet it has a lower $[AaBb]$ at all values of $[a]_T$. The general reason for these "unexpected" outcomes is that the actual 3-parameter system gives more complex curves than can be modeled using a single parameter. The specific reason is that the concentration at which *AB* is saturated with ligand, which occurs at very high $[a]_T$, can be very different for the different curves. For curves where $K_B$ is the same, as in Fig. 2B, the curve with the highest cooperativity factor has the highest saturation concentration of *AB*, regardless of $K_A$.

For example, the black and blue curves have different Kd's but the same cooperativity factor, and therefore saturate at the same level ([*AaBb*] = 0.75; see Fig. 2B in Peacock and Jaynes [2] for the general formula). This is because at very high [*a*], $K_A$ becomes irrelevant. Given that both $[AB]_T$ and $[b]_T$ are constant, the amount of *b* incorporated into *AaBb*, in equilibrium with *b* and *AaB*, at very high [*a*] depends only on the Kd for dissociation of *b* from *AaBb*, namely $K_B/n$.

From these examples, we can see that neither ranking the amounts of ternary complex formed at any one $[a]_T$ nor measuring an apparent Kd in this type of experiment is predictive of relative complex formation overall. (The criteria used to define apparent Kd's are given in Fig. 2B of Peacock and Jaynes [2].) Curves like these will cross each other under conditions that we can glean from the governing equations (given in Fig. 2A in Peacock and Jaynes [2] and described in Fig. 2C of Peacock and Jaynes [2]). Fig. 2C in Peacock and Jaynes [2] also gives an expression for the apparent relative affinity measured in SELEX-seq experiments of the type described in Riley et al., 2014 [47], in terms of the Kd's and cooperativity factors of two cooperating ligands on two composite binding sites.

Another important issue when using this assay is illustrated in Fig. 2C. Two pairs of curves are shown, representing two different concentrations of the "fixed" ligand. As seen by comparing the two pairs of curves, one binding site forms more ternary complex at one fixed concentration, while the other does so at the other fixed concentration, throughout most of the concentration range of the varying ligand. That is, the relative binding behavior reverses in the two cases. The red curve represents a more cooperative site, and at higher concentrations beyond the range shown, it forms more ternary complex than does the less cooperative site, regardless of which concentration of fixed ligand is used.

Here, it is important to note that the formulae and concepts developed throughout this work depend only on relative, not absolute, concentration units. Therefore, we have not specified units for either concentrations or Kd's, except for the specific example developed in Section 3.4 below. Furthermore, cooperativity factors are inherently unitless. Although the absolute nuclear concentrations of most transcription factors have not been established, they are generally thought to be in the range of nanomolar to micromolar [48]. This is also thought to be the concentration range for their functional set of binding sites, and for their individual Kd's in binding to those sites. Therefore, it would be appropriate in applications involving these molecules to consider our values for concentrations and Kd's to have units in the picomolar to nanomolar (nM) range.

Cooperativity for transcription factors has been quantified in very few cases. For λ phage *cro* and *c*I repressors, the measured interaction free energies correspond to cooperativity factors of <200 [49, 50]. (Each change in Kd of 10-fold corresponds to a change in interaction free energy of about 1.37 kcal/mol). Enhanceosomes in mammalian systems involve a number of cooperatively binding proteins [51–53]. Although in several of these cases, synergistic increases in site occupancy were attributed to cooperative binding, no quantitation of cooperativity was reported. However, interactions between transcription factors whose nuclear concentrations are below about 100 nM can result in cooperativity factors up to about $10^6$ without leading to much dimerization in solution. Thus, the cooperativity factors that we have used here are all physiologically plausible.

Clearly, in order to model the system accurately, we need to measure more than one parameter. In Section 3.5, we consider the commonly used method to measure individual Kd's and compare it with a novel method, then go on to show how these can then be used to accurately determine the cooperativity factor *n*. Once these three parameters are known, we can predict the binding behavior, and specifically the relative amounts of each complex, at all concentrations of the components.

Before doing this, we consider another method commonly used to compare the affinities of different combinations of proteins and binding sites, competition experiments, where an unlabeled oligo competes with a labeled oligo for binding by fixed amounts of the proteins [54]. We will consider an example of this experiment in some detail, both to illustrate its limitations and to resolve an apparent paradox, leading to new biological insights.

### 3.4 A case study: insights from an analysis of competition assays

In a paper published in 2012 [38], Fujioka et al. characterized several cooperative binding sites for the Engrailed (En) protein within the *sloppy-paired* (*slp*) locus of Drosophila. En shows strongly cooperative binding to each of these sites with a cofactor complex, which contains one molecule each of the proteins Exd and Hth. This cofactor complex forms in solution and can be co-purified as a single complex when the proteins are co-expressed in bacteria. This stable complex has therefore been treated in binding studies as a single entity, Exd/Hth [35,37,46,54–57]. We compared the apparent affinities of four binding sites from *slp*, and found that two of them showed strong binding by En-Exd/Hth, while the other two were much weaker. These assessments were based on a combination of both direct binding assays and competition binding experiments like those described above. In all cases, ternary complex formation was monitored using gel shift analysis. Note that we refer to the complex containing En, Exd/Hth, and an oligo consisting of the cooperatively bound (composite) site as a ternary complex, because we model Exd/Hth as a single entity for the purposes of studying its cooperativity with En. Binding by the individual proteins was found to be relatively low, and in some cases was undetectable, consistent with a high degree of cooperativity in the binding. However, an apparent paradox was uncovered in that the two lower affinity sites showed distinctly different functional characteristics *in vivo*, despite very similar behaviors in the binding assays. The functional assay involved repression of a reporter transgene in En-expressing cells in the developing Drosophila embryo, which is dependent on a functional binding site for En and Exd/Hth [38]. We speculated that the subtle differences we observed in their apparent cooperativity might be responsible for their distinct functional potencies: the more cooperative site gave more complete repression *in vivo*. Despite this difference, which emerged from a limited set of direct binding assays, the competition assays, considered a good way to quantify relative affinities, showed no difference between them.

In modeling studies represented in Fig. 3, we revisit this issue by using the results from that work [38] to estimate individual Kd's and cooperativity factors for these sites. We model the binding using these parameters, and make several noteworthy discoveries that may explain both the difference in function of the low-affinity sites and why the competition assays did not reveal a distinction between them. These results have important implications for the limitations of these assays, and provide guidelines for their effective use. We note that these results are based on published studies, and it is not our purpose here to establish new biological principles. Rather, we illustrate how quantifying both Kd's and cooperativity factors allow exploration of binding behavior that is outside the concentration range used in the *in vitro* experiments themselves. This in turn can lead to new biological insights, including novel hypotheses that can be tested in subsequent studies. In particular, our analysis is not meant to either test or validate the assumption that Exd and Hth function strictly as a single unit when cooperating with En.

The competition binding assays yielded "competition curves" for each of the sites [38]. In this assay, as mentioned above, a labeled oligo containing the highest affinity site is bound by unlabeled proteins, with fixed concentrations. Increasing amounts of an unlabeled competitor oligo are added, which carries either the same sites as the labeled oligo or different sites, and the decrease in binding to the labeled oligo is quantified. Theoretical curves that closely match the published curves are presented in Fig. 3A for three of the sites. For simplicity, we use only one of the two high-affinity sites, which behaved similarly both in the binding studies and *in vivo*. To produce these curves, we used the limited set of direct binding studies available to estimate a range for the individual Kd's and *n* for each site. We then refined these estimates to produce competition curves that match the data well. Although we do not consider these refined estimates to be precise, they nonetheless suggest a novel hypothesis as to why one of the sites functions better *in vivo*. More generally, the modeling results illustrate why the assays can be misleading, and provide guidance for their effective interpretation and use.

The main conclusions from these studies are 1) the two lower affinity sites have distinct binding behaviors that may explain their functional differences *in vivo*, 2) the competition binding studies did not reveal these differences because of the particular range of concentrations used, and 3) at those concentrations, the two lower affinity sites competed very similarly, but for different reasons. We now describe the results in some detail, to justify and fill out these conclusions. We then describe the lessons learned, one of which applies specifically to the interpretation of competition assays, while another reinforces the lesson from prior sections that in the face of the complexities of cooperative binding, even to two sites, it is necessary to measure multiple parameters in order to model the system effectively. This then provides the impetus to explore novel ways of measuring those parameters, presented in Section 3.5.

Fig. 3A shows that unlabeled oligos carrying the two low-affinity sites compete very similarly (red and blue) for binding to a labeled oligo carrying the high-affinity site. Of course, the high-affinity site itself competes much more effectively (black). Fig. 3B shows that the two low-affinity sites compete similarly for very different reasons, in terms of the complexes that they form as their concentrations increase. Although they each initially form more ternary complex (solid blue and red) than single-protein complexes, the oligo with the lower *n* (blue) binds relatively more of the single-protein complexes (dotted and dashed curves). This difference is magnified at higher concentrations, where the less cooperative oligo forms more of each of the single-protein complexes (blue dotted and dashed curves) than it does the ternary complex (solid blue), while the more cooperative oligo continues to form much more ternary complex (solid red). So, while the net result in the competition assay appears the same, this is specific to the choice of concentrations of labeled oligo and proteins. At other concentrations, differences would be more apparent. Rather than illustrating this, we show in Fig. 3C the differences in direct binding by these two oligos, which may explain their differences in function.

Fig. 3C, top, shows direct binding curves of the type in Fig. 2, $[AaBb]$ vs. $[a]_T$ with both $[b]_T$ and $[AB]_T$ constant. Here we see that more ternary complex is formed on the “red” site at all $[a]_T$, compared to the blue curve. The more cooperative low-affinity site, which forms more ternary complex (solid red), is the one that functions better *in vivo*. In fact, its function *in vivo* is more like that of the high-affinity site than it is the other low-affinity site [38]. This may be explained by the fact that the amounts of ternary complex formed by the two higher-functioning sites (solid black and red) become more similar at high $[a]_T$, and more distinctly different from the lower-functioning site (solid blue). That is, the red curve approaches the black curve, and separates from the blue curve, at high $[a]_T$. This behavior is due to the higher cooperativity of the better-functioning site, as explained above for Fig. 2: the saturation value depends more on the cooperativity, while the relative behavior at low $[a]_T$ is more dependent on $K_A$.

Fig. 3C (middle and bottom graphs) illustrates the underlying reason for the “surprisingly strong showing” by the lower-functioning site in the competition assay. At a given set of concentrations, it forms more single-protein complexes (blue dashed and dotted) than does either of the other two sites (red and black). The results of a competition experiment depend on the total amount of each ligand bound by a site, rather than only the amount of ternary complex. Thus, less ternary complex is made up for by the formation of more single-protein complexes. This is again consistent with the lower cooperativity of the lower-functioning site (it forms relatively less ternary complex and more of the single-protein complexes).

In the competition assays, the total amount of unlabeled, competing binding site goes well beyond the range used in direct binding assays for the same site, typically up to hundreds of fold more. At such high concentrations, single protein complexes can dominate, even though very little of them form in direct binding assays. At the concentrations used in these experiments, this was the case. Therefore, only if sites are independently known to have either similar Kd’s for formation of each single-ligand complex, or to have similar cooperativity factors, can we expect competition assays to reveal a simple set of “relative affinities”.

As these examples emphasize, modeling the binding of cooperating ligands using a single parameter can only have predictive power for occupancies of sites over a limited concentration range. This limitation should be taken into account when interpreting experiments based on high-throughput methodologies. An example is given in Fig. 2C of Peacock and Jaynes [2].

### 3.5 A comprehensive method for modeling relative complex formation over a wide range of concentrations

The foregoing argues for measuring individual Kd's and a cooperativity factor in order to model binding by a cooperating pair of ligands. To this end, we first describe a less well-known method that has significant advantages over standard methods for determining an individual Kd. We then show how the cooperativity factor can be determined once both individual Kd's are known. In Section 5 of Peacock and Jaynes [2], this methodology is extended to find both the cooperativity factor and the second Kd when only one of the individual Kd's can be accurately determined using single-ligand binding.

Standard methods have been described for determining individual Kd's that involve simply measuring complex formation as a function of ligand concentration (the "saturation binding" method; see e.g. [41]). However, in cases where cooperativity is high, individual Kd's can be challenging to determine accurately, particularly in cases where either the available amount of ligand or its tendency to aggregate precludes obtaining data at high ligand concentrations. Our alternative method uses competition assays, similar to those illustrated in Fig. 3A, except that only a single ligand is used to determine its individual Kd.

Using competition assays to determine an individual Kd has significant advantages, particularly in the case of DNA binding proteins. Preparation of proteins for binding assays often involves concentrating them in a way that can cause denaturation, and for this and other reasons, the fraction of protein that is active in binding may be difficult to determine. In such cases, directly measuring the amount of protein provides only an upper limit on its effective concentration. Using a competition binding assay, we can straightforwardly determine both the Kd and the active concentration of ligand from the same data set. We note in this context that recently developed high-throughput methods for comparing DNA affinities [58] do not always distinguish between absolute and active concentrations of ligand. Further, it is important that these high-throughput methods include exemplars for entrainment and validation that have been characterized by methods grounded in solution biochemistry, such as the following.

Fig. 4A shows examples of binding curves from modeling this type of competition experiment, in which constant total amounts of labeled oligo ($[A]_T$) and ligand ($[a]_T$) are used in combination with increasing amounts of unlabeled oligo ($[U_A]_T$), and the resulting ligand-substrate complex ($[Aa]$) is quantified. $[Aa]$ decreases as $[U_A]_T$ is increased, while all other quantities are constant. $[A]_T$ is known, and $[a]_T$ and $K_A$ are determined as parameters in a non-linear regression analysis. Values for constants were chosen to illustrate why the approach can yield these two parameters independently. When any family of curves representing different Kd's and ligand concentrations start at the same point (i.e., they give the same amount of complex without competitor oligo), they have significantly different shapes, and so diverge from their common starting point, no matter where that starting point is. The differences between the 3 colors is a 10% change in $[a]_T$. The values of $K_A$ are adjusted for each curve to give the same initial value of $[Aa]$. This difference in shape can be captured during regression analysis to give the two parameters independently. Fig. 4A of Peacock and Jaynes [2] gives a derivation of the expressions used. These expressions can also be used in a high-throughput analysis to determine the relative affinities of related binding sites, where the highest affinity site is labeled, and measurements are made using a panel of unlabeled sites of how well they compete for binding to a fixed amount of protein, as described in Hallikas, et al., 2006 [59]. Fig. 4A of Peacock and Jaynes [2] gives an exact expression for determining the relative affinity in such an experiment, and also a simple approximation for the case where only a small fraction of labeled oligo is bound without competitor (which is different from that given in Hallikas et al. and has the correct limit behavior).

We tested the ability of this method to give precise values for the parameters $K_A$ and $[a]_T$ under different conditions, and compared it to the traditional saturation binding method. First, a few words about the latter method. It is possible, in principle, to use saturation binding to determine both $K_A$ and $[a]_T$ simultaneously. This can be done by varying $[a]_T$ in a systematic way (for example by diluting a stock solution) and measuring the resulting $[Aa]$, without initially knowing the actual values of $[a]_T$. If we let the reference value of $[a]_T$ be $[a]_{T0}$, and each experimental value be $[a]_{T0}/\Delta$ (Δ is the “dilution factor”, which varies for each data point, while $[a]_{T0}$ is a constant to be determined from regression analysis) then the relevant formula is $\Delta=[a]_{T0}/\left\{[Aa]*(\{K_A/([A]_T-[Aa])\}+1)\right\}$. From the data set {(Δ, $[Aa]$)}, and knowing $[A]_T$, we can use regression analysis to simultaneously find the two parameters $K_A$ and $[a]_{T0}$. Although this can in principle work effectively, it requires using a set of experimental conditions that are unknowable before the Kd is determined. The optimal conditions are when $[A]_T \sim 10{*}K_A$ (of course, $K_A$ is initially unknown). Even under these optimal conditions, data that are precise to within about 10% are required to determine $K_A$ within about 50% and $[a]_T$ within about 10% (Fig. 4B).

In contrast, using the competition method described above, it is possible to use data accurate to within ~10% to determine $K_A$ within about 20% and $[a]_T$ within about 7% (Fig. 4B). Importantly, the competition method provides this level of precision as long as the $[A]_T$ used is less than or ~ $K_A$. This contrasts with the saturation binding method, which becomes much less effective at determining $K_A$ when $[A]_T$ is either above or below $10{*}K_A$ by several-fold or more. The flexibility and resolving power of the competition method allows us to define an optimal approach to precisely determining both $K_A$ and $[a]_T$: use the lowest $[A]_T$ that allows precise quantitation of $[Aa]$, and the highest $[a]_T$ available, up to a $[a]_T$ that gives $[Aa] \sim [A]_T/2$ without competitor $U_A$. Data taken under these conditions, and with increasing $U_A$ so that $[Aa]$ is reduced to 1/3 or less of its initial value without $U_A$, will give the most precise value practicable for $K_A$, along with a somewhat more precise value for $[a]_T$.

These conclusions are put into context in Fig. 4B, which shows the results of a Monte Carlo analysis of the two methods. A random error was introduced into data sets, and regression analysis was used to find $K_A$ and $[a]_T$ as parameters from the data. The resulting errors in the parameters are shown, as a function of the chosen value for $[A]_T$. The value of $K_A$ was fixed at 1. This is justified by the fact that it is only the relative values of $[A]_T$, $[a]_T$, and $K_A$, and not their absolute values, which determine the shapes of the curves, and therefore how precisely the parameters can be determined. As Fig. 4B illustrates, when $[A]_T$ exceeds $K_A$ by more than 10-fold ($[A]_T$ >10 in this case), determination of $K_A$ becomes increasingly unreliable (approaching or exceeding 100% error at least 5% of the time) with both methods.

With $[A]_T$ between $K_A$ and 10*$K_A$, both methods provide similarly reliable estimates of both parameters. Importantly, with $[A]_T < K_A$, the competition method becomes more reliable, while the saturation method fails completely. These results support the strategy given in the previous paragraph for finding the two parameters simultaneously to the best possible precision under a wide variety of circumstances. Similar results are obtained with different errors in the input data sets (1% and 5%, Fig. 4B of Peacock and Jaynes [2]). Errors at the 50th percentile in the error distribution (Fig. 4B of Peacock and Jaynes [2]) follow the same qualitative pattern as those at the 95th percentile (Fig. 4B), illustrating that the overall error distribution is similar in all cases as a function of the chosen $[A]_T$.

For both the competition and the saturation binding methods, a more precise value for $[a]_T$ and a less precise value for $K_A$ are obtained when higher values of $[A]_T$ are used above ~10*$K_A$. The precision for $K_A$ achievable within an experiment never exceeds that for $[a]_T$. This is because in both formulae, $K_A$ is divided by ($[A]_T$ –$[Aa]$), whereas $[a]_T$ (or $[a]_{T0}$) is divided by $[Aa]$, which is typically lower when averaged over the data points than is ($[A]_T$ – $[Aa]$). Therefore, a smaller change in $[a]_T$ compensates for a larger change in $K_A$. This causes the bounds placed on $[a]_T$ by the data to be more stringent than those placed on $K_A$.

For curve fitting to determine an individual protein’s Kd and its concentration by the above methods, we can use freely available software (e.g., at “statpages.org/nonlin.html”), along with the expressions given above and in Fig. 4A of Peacock and Jaynes [2] (which also includes their derivations). Fig. 4A of Peacock and Jaynes [2] also includes equations to provide initial guesses for the parameters from one or two data points, which are sometimes needed for the regression algorithm to converge.

It is often useful, from an experimental point of view, to include non-specific DNA in such binding experiments, to minimize the effects of contaminating DNA binding proteins that may come through the purification process. Therefore, we developed an alternative methodology which allows accurate determination of $[a]_T$ and $K_A$, as well as the Kd for non-specific binding, as parameters from curve fitting. The expressions used for this purpose, along with derivations and a description of how to analyze the data, are given in Fig. 4C of Peacock and Jaynes [2]. As an illustration of why this can work, if non-specific DNA is included in the experiment shown in Fig. 4A, the apparent Kd’s change without a significant change in the shapes of a curves, as long as the ratio of the non-specific Kd to each of the specific Kd’s is much greater than 1. So, the set of curves shown there is indistinguishable from that resulting from the inclusion of non-specific DNA with a Kd of 1000 at a concentration of 4000, with the same values for $[A]_T$ (= 2), and $[a]_T$ (= 6, 5.4, and 6.6), but with all of the $K_A$’s reduced by a factor of 5.

Once individual Kd’s have been determined for both cooperating proteins, $n$ for the composite site can be found readily. A simple method is illustrated in Fig. 4C. Here, the same equation used to generate curves in Figs. 3B,C and 4C is used to show how a 10% change in $n$ affects the amount of ternary complex that forms as one ligand increases in concentration, while the total amount of the other is held constant. The same expression can be used in regression analysis to determine $n$ from data of this kind. The upper panel of Fig. 4C shows 3 families of 3 curves each. Each family has a different value for $n$, and within each subfamily (those lying close together), this difference is 10%, either above (blue) or below (red) the middle value (black). The uppermost black curve in this graph uses *n* = 500, and it is the same as the black curve in Fig. 3C, upper panel. The other two subfamilies have *n* reduced by 10 and 100-fold, but use the same Kd's. The relative amount of separation within each subfamily suggests how precisely *n* can be determined from data in such an experiment. The upper panel shows that the higher *n* value will be relatively difficult to determine accurately with the chosen concentrations of oligo and fixed protein. However, if we reduce these values, the curves for *n* = 500 ± 10% are more widely separated, as illustrated in the lower graph of Fig. 4C (both $[AB]_T$ and $[b]_T$ are reduced by 10-fold relative to the upper graph). In this case, it becomes easier to accurately determine *n* = 500, while it may become more difficult to determine the lower *n* values used here.

In order to get the most accurate determination of *n*, $[AB]_T$ and $[b]_T$ should be chosen to allow ternary complex formation to approach 50% saturation, but not exceed it. However, if this is not achievable with a $[AB]_T$ that is high enough to allow accurate quantitation of ternary complex, it may be best to use the lowest $[AB]_T$ (and $[b]_T$, which should be comparable) that does allow accurate quantitation, and also results in less than 50% saturation at the highest achievable $[a]_T$. The latter, of course, may be limited by the amount of available ligand. For regression analysis to determine *n*, we use the expression given in Fig. 4C (and derived in Fig. 2A of Peacock and Jaynes [2]).

The method described above requires that each individual Kd be measurable. However, for some ternary complexes, one of the ligands is not observed to bind alone, even at concentrations much higher than those required for strong, cooperative binding. A well-known example of this is described by Jin et al., 1999 [60], involving cooperative binding by the yeast transcription factors a1 and α2. Binding by α2 alone was seen, and addition of a1 resulted in much greater complex formation, suggestive of highly cooperative binding. However, even at the highest concentrations tested, no binding by a1 alone was observed. In order to extend our method to this type of situation, we devised equations for curve fitting to find both the Kd of the weakly binding ligand and the cooperativity factor from gel shift analyses similar to those presented by Jin et al. [60]. This method is described in Section 5 of Peacock and Jaynes [2].

### 3.6 Generalizing the methods to more than two binding sites and cooperating ligands

How can this methodology help us in understanding the functions of more complex composite binding sites, such as those often found in higher eukaryotic genes? Once the Kd's and pairwise cooperativity factors have been determined for pairs of individual sites that make up the composite site, it is straightforward to model them as an interacting network of pairwise interactions. In a typical "Boltzmann" statistical thermodynamic version of such a model ([24] and references therein), the free energy differences induced by each pairwise interaction are combined to give a relative free energy, and therefore a relative occupancy under any specified conditions, for each possible complex. The Kd's and cooperativity factor as defined above naturally feed into this type of model, because they can be readily associated with free energy differences among the various complexes. Although the behavior of such sites has been modeled without information about the pairwise interaction parameters (e.g., [24]), including those parameters would likely yield more meaningful, mechanistic models with wider predictive power that extends well beyond the range of the experimental data [61].

For DNA binding proteins in particular, it may be common for pairwise interactions to dominate the system, mediated by separable protein-DNA and protein-protein interaction domains. For example, in the cooperative interactions between λ phage *cro* and *c*I repressors, pair-wise interactions between nearest neighbors appear to dominate [49, 50]. In such cases, we can measure the individual protein-DNA Kd's and pairwise cooperativity factors, to fully describe the behavior of complexes involving several DNA binding proteins. In these cases, the free energy of dissociation of each of the ligands from the 3-ligand complex can be accounted for by the dissociation energy of the two 2-ligand complexes which contain that ligand, so the system as a whole is solved by knowing each of the pairwise Kd's and cooperativity factors. We now summarize the relationships among these quantities (which are measurable using the procedures described above) and the dissociation constants and cooperativity factor of a 3-ligand complex. Details are provided in Fig. 6 of Peacock and Jaynes [2].

We now need subscripts for each pairwise cooperativity factor to distinguish which pair of ligands it is associated with, as well as another cooperativity factor associated with the 3-ligand complex. This factor is an independent quantity in the general case where additional free energy (positive or negative) may be associated with the formation of the 3-ligand complex beyond that associated with the formation of each 2-ligand complex.

As derived for a 2-ligand complex containing *a* and *b* in Fig. 2A:

$$n_{AB}=[ABC]*[AaBbC]/([AaBC]*[ABbC])$$

and, by analogy,

$$\begin{gathered}n_{AC}=[ABC]*[AaBCc]/([AaBC]*[ABCc])\\ n_{BC}=[ABC]*[ABbCc]/([ABbC]*[ABCc]).\end{gathered}$$

We can find $n_{ABC}$ to be:

$$\begin{gathered}n_{ABC}=[ABC]^2*[AaBbCc]/([AaBC]*[ABbC]*[ABCc])\\ n_{ABC}=[AaBbCc]*K_A*K_B*K_C/([ABC]*[a]*[b]*[c])\end{gathered}$$

where

$$\begin{gathered}K_A=[ABC]*[a]/[AaBC]\\ K_B=[ABC]*[b]/[ABbC]\\ K_C=[ABC]*[c]/[ABCc].\end{gathered}$$

Now, how is the “new” cooperativity factor for the 3-ligand complex related to those of the 2-ligand complexes? Complete dissociation of the 3-ligand complex involves the sum of 3 free energies. For one possible dissociation route, these are: the free energy change when ligand *a* dissociates, that when ligand *b* dissociates from the 2-ligand complex, and that when ligand *c* dissociates from the single-ligand complex.

Because of the basic relationship between the standard Gibbs free energy and the Kd: $\Delta \mathrm{G}^0 = -\mathrm{R} * \mathrm{T} * \ln(\mathrm{Kd})$, where R is the gas constant and T is the absolute temperature, adding these 3 free energies is equivalent to multiplying together 3 dissociation constants: that governing the dissociation of ligand *a* from the 3-ligand complex, that governing the dissociation of ligand *b* from the 2-ligand complex containing ligands *b* and *c*, and that governing the dissociation of ligand *c* to release the free DNA.

This product is:

$$\begin{aligned}&([ABbCc]*[a]/[AaBbCc])*([ABCc]*[b]/[ABbCc])*([ABC]*[c]/[ABCc])\\&=[ABC]*[a]*[b]*[c]/[AaBbCc]\\&=K_A*K_B*K_C/n_{ABC},\end{aligned}$$

where the last equality comes from the last expression above for *nABC*. So, 1/*nABC* represents the “extra” free energy in the complex that is due to cooperative binding; i.e., this “extra” $\Delta G^0 = R * T * \ln(n_{ABC})$. If there is no cooperativity, $n_{ABC} = 1$, and the total free energy of dissociation is just the sum of those for each ligand individually. From the definitions above, we have, for the Kd that governs the dissociation of ligand *a* from the 3-ligand complex:

$$\begin{aligned}&[ABbCc]*[a]/[AaBbCc]=([ABC]*[a]/[AaBC])*\{[ABC]*[ABbCc]/([ABbC]*[ABCc])\}\\&/\left\{[ABC]^2*[AaBbCc]/([AaBC]*[ABbC]*[ABCc])\right\}\\&=K_A*n_{BC}/n_{ABC},\end{aligned}$$

and similarly for the other dissociation constants from the 3-ligand complex.

Now, what do we expect for *nABC* in the case where the free energies of interaction within the complex consist solely of those found within the respective 2-ligand complexes? In this case, as derived in Fig. 6 of Peacock and Jaynes [2],

$$n_{AB}*n_{AC}*n_{BC}=n_{ABC},$$

and we already have enough information from analysis of the 2-ligand complexes to predict the behavior of the entire 3-ligand system (see Fig. 6 of Peacock and Jaynes [2], which also provides a straightforward method for testing whether this relationship holds in any particular case).

We can follow an analogous procedure to characterize a 4-ligand system. If all of the interactions leading to cooperativity are contained within pairwise interaction domains that are not significantly affected by higher-order complex formation, then the sum of the free energies from the pairwise interactions equals the total cooperative free energy of the entire complex, and:

$$n_{ABCD}=n_{AB} * n_{AC} * n_{BC} * n_{AD} * n_{BD} * n_{CD}$$

Generally, for *j* ligands binding to distinct sites on a substrate and cooperating solely through pairwise interactions, the cooperativity factor is the product of the

$$j * (j-1)/2$$

possible pairwise cooperativity factors, which can each be measured by studying the ternary complex containing those two ligands, using the methods given here, and in Peacock and Jaynes [2].

## 4. SUMMARY AND CONCLUSIONS

Most currently used methods for quantifying cooperative binding by transcription factors to DNA do not provide the means to accurately predict binding behavior over a wide range of concentrations. The Hill equation and Hill plots are only useful for quantifying cooperativity when binding sites are equivalent, which is rarely the case for DNA binding proteins, as they typically bind to a variety of sites with different affinities. Even in the rare cases where cooperating sites are equivalent, different modes of cooperativity are possible, and these have sufficiently different behaviors in Hill plots as to make quantifying the cooperativity difficult.

In recent years, two main approaches have been used to compare cooperative binding of either 1) two different proteins to the same sites or 2) the same two proteins to different sites. Both of these, while they provide some useful information, have serious limitations for predicting binding behavior over a wide concentration range. One involves holding the concentrations of both a binding site oligo and one protein constant while varying that of the second protein. We have shown that this can result in binding curves that cross, precluding the general usefulness of this approach in isolation to accurately model binding behaviors of cooperating proteins. However, this approach is useful for quantifying cooperativity once individual proteinbinding site Kd’s have been determined.

A second approach is to compare the ability of unlabeled oligos containing different sites to compete for binding (by two cooperating proteins) to a labeled binding site-containing oligo. We have shown that this approach, too, provides only partial information about the system and can therefore give misleading results. For example, two oligos can compete very similarly under one set of conditions, while the occupancies of these sites can be very different under different conditions.

This analysis of existing methods argues for a more comprehensive approach that can use experimental data obtained over a limited range of concentrations to predict binding behavior over the full concentration range. We therefore developed two new tools to achieve this end. The first tool is a new approach to determining individual protein-binding site Kd's. Active protein concentrations must be determined in order to obtain accurate Kd's, and our approach allows the simultaneous determination of both of these, as parameters in non-linear regression analysis, using data from oligo competition assays. We described an optimized approach to give maximum accuracy, which mandates using the lowest concentration of labeled oligo which allows robust quantitation, along with an amount of protein that gives around 50% occupancy. Holding both of these constant, increasing amounts of unlabeled oligo identical in sequence to the labeled oligo are added. Quantifying the resulting reduced binding to the labeled oligo gives a data set that is fed into freely available software (e.g., at "statpages.org/nonlin.html"), along with the equation we have derived. This can have very significant advantages over previously described methods that involve saturation binding (increasing concentrations of protein). One advantage is that our method typically requires lower protein concentrations, avoiding problems of precipitation or aggregation at high concentrations. The second advantage is that both active protein concentration and Kd can be accurately determined without prior knowledge of either parameter.

The second new tool is the means to extract, via regression analysis, the cooperativity factor for binding by a pair of proteins once the individual protein–site Kd's have been determined. This involves holding the concentrations of a labeled oligo carrying the two cooperating binding sites and one of the proteins constant, while measuring cooperative complex formation as the other protein concentration is varied. Armed with the cooperativity factor and the two individual Kd's, binding behavior can be predicted and compared over the full concentration range of each interacting species.

We also provide modifications to these methods that extend their applicability in special cases. First, in cases where contaminating DNA binding proteins co-purify with the specific protein being studied, it is often advantageous to include non-specific DNA in the experiments used to measure the Kd, which reduces significantly the inaccuracy that can result from these contaminating proteins competing for binding to the labeled oligo. We describe the method and give equations for regression analysis to find the specific Kd and the non-specific Kd along with the active protein concentration. Second, we provide a means to extract both Kd's and the cooperativity factor (as parameters in regression analysis) for cases where only one Kd can be measured directly, which may occur when one cooperating protein binds very weakly on its own. This requires simultaneously measuring both ternary complex formation and the accompanying single-protein complex as the weakly binding protein's concentration is varied.

We provide a summary of the main formulae used in this paper, along with their applications in the methodology developed here, in Fig. S1.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research was supported by the National Institutes of Health, Awards R01GM050231 and R01GM117458 to JBJ. We give a special thanks to Miki Fujioka for the graphical abstract.

## References

1. Hill AV. The combinations of haemoglobin with oxygen and with carbon monoxide I. Biochem J. 1913; 7:471–480. [PubMed: 16742267]
2. Peacock J, Jaynes JB. Mathematical toolkit for quantitative analysis of cooperative binding of two or more ligands to a substrate. MethodsX. submitted.
3. Wyman, J., Gill, SJ. Binding and Linkage: Functional Chemistry of Biological Macromolecules. University Science Books; 1990.
4. Weiss JN. The Hill equation revisited: uses and misuses. FASEB J. 1997; 11:835–841. [PubMed: 9285481]
5. Stefan MI, Le Novère N. Cooperative Binding. PLoS Comp Biol. 2013; 9:e1003106.
6. Freire, Ernesto, Schön, Arne, Velazquez-Campoy, Adrian. Methods in Enzymology. Vol. 455. Elsevier Inc; 2009. Isothermal Titration Calorimetry: General Formalism Using Binding Polynomials; p. 127-155.(Chapter 5)
7. Bardsley WG. Factorability of the Allosteric Binding Polynomial and Graphical Manifestations of Cooperativity in Third Degree Saturation Functions. J Theor Biol. 1977; 67:407–431. [PubMed: 904322]
8. Acerenza L, Mizraji E. Cooperativity: A unified view. Biochimica et Biophysica Acta. 1997; 1339:155–166. [PubMed: 9165110]
9. Haiech J, Gendrault Y, Kilhoffer M-C, Ranjeva R, Madec M, Lallement C. A general framework improving teaching ligand binding to a macromolecule. Biochimica et Biophysica Acta (BBA) - Mol Cell Res. 2014; 1843:2348–2355.
10. Garcés JL, Acerenza L, Mizraji E, Mas F. A Hierarchical Approach to Cooperativity in Macromolecular and Self-Assembling Binding Systems. Journal of Biological Physics. 2008; 34:213–235. DOI: 10.1007/s10867-008-9116-x [PubMed: 19669504]
11. Garcés JL, Rey-Castro C, David C, Madurga S, Mas F, Pastor I, Puy J. Model-Independent Link between the Macroscopic and Microscopic Descriptions of Multidentate Macromolecular Binding: Relationship between Stepwise, Intrinsic, and Microscopic Equilibrium Constants. The Journal of Physical Chemistry B. 2009; 113:15145–15155. DOI: 10.1021/jp9041815
12. McGhee JD, von Hippel PH. Theoretical Aspects of DNA-Protein Interactions: Co-operative and Non-co-operative Binding of Large Ligands to a One-dimensional Homogeneous Lattice. J Mol Biol. 1974; 86:469–489. [PubMed: 4416620]
13. Bardsley WG, Waight RD. Factorability of the Hessian of the Binding Polynomial. The Central Issue Concerning Statistical Ratios Between Binding Constants, Hill Plot Slope and Positive and Negative Co-operativity. J Theor Biol. 1978; 72:321–372. [PubMed: 661345]
14. Briggs WE. Cooperativity and Extrema of the Hill Slope for Symmetric Protein-Ligand Binding Polynomials. J Theor Biol. 1984; 108:77–83. [PubMed: 6748683]
15. Briggs WE. The Relationship between Zeros and Factors of Binding Polynomials and Cooperativity in Protein-Ligand Binding. J Theor Biol. 1985; 114:605–614. [PubMed: 4021509]
16. Kowalczykowski SC, Paul LS, Lonberg N, Newport JW, McSwiggen JA, von Hippel PH. Cooperative and Noncooperative Binding of Protein Ligands to Nucleic Acid Lattices: Experimental Approaches to the Determination of Thermodynamic Parameters. Biochemistry. 1986; 25:1226–1240. [PubMed: 3486003]
17. Yuan D, Ma X, Ma J. Recognition of Multiple Patterns of DNA Sites by Drosophila Homeodomain Protein Bicoid. J Biochem. 1999; 125:809–817. [PubMed: 10101296]

18. Courey AJ. Cooperativity in transcriptional control. Curr Biol. 2001; 11:R250–R252. [PubMed: 11413011]

19. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of Specificity in Protein-DNA Recognition. Annu Rev Biochem. 2010; 79:233–69. [PubMed: 20334529]

20. Barozzi I, Simonatto M, Bonifacio S, Yang L, Rohs R, Ghisletti S, Natoli G. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. Mol Cell. 2014; 54:844–857. [PubMed: 24813947]

21. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. Absence of a simple code: How transcription factors read the genome. Trends Biochem Sci. 2014; 39:381–399. [PubMed: 25129887]

22. Mao C, Carlson NG, Little JW. Cooperative DNA-Protein Interactions: Effects of Changing the Spacing Between Adjacent Binding Sites. J Mol Biol. 1994; 235:532–544. [PubMed: 8289280]

23. Liu Z, Little JW. The Spacing Between Binding Sites Controls the Mode of Cooperative DNA-protein Interactions: Implications for Evolution of Regulatory Circuitry. J Mol Biol. 1998; 278:331–338. [PubMed: 9571055]

24. Parker DS, White MA, Ramos AI, Cohen BA, Barolo S. The cis-Regulatory Logic of Hedgehog Gradient Responses: Key Roles for Gli Binding Affinity, Competition, and Cooperativity. Science Signaling. 2011; 4(176):ra38.doi: 10.1126/scisignal.2002077 [PubMed: 21653228]

25. Gillitzer E, Chen G, Stenlund A. Separate domains in E1 and E2 proteins serve architectural and productive roles for cooperative DNA binding. EMBO J. 2000; 19:3069–3079. [PubMed: 10856250]

26. Henriksson-Peltola P, Sehlén W, Haggård-Ljungquist E. Determination of the DNA-binding kinetics of three related but heteroimmune bacteriophage repressors using EMSA and SPR analysis. Nucl Acids Res. 2007; 35:3181–3191. DOI: 10.1093/nar/gkm172 [PubMed: 17412705]

27. Dodd IB, Shearwin KE, Perkins AJ, Burr T, Hochschild A, Egan JB. Cooperativity in long-range gene regulation by the λ CI repressor. Genes Dev. 2004; 18:344–354. [PubMed: 14871931]

28. Neo SJ, Su X, Thomsen JS. Surface Plasmon Resonance Study of Cooperative Interactions of Estrogen Receptor α and Transcriptional Factor Sp1 with Composite DNA Elements. Anal Chem. 2009; 81:3344–3349. [PubMed: 19331400]

29. Melzer R, Verelst W, Theißen G. The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in 'floral quartet'-like complexes in vitro. Nucl Acids Res. 2009; 37:144–157. DOI: 10.1093/nar/gkn900 [PubMed: 19033361]

30. Majka J, Speck C. Analysis of Protein–DNA Interactions Using Surface Plasmon Resonance. Adv Biochem Engin/Biotechnol. 2007; 104:13–36. DOI: 10.1007/10_026

31. Baillat D, Bègue A, Stéhelin D, Aumercier M. ETS-1 Transcription Factor Binds Cooperatively to the Palindromic Head to Head ETS-binding Sites of the Stromelysin-1 Promoter by Counteracting Autoinhibition. J Biol Chem. 2002; 277:29386–29398. [PubMed: 12034715]

32. Matos RG, Barbas A, Arraiano CM. Comparison of EMSA and SPR for the Characterization of RNA–RNase II Complexes. Protein J. 2010; 29:394–397. DOI: 10.1007/s10930-010-9265-1 [PubMed: 20589527]

33. Burgess, Darren J. Shaping up transcription factor binding. Nat Rev Genet. 2015; 16:258–259. DOI: 10.1038/nrg3944

34. Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, Rohs R, Mann RS. Deconvolving the Recognition of DNA Shape from Sequence. Cell. 2015; 161:307–318. [PubMed: 25843630]

35. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordân R, Rohs R. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci USA. 2015; 112:4654–4659. [PubMed: 25775564]

36. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, Mann RS. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell. 2011; 147:1270–1282. [PubMed: 22153072]

37. Hu P, Shen Z, Tu H, Zhang L, Shi T. Integrating multiple resources to identify specific transcriptional cooperativity with a Bayesian approach. Bioinformatics. 2014; 30:823–830. DOI: 10.1093/bioinformatics/btt596 [PubMed: 24192543]

38. Fujioka M, Gebelein B, Cofer ZC, Mann RS, Jaynes JB. Engrailed cooperates directly with Extradenticle and Homothorax on a distinct class of homeodomain binding sites to repress sloppy paired. Dev Biol. 2012; 366:382–392. DOI: 10.1016/j.ydbio.2012.04.004 [PubMed: 22537495]

39. Chen Q-H, Bylund David B. A Mathematical Theory of Competitive Binding Assays. Receptors and Signal Transduction. 1997; 7:73–84. [PubMed: 9392436]

40. Saltelli A, Ratto M, Tarantola S, Campolongo F. Sensitivity Analysis for Chemical Models. Chemical Reviews. 2005; 105:2811–2828. DOI: 10.1021/cr040659d [PubMed: 16011325]

41. Motulsky, H., Arthur, C. Fitting Models to Biological Data using Linear and Nonlinear Regression. GraphPad Software Inc; San Diego, CA: 2003.

42. Churion, K., Liu, Y., Hsiao, H., Matthews, KS., Bondos, SE. Measuring Hox-DNA Binding by Electrophoretic Mobility Shift Analysis. In: Graba, Y., Rezsohazy, R., editors. Hox Genes: Methods and Protocols, Methods in Molecular Biology. Vol. 1196. p. 211-230.(Chapter 13)© Springer Science+Business Media New York 2014

43. Wyman, Jeffries, Jr. Linked Functions and Reciprocal Effects in Hemoglobin: A Second Look; Advances in Protein Chemistry. 1964. p. 223-286.http://www.sciencedirect.com/science/article/pii/S0065323308601904?via%3Dihub

44. Di Cera E. Site-specific thermodynamics: understanding cooperativity in molecular recognition. Chemical Reviews. 1998; 98:1563–1592. [PubMed: 11848942]

45. Sackett DL, Saroff HA. The multiple origins of cooperativity in binding to multi-site lattices. FEBS Letters. 1996; 397:1–6. [PubMed: 8941702]

46. Noro B, Lelli K, Sun L, Mann RS. Competition for cofactor-dependent DNA binding underlies Hox phenotypic suppression. Genes Dev. 2011; 25:2327–2332. DOI: 10.1101/gad.175539.111 [PubMed: 22085961]

47. Riley*, TR., Slattery*, M., Abe, N., Rastogi, C., Liu, D., Mann, RS., Bussemaker, HJ. SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes. In: Graba, Y., Rezsohazy, R., editors. Hox Genes: Methods and Protocols, Methods in Molecular Biology. Vol. 1196. p. 255-278.(Chapter 16)© Springer Science +Business Media New York 2014 [*these authors made equal contributions]

48. Biggin MD. Animal Transcription Networks as Highly Connected, Quantitative Continua. Dev Cell. 2011; 21:611–626. DOI: 10.1016/j.devcel.2011.09.008 [PubMed: 22014521]

49. Shea MA, Ackers GK. The $O_R$ Control System of Bacteriophage Lambda, A Physical-Chemical Model of Gene Regulation. J Mol Biol. 1985; 181:211–230. [PubMed: 3157005]

50. Darling PJ, Holt JM, Ackers GK. Coupled Energetics of λ cro Repressor Self-assembly and Site-specific DNA Operator Binding II: Cooperative Interactions of cro Dimers. J Mol Biol. 2000; 302:625–638. [PubMed: 10986123]

51. Kim TK, Maniatis T. The Mechanism of Transcriptional Synergy of an In Vitro Assembled Interferon-β Enhanceosome. Mol Cell. 1997; 1:119–129. [PubMed: 9659909]

52. Carey M. The Enhanceosome and Transcriptional Synergy. Cell. 1998; 92:5–8. [PubMed: 9489694]

53. Merika M, Thanos D. Enhanceosomes. Curr Opin Genet Dev. 2001; 11:205–208. [PubMed: 11250145]

54. Uhl JD, Cook TA, Gebelein B. Comparing anterior and posterior Hox complex formation reveals guidelines predicting *cis*-regulatory elements. Dev Biol. 2010; 343:154–166. [PubMed: 20398649]

55. Mann RS, Lelli KM, Joshi R. Hox Specificity: Unique Roles for Cofactors and Collaborators. Curr Top Dev Biol. 2009; 88:63–101. [PubMed: 19651302]

56. Gebelein B, Culi J, Ryoo HD, Zhang W, Mann RS. Specificity of Distalless repression and limb primordial development by abdominal Hox proteins. Dev Cell. 2002; 3:487–498. [PubMed: 12408801]

57. Gebelein B, McKay DJ, Mann RS. Direct integration of Hox and segmentation gene inputs during Drosophila development. Nature. 2004; 431:653–659. [PubMed: 15470419]

58. Riley TR, Lazarovici A, Mann RS, Bussemaker HJ. Building accurate sequence-to-affinity models from high-throughput *in vitro* protein-DNA binding data using FeatureREDUCE. eLife. 2015; 4:e06397.doi: 10.7554/eLife.06397 [PubMed: 26701911]

59. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. Cell. 2006; 124:47–59. [PubMed: 16413481]

60. Jin Y, Zhong H, Vershon AK. The Yeast a1 and α2 Homeodomain Proteins Do Not Contribute Equally to Heterodimeric DNA Binding. Mol Cell Biol. 1999; 19:585–593. [PubMed: 9858582]

61. Andersen PS, Schuck P, Sundberg EJ, Geisler C, Karjalainen K, Mariuzza RA. Quantifying the Energetics of Cooperativity in a Ternary Protein Complex. Biochemistry. 2002; 41:5177–5184. DOI: 10.1021/bi0200209 [PubMed: 11955066]

## HIGHLIGHTS

- Hill plots remain prominent in biology, but can mask cooperativity
- Effective modeling of binding by two ligands requires the use of 3 parameters
- We develop novel ways to find these parameters for two cooperating ligands
- We show how they can be used to enhance the power of established methods
- We describe how this framework can be extended to multiple cooperating ligands

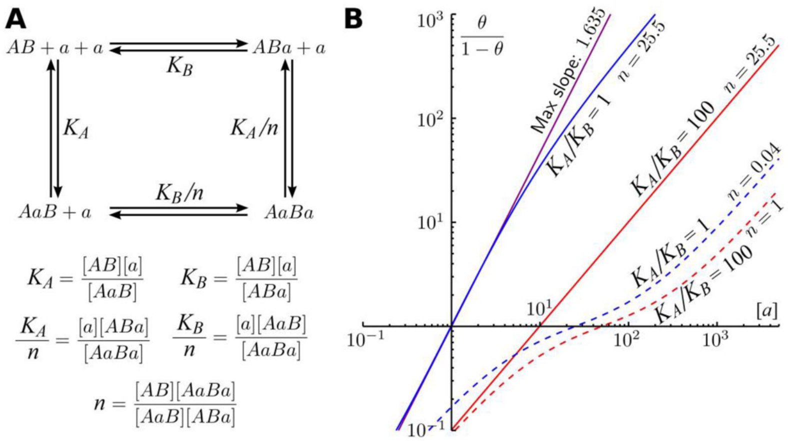


**Fig. 1. Barriers to determining cooperativity from Hill plots**

**A:** Model for cooperative binding to *AB* (substrate with two distinct binding sites) by ligand *a*. The ternary complex *AaBa* can dissociate in two ways, losing *a* from either the *A* or the *B* site first. Defining the Kd's for the ternary complex as $K_A/n$ and $K_B/n$ reduces the number of variables, because, from the definitions of the Kd's (below the line), $K_A$ divided by $K_A/n$ gives the same thing as $K_B$ divided by $K_B/n$. **B.** Hill plots for two binding sites with the same or different Kd's. From the model in A, the ratio (fractional occupancy)/(1 − fractional occupancy), which is $(K_A+K_B+2n[a])[a]/\{2K_AK_B+(K_A+K_B)[a]\}$ (derived in Fig. 1A of Peacock and Jaynes [2]), was used to generate Hill plots. Concentration units (for Kd's and [*a*]) are arbitrary. The case where $K_A = K_B = 5$ and $n = 25.5$ is shown as a solid blue curve, along with a tangent line (purple) at the point of maximum slope. Also shown are: the same equivalent sites, but with negative cooperativity ($n = 0.04$, dashed blue), and the case of two non-equivalent sites ($K_A = 5$, $K_B = 500$), either with positive cooperativity ($n = 25.5$, solid red) or with no cooperativity ($n = 1$, dashed red). Note the similarity in shape of the plots for equivalent sites with negative cooperativity and for non-equivalent sites without cooperativity. Also note that for two non-equivalent sites, when *n* approaches the value $(K_A+K_B)$^2/4$K_AK_B$ (derived in Fig. 1A of Peacock and Jaynes [2]), the plot approaches a straight line of slope 1, which is indistinguishable from equivalent sites with no cooperativity. Thus, without prior knowledge that sites are equivalent, Hill plots are at best ambiguous for identifying cooperativity.

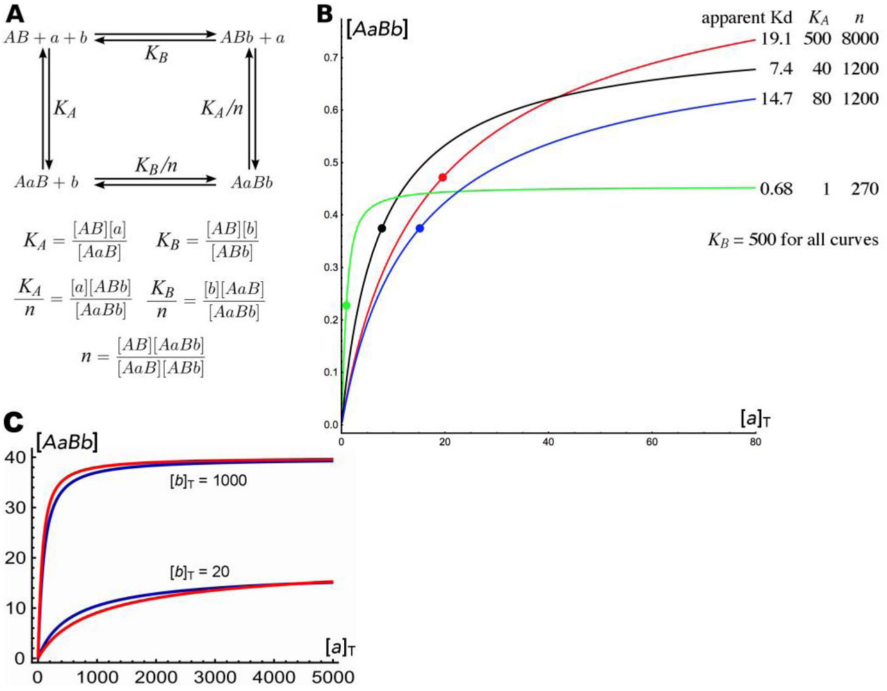
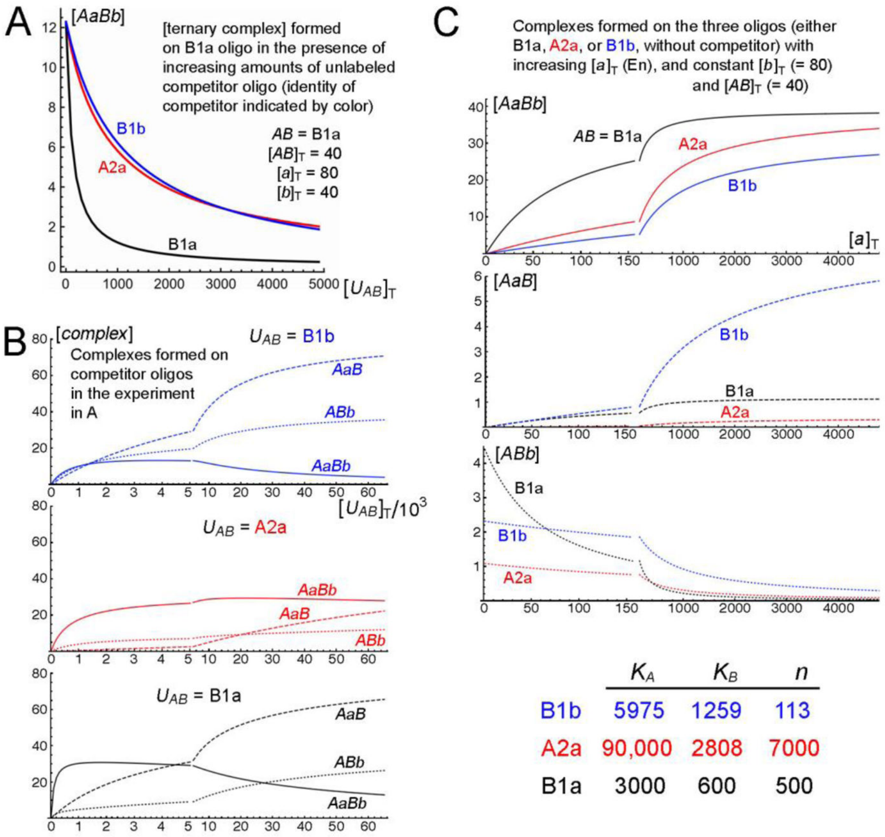


**Fig. 2. Two cooperating proteins binding to two different sites**

**A:** schematic of binding equilibria. Either protein can bind first. The upper path from left to right represents initial binding by protein *b*. The equilibrium concentration of the *ABb* single-protein complex is governed by its Kd, $K_B$. It can then bind protein *a* to form the ternary complex *AaBb*. The alternative pathway to the ternary complex is similarly diagrammed on the left. As the definitions of the various dissociation constants below show, not all 4 are independent. If we divide $K_B$ by the Kd that governs the dissociation of *b* from *AaBb*, we get the same quantity that we get if we divide $K_A$ by the Kd that governs the dissociation of *a* from *AaBb*. It is therefore convenient to define this ratio as the cooperativity factor *n*. **B:** graphs of [*AaBb*] as a function of increasing $[a]_T$, holding $[b]_T$ constant. For all graphs, $[AB]_T = 1$, $[b]_T = 2$, and $K_B = 500$, while $K_A$'s and cooperativity factors vary. The apparent Kd (based on a single-site model, see Fig. 2B in Peacock and Jaynes [2]) is the [*a*] at the point of half-maximal [*AaBb*], which is marked by a dot for each curve. Note that the relative amounts of ternary complex depend strongly on $[a]_T$. The green curve, which has the lowest apparent Kd (0.68), actually shows the lowest [*AaBb*] at high $[a]_T$. This is due to its relatively low *n*, which determines the [*AaBb*] at saturation with protein *a*, independent of $K_A$. The black curve crosses the red curve, and also shows less binding at high $[a]_T$ due to a lower *n*. The blue curve does not cross the red curve, and has a lower [*AaBb*] at all values of $[a]_T$, despite having a lower apparent Kd! So, a ranking of apparent Kd's from this type of experiment is not predictive of relative ternary complex formation overall. Derivations of expressions relating [*AaBb*] to $[a]_T$ (and to [*a*], [*AaB*], and [*AB*]) are given in Fig. 2A of Peacock and Jaynes [2]. Derivations of expressions for $[AaBb]_{max}$ and for finding the apparent Kd are given in Fig. 2B of Peacock and Jaynes [2]. **C:** relative ternary complex formation can be qualitatively different depending on the fixed $[b]_T$ chosen for the experiment. The two pairs of curves (upper and lower) represent [*AaBb*] formed on two sites (red and dark blue) as a function of increasing $[a]_T$, differing only in the fixed $[b]_T$. Note that at the lower $[b]_T$, the site represented by the dark blue curve ($K_A$ = 5975, $K_B$ = 400, $n$ = 113) forms more ternary complex throughout most of the experimental range, while at the higher $[b]_T$, the site represented by the red curve ($K_A$ = 90,000, $K_B$ = 2808, $n$ = 7000) forms more over the entire range. This illustrates another limitation of modeling cooperative binding using a single parameter. **NOTE:** Concentration units are not specified, because in all cases, these units (which includes the concentrations of ligands and substrate, as well as Kd's) can be factored out of the governing equations, and do not affect the shapes of curves, or any of the conclusions.



| | $K_A$ | $K_B$ | $n$ |
|---|---|---|---|
| B1b | 5975 | 1259 | 113 |
| A2a | 90,000 | 2808 | 7000 |
| B1a | 3000 | 600 | 500 |

**Fig. 3. Competition curves measuring ternary complex have limited predictive power**

Concentrations and Kd’s are in nM. **A:** Ternary complex as a function of competitor. For each curve, the [*AaBb*], labeled ternary complex, is graphed as a function of an unlabeled competitor. The lower (black) curve is self-competition by the high-affinity site (B1a [38]), where the unlabeled competitor, $U_{AB}$, is the same DNA sequence as the labeled binding site, *AB*. The other two curves show competition with two binding site oligos that have very different Kd’s and cooperativity factors (*n*), yet compete similarly for ternary complex formation by labeled B1a oligo. Note that self-competition is much more effective at all concentrations shown than is competition by either of the other oligos, while the other two oligos compete very similarly over a wide concentration range.

**B:** Forms of competitor oligo in the competition experiment. Each graph shows the 3 bound forms for one of the competitor oligos. In each case, the solid line shows [*AaBb*], the dashed line shows [*AaB*], and the dotted line shows [*ABb*]. The upper panel shows the less cooperative low-affinity site (B1b, blue), the middle panel shows the more cooperative low-affinity site (A2a, red), and the lower panel shows the high-affinity site (B1a, black). The left and right sections of each curve show two different ranges of $[a]_T$, on two different scales. Note that at the higher concentrations of competitor, B1b forms mostly single-protein complexes, while A2a forms mostly ternary complex, reflecting its much higher cooperativity. At concentrations well beyond the range shown, all of each protein is incorporated into single-protein complexes, as the proteins are distributed over a vast excess of oligo. For oligo B1b (blue), we see the approach to this limit, while for oligo A2a (red), this approach is beyond the range shown.

**C:** Concentrations of complexes as a function of $[a]_T$, holding $[b]_T$ constant (no competitor). Using the same oligos as in A (and B), the concentrations of the various protein-containing forms are graphed for a similar experiment as in Fig. 2. The left and right sections of each curve show two different ranges of $[a]_T$, on two different scales. The top graph shows [*AaBb*] for each oligo, color-coded as in A and B. Note that despite the similarity of the blue and red competition curves in A, the oligo with the higher value of *n* (red) forms more ternary complex at all $[a]_T$, and the red curve approaches the black curve at high $[a]_T$. This provides a plausible explanation for the *in vivo* behavior of the binding sites represented by the blue and red curve: the one with the higher *n* (red) is more potent. It is more similar to the black curve than to the blue curve at high $[a]_T$, suggesting that the ability to form ternary complexes at high $[a]_T$ may explain the relative functionality of these binding sites *in vivo*. The middle and bottom graphs show [*AaB*] (dashed) and [*ABb*] (dotted), respectively, for each oligo, color coded as above. As seen in the competition experiment in B, the less cooperative oligo forms more binary complexes (blue) than does the more cooperative oligo (red), especially *AaB* at high $[a]_T$, due to its having a lower Kd for binding each of the proteins. A similar phenomenon occurs at high concentrations of these oligos in the competition experiment: the less cooperative site sequesters more of each protein individually, while it forms less ternary complex than does the more cooperative site. These complexes are invisible in a competition assay, because the competitor oligo is unlabeled. For derivations of equations that can be used to generate these graphs, see Fig. 3A in Peacock and Jaynes [2]. For derivations of equations for graphing the total occupancy by each protein as a function of $[a]_T$, see Fig. 3B in Peacock and Jaynes [2].

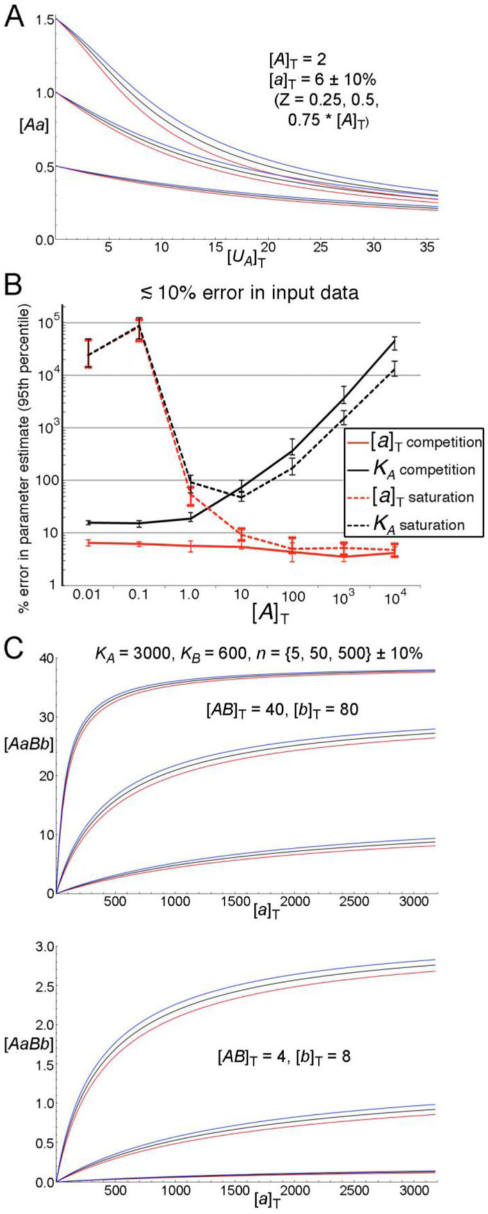


**Fig. 4. Illustrations of curve-fitting equations and error comparison**

**A:** Families of competition curves with different values of $[a]_T$ and $K_A$. Labeled binary complex, $[Aa]$, is graphed as a function of increasing total unlabeled binding site, $[U_A]_T$, with constant amounts of both labeled binding site, $[A]_T$, and total ligand, $[a]_T$. The applicable formula is $[U_A]_T = [A]_T * ([a]_T/[Aa] - K_A/([A]_T - [Aa]) - 1)$. Sets of data points $\{([U_A]_T, [Aa])\}$ along with known $[A]_T$ can be used to find both $K_A$ and $[a]_T$ as parameters using freely available curve fitting software (see text). The values used here are: $[A]_T = 2$ for all curves; $[a]_T = 6$, 5.4, and 6.6, and $K_A = \{1.5, 5, 16.5\}, \{1.3, 4.4, 14.7\}, \{1.7, 5.6, 18.3\}$, for the black, red, and blue curves, respectively. The values of $K_A$ were adjusted to give 3 sets of 3 curves each with the same 3 initial values (without competitor), 0.5, 1.0, and 1.5. Note that each set of curves with the same starting value diverges significantly as competitor increases. **B:** Performance of competition and saturation binding methods for simultaneously finding $[a]_{\mathrm{T}}$ and $K_A$. Monte Carlo analysis (100 trials are represented by each data point) of the accuracy of curve fitting to find both $[a]_{\mathrm{T}}$ and $K_A$ as parameters was run with 15-point data sets at 7 different $[A]_{\mathrm{T}}$ using either competition (fixed $[a]_{\mathrm{T}}$, varying $[U_A]_{\mathrm{T}}$, as illustrated in A) or standard saturation binding (varying $[a]_{\mathrm{T}}$, no competitor). Data sets were generated by introducing random errors into calculated values of $[Aa]$. These errors were randomly drawn from a normal distribution (centered on zero) such that 95% of the errors were within ±10% of the actual value (standard deviation = 5%, mean error = 4.0%, median error = 3.4%). Percent errors are shown in the values found for each parameter ($[a]_{\mathrm{T}}$ and $K_A$) using least-squares non-linear regression. These percent errors were ranked by increasing absolute value, and the 95th largest (out of 100) plotted, with errors bars extending between the 90th and 99th largest. These error bars represent a 95% confidence interval for the true value of the 95th error percentile, based on standard statistical analysis. Note that the best estimate for $K_A$ is provided by the competition method at low $[A]_{\mathrm{T}}$, which simultaneously provides a precise estimate for $[a]_{\mathrm{T}}$. See text for further explanation. **C:** Family of curves with different values of *n*. $[AaBb]$ is graphed as a function of $[a]_{\mathrm{T}}$, holding constant $[b]_{\mathrm{T}}$ and $[AB]_{\mathrm{T}}$, for 3 different values of the cooperativity factor ($n$). $K_A = 3000$, $K_B = 600$, $n = \{5, 50, 500\}$ for the black curves, $n = \{4.5, 45, 450\}$ for the red curves, and $n = \{5.5, 55, 550\}$ for the blue curves. The uppermost black curve corresponds to the black curve in Fig. 3C, top. Once $K_A$, $K_B$, and $[b]_{\mathrm{T}}$ are determined, the formula used to draw these curves can be used to find $n$ from sets of data points $\{([AaBb], [a]_{\mathrm{T}})\}$ using freely available software (see text). The formula is

$$[a]_{\mathrm{T}}=[AaBb]+[AaBb]\,\Bigl\{[AB]_{\mathrm{T}}+K_B-[b]_{\mathrm{T}}+\mathbf{Sqrt}\Bigl[([AB]_{\mathrm{T}}+K_B-[b]_{\mathrm{T}})\hat{}2+4\,K_B\,([b]_{\mathrm{T}}-[AaBb]+[AaBb]/n)\Bigr]\Bigr\}\,/\,\Bigl\{2\,n\,\Bigl([b]_{\mathrm{T}}-$$

$$+K_A\,[AaBb]\,\Bigl\{[AB]_{\mathrm{T}}+K_B+[b]_{\mathrm{T}}-2\,[AaBb]+\mathbf{Sqrt}\Bigl[([AB]_{\mathrm{T}}+K_B-[b]_{\mathrm{T}})\hat{}2+4\,K_B\,([b]_{\mathrm{T}}-[AaBb]+[AaBb]/n)\Bigr]\Bigr\}\,/\,\Bigl\{2\,n\,\Bigl(([A$$

Note that the upper set of curves in the upper graph (which differ among themselves only by a change in *n* of 10%, like each set of 3 closely situated curves) are very close together, making it difficult to determine this $n$ (= 500) using these values of $[AB]_{\mathrm{T}}$ and $[b]_{\mathrm{T}}$. However, when both are reduced by a factor of 10 (as shown in the lower graph), the upper set of curves (again representing $n$ = 500) diverge more. Thus, $n$ values that result in saturation of the probe ($AB$) can be more precisely determined by reducing its concentration, along with that of the fixed $[b]_{\mathrm{T}}$ (which is optimal for determining $n$ when it is similar to $[AB]$).

Peacock and Jaynes, 2017
Fig. S1.

**CONTENTS**

## Fig. 1:

For constructing a Hill plot for two sites (details given in Fig. 1A of Peacock and Jaynes [2]), first, fractional occupancy / (1 – fractional occupancy) is:

$$\frac{\phi}{1-\phi} = \frac{(k+L+2nf)f}{2kL+(k+L)f}$$

where *k* and *L* are the two individual site Kd's,
*n* is the cooperativity factor, and
*f* is the [free ligand].

Specializing this to two equivalent sites:

$$\frac{\phi}{1-\phi} = \frac{(k+nf)f}{k(k+f)}$$

The Hill plot is a straight line when:

🟦 $n = \frac{(k+L)^2}{4kL}$

in which case the apparent Kd is:

🟩 $k' = \frac{2kL}{k+L}$

That is, the plot is the same as one with Kd = *k*' and *n* = 1 (no cooperativity).

**Fig. 1B-I of Peacock and Jaynes [2]:**

The fractional occupancy of a set of *Z* equivalent sites **with 1-step cooperativity** is:

$$\theta = \frac{s\,(1+ns)^{Z-1}}{1+\frac{1}{n}\,(\,(1+ns)^{Z}-1)} = \frac{ns\,(1+ns)^{Z-1}}{n+\,(1+ns)^{Z}-1}$$

where $s = fr/k$, which is the free concentration of ligand divided by the equilibrium dissociation constant without cooperativity ($k/r$), and
$f$ is the free [ligand],
$r$ is the (microscopic) forward rate constant for binding of a ligand molecule to the complex,
$k$ is the (microscopic) dissociation rate constant for the singly bound complex,
$k/n$ is the (microscopic) dissociation rate constant for any ligand molecule from the complex except the first.
So,

$$\frac{\theta}{1-\theta} = \frac{ns\,(1+ns)^{Z-1}}{n-1+\,(1+ns)^{Z-1}} = \frac{ns}{(n-1)\,(1+ns)^{1-Z}+1}$$

The slope of the resulting Hill plot is:

$$1+\frac{(n-1)\,(Z-1)\,ns}{(n-1)\,(1+ns)\,+\,(1+ns)^{Z}}$$

The condition for maximum slope of the Hill plot is:

$$(n-1)\,(1+ns)^{1-Z}+1\,=\,(Z-1)\,ns$$

This expression can be solved explicitly for the variable *s* only for *Z* = 2, 3, or 4.
However, it can be used to get a simple formula for the maximum slope of a Hill plot with 1-step cooperativity
in terms of the value of *s* where the slope is maximum, by incorporating the condition for max. slope into the formula for max. slope.

The slope then reduces to the following (*s* here = the value of *s* which satisfies the condition for max. slope):

$$\frac{Zns}{1+ns} = \frac{Z}{1+\frac{1}{ns}}$$

The limit of this, as *n* goes to infinity, is *Z*.
This is because (*ns*) also goes to infinity as *n* goes to infinity (even though *s* goes to zero as *n* goes to infinity).

The fractional occupancy at the point of max. slope is 1 / (# of sites):

$$\theta = \frac{1}{Z}$$

The slope of a Hill plot **with progressive cooperativity**
(where the cooperativity increases at each binding step by the same factor, up to a maximum of $n$) is:

$$\frac{\sum_{i=1}^{Z} \frac{i}{(Z-i)!\,(i-1)!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^{i}}{\sum_{i=1}^{Z} \frac{1}{(Z-i)!\,(i-1)!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^{i}} - \frac{\sum_{i=1}^{Z-1} \frac{1}{(Z-i-1)!\,(i-1)!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^{i}}{\sum_{i=0}^{Z-1} \frac{1}{(Z-i-1)!\,i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^{i}}$$

This is a maximum at 50% occupancy, which is also where $s = n$^(–1/2).
So, the maximum slope is:

$$\frac{\sum_{i=1}^{Z} \frac{i}{(Z-i)!\,(i-1)!} n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}{\sum_{i=1}^{Z} \frac{1}{(Z-i)!\,(i-1)!} n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}} - \frac{\sum_{i=1}^{Z-1} \frac{1}{(Z-i-1)!\,(i-1)!} n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}{\sum_{i=0}^{Z-1} \frac{1}{(Z-i-1)!\,i!} n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}$$

**Fig. 2:**

Definitions of variables:
$h$ = [free probe DNA, "hot" probe]
$H$ = total ["hot" DNA]
$f$ = [free protein]
$a$ = [protein – site 1 complex] = [($fh$)]
$b$ = [site 2 – protein complex] = [($hf$)]
$A$ = [ternary complex] = [($fhf$)]

Dissociation (equilibrium) constants, including cooperativity factor $n$:
$k$ = dissociation constant of ($fh$), site 1 binary complex
$L$ = dissociation constant of ($hf$), site 2 binary complex
$k/n$ = dissociation constant of protein from site 1 of ternary complex ($fhf$)
$L/n$ = dissociation constant of protein from site 2 of ternary complex ($fhf$)

In order to get an expression connecting $A$ and $p$, with q constant, given $H$, $k$, $L$, and $n$, we can use the fact that $p = A + a + f$. So, from expressions for $a$ and $f$ that involve only the known quantities, we get the desired connection.

Such an expression for *a* is (from Fig. 2A of Peacock and Jaynes [2]):

🟦 $$a = \frac{A}{n}\left(\frac{H+L-q+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left(q-A+\frac{A}{n}\right)}\right)$$

and for *f* is:

🟩 $$f = \frac{kA}{n}\left(\frac{H+L+q-2A+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left((H-A)\,(q-A)-\frac{LA}{n}\right)}\right)$$

Since *f* goes to infinity as *p* goes to infinity, the saturation value of *A* (as *p* goes to infinity) occurs when the denominator of this expression is 0:

🟥 $$(H-A)\,(q-A)-\frac{LA}{n} = 0$$

or

🟦 $$A = \frac{1}{2}\left(H+q+\frac{L}{n}-\sqrt{\left(H+q+\frac{L}{n}\right)^2-4qH}\right)$$

Substituting the expressions for *a* and *f* above into $p = A + a + f$ gives:

🟦 $$p = A+\frac{A}{n}\left(\frac{H+L-q+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left(q-A+\frac{A}{n}\right)}\right)+\frac{kA}{n}\left(\frac{H+L+q-2A+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left((H-A)\,(q-A)-\frac{LA}{n}\right)}\right)$$

This can be used to graph *A* as a function of *p*, at constant *q* and *H*, given *k, L*, and *n*.
It can also be used to find *n* from data points (*A, p*) once the other variables have been determined, using curve fitting procedures.

Similar expressions for other variables are:

⬜ $$b = \frac{1}{2}\left(H+L+q-2A-\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}\right)$$

🟪 $$g = \frac{2L\left(q-A+\frac{A}{n}\right)}{H+L-q+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}$$

■ $$h = \frac{(q-A)\left(H-L-q+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}\right)-2\frac{LA}{n}}{2\left(q-A+\frac{A}{n}\right)}$$

When do two binding curves of [ternary complex] (*A*) as a function of [protein1] (*p*) cross each other?

From Fig. 2C of Peacock and Jaynes [2]:

A more cooperative site always saturates at a higher [protein1], so when the curves cross, the more cooperative site will usually have a lower initial slope. If *k* and *L* are the two Kd's for the site with cooperativity factor *n*, and $\mathbb{k}$ and $\mathbb{L}$ are the two Kd's for the site with cooperativity factor $\mathbb{m}$, then *n* must be < :

□ $$\frac{\left(\frac{k}{H}+1\right)\left(H+L+\sqrt{(H-q+L)^2+4qL}\right)+q\left(\frac{k}{H}-1\right)}{\left(\frac{\mathbb{k}}{H}+1\right)\left(H+\mathbb{L}+\sqrt{(H-q+\mathbb{L})^2+4q\mathbb{L}}\right)+q\left(\frac{\mathbb{k}}{H}-1\right)}\mathbb{m}$$

**Fig. 3:**

From Fig. 3A of Peacock and Jaynes [2]:

Definitions of variables:

*A* = ["hot" ternary complex]
*H* = total ["hot" DNA]
*U* = total [unlabeled DNA]
*f* = [free protein1]
*p* = total [protein1]
*q* = total [protein2]

Dissociation (equilibrium) constants:

for "hot" complexes, including cooperativity factor *n*:
*k* = dissociation constant protein1 – site1 binary complex
*L* = dissociation constant of site2 – protein2 binary complex
*k*/*n* = dissociation constant of protein1 from site1 of ternary complex
*L*/*n* = dissociation constant of protein2 from site2 of ternary complex

for unlabeled complexes, including cooperativity factor *m*:
*Q* = dissociation constant of protein1 – site1 binary complex
*R* = dissociation constant of site2 – protein2 binary complex
*Q*/*m* = dissociation constant of protein1 from site1 of ternary complex
*R*/*m* = dissociation constant of protein2 from site2 of ternary complex

In order to graph the various species as a function of $U$ (total [specific competitor]) parametrically, using $f$ as the parameter, we need each species as a function of $f$ and constants ($H, p, q, k, L, n, Q, R$, and $m$). First, we get $A$ as a function of $f$ and these constants, by "turning around" the expression originally derived to get $f$ as a function of $A$ and constants (given below).

The expression connecting $A$ and $f$ is cubic in $A$; see Fig. 3A of Peacock and Jaynes [2] for the explicit solution for $A$ as a function of $f$ from the following expression, which = 0:

$$\left(\left(m-\frac{R}{L}\right)(k+nf)^2-(n-1)\,(mf\,(2k+nf)+kQ)\right)\frac{k+nf}{nf\,(k+f)}A^3$$
$$+\left(\left(\frac{(q+2H)\,R}{L\,(k+f)}\right)(k+nf)^2-\left(m\,(f-p+q)+Q-R-p\frac{Q}{f}+2H\frac{Q+mf}{k+f}\right)(k+nf)+(nQ-mk)\,H-mL\,(k+f)\right)A^2$$
$$+\left(\frac{f}{k+f}\left(H\,(Q+mf)-\frac{R\,(H+2q)}{L}(k+nf)\right)+mf^2+(Q-R+m\,(q-p))\,f-pQ\right)nHA+\frac{qR\,(Hnf)^2}{L\,(k+f)}$$

We plug that solution for $A$ as a function of $f$ into the following formulae, for each species as a function of $A$ and $f$, then graph each of them vs. $U$ parametrically (with $f$ as parameter). The solution to the cubic, $A$ as a function of $f$, is substituted for $A$ in each case, and the parametric variable (such as $t$, in "Graphing Calculator") is substituted for $f$ (in the following and in the solution to the cubic, so that $A(f)$ becomes $A(t)$).

From Fig. 3A of Peacock and Jaynes [2]:

🟥 $$U=\left(pk-kA\left(1-\frac{1}{n}\right)+(p-k-H)\,f-f^2\right)\left(\frac{nHRf\,(Q+f)+((mL-nR)\,f^2+((mL-R)\,k+(L-nR)\,Q)\,f+(L-R)\,Qk)\,A}{(nHRf\,(k+f)+((mL-nR)\,f^2+((2mL-(n+1)\,R)\,f+(mL-R)\,k)\,k)\,A)\,f}\right)$$

similar expressions for the other variables (in terms of $A, f$, and constants) are:

⬛ $$a=\frac{Hf-\left(f+\frac{k}{n}\right)A}{k+f}$$

🟪 $$b=\frac{kA}{nf}$$

⬜ $$h=\frac{H-A\left(1+\frac{k}{nf}\right)}{1+\frac{f}{k}}$$

🟥 $$g=\frac{LA\,(k+f)}{(H-A)\,nf-kA}$$

🟦 $$B = \frac{q - \frac{L\,(k+f)}{\left(\frac{H}{A} - 1\right) nf - k} - \left(\frac{k}{nf} + 1\right) A}{\frac{Q}{mf} + 1}$$

⬜ $$c = \frac{p - f - \frac{kA\left(1 - \frac{1}{n}\right) + Hf}{k+f}}{1 + \frac{m}{R}\,\frac{L\,(k+f)}{\left(\frac{H}{A} - 1\right) nf - k}}$$

🟦 $$d = \frac{q - \frac{L\,(k+f)}{\left(\frac{H}{A} - 1\right) nf - k} - \left(\frac{k}{nf} + 1\right) A}{1 + \frac{mf}{Q}}$$

🟨 $$u = R\,\frac{\frac{q - \left(\frac{k}{nf} + 1\right) A}{L\,(k+f)}\left(\left(\frac{H}{A} - 1\right) nf - k\right) - 1}{1 + \frac{mf}{Q}}$$

When the competitor is the same as probe (except for being unlabeled), the cubic solution fails, because the coefficient of *A*-cubed in the expression above vanishes (*Q, R*, and *m* become *k, L,* and *n*, respectively), and a simpler set of formulae are both necessary and easier to use. The quadratic in *A* as a function of *f* is:

🟩 $$\left(q\,\frac{(k+nf)^2}{k+f} - nL\,(k+f) - (k+nf)\left((k+nf)\left(1 - \frac{p}{f}\right) + nq - L\right)\right) A^2 + \left(nf^2 + (k - L + n\,(q-p))\,f - pk - \frac{2qf\,(k+nf)}{k+f}\right) nHA + \frac{q\,(Hnf)^2}{k+f} = 0$$

Derivations of the following specialized forms (when competitor is simply unlabeled probe) are available on request;

⬛ $$U = \frac{pk + (p-k)f - f^2}{f + \left(1 - \frac{1}{n}\right)\frac{kA}{H}} - H$$

$a$, $b$, $g$, and $h$ are the same as above, since they don't contain $Q$, $R$, or $m$, while:

🟩 $$B = \frac{q - \frac{L\,(k+f)}{\left(\frac{H}{A} - 1\right) nf - k}}{\frac{k}{nf} + 1} - A$$

🟦 $$c = \frac{p - f - \frac{kA\left(1 - \frac{1}{n}\right) + Hf}{k+f}}{1 + n\frac{k+f}{\left(\frac{H}{A} - 1\right) nf - k}}$$

🟨 $$d = \frac{q - \frac{L\,(k+f)}{\left(\frac{H}{A} - 1\right) nf - k}}{1 + \frac{nf}{k}} - \frac{kA}{nf}$$

🟦 $$u = \frac{\frac{q - \left(\frac{k}{nf} + 1\right) A}{k+f}\left(\left(\frac{H}{A} - 1\right) nf - k\right) - L}{1 + \frac{nf}{k}}$$

An alternate approach to graphing *A* as a function of *U* is to do this non-parametrically,
but this requires the solution of a quartic expression in *A* (as a function of *f* and constants) rather than "merely" a cubic one,
which some software packages cannot handle, or cannot do so over the full range of the constants;
this solution is given in Fig. 3A of Peacock and Jaynes [2].

**From Fig 3B of Peacock and Jaynes [2]:**

*T* = total [bound protein1] = *a* + *A*
*S* = total [bound protein2] = *b* + *A*
To graph total occupancy by proteins 1 or 2 (*S* or *T* becomes *y*) as a function of total [protein1] (*p* becomes *x*):

🟩 $$p = (LS - (H-S)\,(q-S))\left(\frac{k}{n\,(H-S)\,(q-S) - LS} + \frac{\frac{n}{L} + \frac{1}{q-S}}{n-1}\right)$$

and reversing the roles of proteins 1 and 2 gives an implicit expression for *T* as a function of *p*:

🟨 $$q = (kT - ((H-T)\,(p-T)))\left(\frac{L}{n\,(H-T)\,(p-T) - kT} + \frac{\frac{n}{k} + \frac{1}{p-T}}{n-1}\right)$$

**Fig. 4:**

(Details are given in Fig. 4A of Peacock and Jaynes [2]:)
Using data of [single-protein complex] (= *a*) formed on labeled "hot" oligo (total concentration = *H*)
as a function of increasing amounts of competitor oligo of the same sequence as the "hot" oligo, where
*V* = total [unlabeled DNA, Kd = *k*, the same as that of the labeled DNA]

key for single-character notation used here:

| below | main text | description |
|---|---|---|
| $H$ | $[A]_T$ | total concentration of labeled substrate |
| $V$ | | total concentration of specific unlabeled substrate with same dissociation constant as that of labeled substrate |
| $p$ | $[a]_T$ | total concentration of ligand |
| $k$ | $K_A$ | dissociation constant of ligand from labeled complex |
| $a$ | $[Aa]$ | concentration of labeled complex |

Formula for curve fitting using data points (*a*, *V*), with constants *H* (known), and *p* and *k* (to be determined as parameters in curve fitting):

$$V = H\left(\frac{p}{a} - \frac{k}{H-a} - 1\right)$$

In order to get initial estimates for curve fitting for *p* and *k*,
use the value of *a* when *V* = 0, called *Z*, and one other data point (*a, V*) in the formulae:

$$p = \frac{Za}{H}\left(\frac{V\left(1-\frac{a}{H}\right)}{Z-a} + 1\right)$$

and

$$k = \left(\frac{V}{\frac{Z}{a} - 1} - H\right)\left(1 - \frac{Z}{H}\right)\left(1 - \frac{a}{H}\right)$$

To do the same thing as the above but in the presence of a constant amount of non-specific competitor DNA of known concentration (*D*)
(we assume that this is in large excess over the ligand, so that cooperativity on this DNA is not relevant;
details are given in Fig. 4C of Peacock and Jaynes [2]):

*D* = total [unlabeled DNA, non-specific] of dissociation constant *Q*
*Z* = [labeled binary complex] in the absence of specific competitor (that is, when *V* = 0).
This is just the data point (*V, a*) = (0, *Z*), but in practice, it should be determined from multiple trials for maximum accuracy.

For curve fitting with data points (*V, a*), to determine *p* and *k* as parameters, knowing *D, H*, and *Z*:

🟪 $$V = H\left(\frac{p}{a} - \frac{k}{H-a} - 1 - \frac{1}{\dfrac{H-a}{\left(\frac{p}{Z}-1\right)(H-Z)-k} - \dfrac{Z-a}{D\left(1-\frac{Z}{H}\right)}}\right)$$

and using the following to get an initial estimate for *k*, from (*V*=0, *Z*) and some other point (*V*≠0, *a*);
use as the initial estimate for *p* the measured [total protein] (e.g., Bradford assay).
*k* = :

$$\frac{1}{2}\left(Z + a - 2H - V\left(1-\frac{a}{H}\right) - 2p + Hp\left(\frac{1}{a}+\frac{1}{Z}\right) - \sqrt{\left(Z + a - 2H - V\left(1-\frac{a}{H}\right) - 2p + Hp\left(\frac{1}{a}+\frac{1}{Z}\right)\right)^2 - 4\,(H-a)\,(H-Z)\left(\frac{D}{a-Z}\left(\frac{p}{a}-\frac{p}{Z}-\frac{V}{H}\left(1-\frac{a}{H}\right)\right) + \left(\frac{p}{a}-1-\frac{V}{H}\right)\left(\frac{p}{Z}-1\right) + \frac{D}{H}\right)}\right)$$

Once *p* and *k* are determined from this curve fit, they can sometimes be optimized, by first estimating *Q* using *p* and *k* in:

🟩 $$\frac{Q}{k} = \frac{D}{(p-Z)\left(\frac{H}{Z}-1\right)-k} - \frac{1}{\frac{H}{Z}-1}$$

then doing a 3-parameter curve fit to find *p, k*, and *Q* using the following, along with the same data set {(*V, a*)} used initially to get *p* and *k*:

🟦 $$V = \frac{H}{a}\left(p - a - \frac{k}{\frac{H}{a}-1} - \frac{D}{\left(\frac{H}{a}-1\right)\frac{Q}{k}+1}\right) = H\left(\frac{p}{a} - 1 - \frac{k}{H-a} - \frac{D}{(H-a)\,\frac{Q}{k}+a}\right) = H\left(\frac{p}{a} - 1 - \frac{k}{H-a} - \frac{D}{\frac{Q}{k}H - \left(\frac{Q}{k}-1\right)a}\right)$$

[As an aside, it is worth noting that when *Q*/*k* >> 1, this is closely approximated by:

🟪 $$V = H\left(\frac{p}{a} - 1 - \frac{k}{H-a} - \frac{D}{\frac{Q}{k}H - \left(\frac{Q}{k}\right)a}\right) = H\left(\frac{p}{a} - 1 - \frac{k}{H-a} - \frac{D\frac{k}{Q}}{H-a}\right) = H\left(\frac{p}{a} - 1 - \frac{k\left(1+\frac{D}{Q}\right)}{H-a}\right)$$

So, when *Q*/*k* >> 1, adding non-specific competitor has the same effect on the competition curve as increasing *k* by the factor 1 + *D*/*Q*.]

However, the 3-parameter curve fit may fail to converge to accurate values with imprecise data.
In fact, from trial runs that we have done, unless the data are very precise, especially the value of $Z$,
the initial curve fitting to find $p$ and $k$ will likely fail.

A much more robust method involves taking two sets of data, one data set using increasing amounts of
cold competitor oligo that is the same sequence as the probe DNA ($V$),
and the other data set using increasing amounts of extra added non-specific competitor DNA ($W$; the total concentration is now $D + W$),
and measuring the amount of labeled complex formed ($a$ or $\mathring{a}$, respectively).
If we call these data sets $\{(V, a)\}$ and $\{(W, \mathring{a})\}$, respectively, then we can combine the two data sets to generate the data set
$\{(W, \mathring{a}, V, a)\}$, assuming that there are the same number of data points in each.

🟥 $$W = \left( p - \mathring{a} - \frac{\left(\frac{p}{a} - 1 - \frac{V}{H}\right)(H - a) - \dfrac{D}{\frac{Q}{k} + \dfrac{1}{\frac{H}{a} - 1}}}{\frac{H}{\mathring{a}} - 1} \right) \left( \left(\frac{H}{\mathring{a}} - 1\right) \frac{Q}{k} + 1 \right) - D$$

Curve fitting with this expression to find $p$ and $Q/k$ as parameters using such data is quite robust
(as long as the values of $a$ and $\mathring{a}$ are not all identical in the two data sets, in which case the formula collapses, eliminating $p$).

It works best to generate the data set $\{(W, \mathring{a}, V, a)\}$ by combining the lower values for $\mathring{a}$ with the higher values for $a$.
Alternatively, we can generate a larger data set by combining the two sets $\{(V, a)\}$ and $\{(W, \mathring{a})\}$ combinatorially
(each point of one set combined with all the points of the other set). In the trials we have done, this does not generally
give better results than using the smaller data set, but it may do so in some circumstances.

Once $Q/k$ is determined, we can use the above equation to find $p$ and $k$ by curve fitting, or, alternatively,
use the value of $p$ found in the first step and do single-parameter curve fitting to find $k$, using the same data set $\{(V, a)\}$:

🟥 $$V = \left( p - a - \frac{k}{\frac{H}{a} - 1} - \frac{D}{\left(\frac{H}{a} - 1\right) \frac{Q}{k} + 1} \right) \frac{H}{a}$$

Once the two binary complex Kd's and the two active protein concentrations are determined by one of the above methods (either without or with non-specific competitor), *n* can be found by curve fitting using the formula for *p* as a function of *A* given above under "Fig. 2", which is:

$$p = A + \frac{A}{n}\left(\frac{H+L-q+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left(q-A+\frac{A}{n}\right)}\right) + \frac{kA}{n}\left(\frac{H+L+q-2A+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left((H-A)\,(q-A)-\frac{LA}{n}\right)}\right)$$

A set of data points {(*A, p*)} (in the absence of competitor DNA) can be used in the above with the constants *H, q, k*, and *L* known, and *n* as the parameter to be determined.
An initial estimate for *n* can be obtained by plugging values from one or two data points into the above equation, and finding the corresponding *n* as the variable, using a graphing calculator or other means (*n* cannot be solved for explicitly due to the complexity of the expression).

## Section 5 of Peacock and Jaynes [2]:

For the case where the single-protein dissociation constant (*L*) and protein concentration (*q*) are known for only one protein, finding first *n*, from a set of data points {(*b, A*)}, where

*b* = [single-protein complex] containing the protein with known Kd (= *L*) and total concentration (*q*), and
*A* = [ternary complex], both of which vary as the other protein concentration is varied
(*p* varies but is not yet known, and has been eliminated from the system of equations),

use the following formulae, derived in Section 5B of Peacock and Jaynes [2]:

$$A = \frac{1}{2}\left(H+q-2b+\frac{L}{n}-\sqrt{\left(H+q-2b+\frac{L}{n}\right)^2-4\left((H-b)\,(q-b)-Lb\right)}\right)$$

All constants are known except *n*, which is determined as the parameter in curve fitting.
An initial estimate for *n* can be obtained from any data point (or several) using the equation:

$$n = \frac{LA}{(H-A-b)\,(q-A-b)-Lb}$$

Another way to obtain a minimum estimate for *n* comes from the simple relationship between *n* and the various forms of oligo, with *A*, *h*, and *b* measurable:

$$n = \frac{Ah}{ab}$$

Assuming that *a* is not detectable in the experiment, we estimate what the minimum detectable [oligo] is. This gives us an upper limit for *a*. This, inserted into the above expression, gives us a minimum estimate for *n*.

After *n* is determined, the unknowns *k* and *p* can be found from the same data set.
Starting with the general formula:

$$p = \frac{kA}{nb} + \frac{H + (n-1)\,b}{1 + \frac{nb}{A}}$$

we keep track of the set of dilution factors {*Δ*} used (relative to the highest amount of *p* used, which should ideally be enough to "chase" most of *H* into the ternary complex *A*), to get the data set {(*b, A*)}.
This gives us the data set {(*Δ, b, A*)}, which is used in 2-parameter curve fitting with the formula:

$$\Delta = \frac{A}{\mathrm{p0}}\left(\frac{k}{nb} + \frac{H + b\,(n-1)}{A + nb}\right)$$

where p0 is the highest amount of *p* used in the experiment.
*k* and p0 are determined as parameters, knowing *H* and *n*.

For an initial estimate for *p*, we can use the measured concentration (which should represent a maximum estimate).
For an initial estimate for *k*, we can use that estimate for *p*, and then *k* from one or more points (*b, A*) using the rearranged version of the above:

$$\frac{nb}{A}\left(p - \frac{H + (n-1)\,b}{1 + \frac{nb}{A}}\right) = k = p\,\frac{nb}{A} - \frac{H + (n-1)\,b}{\frac{A}{nb} + 1}$$

Once *n*, *k*, and p0 are thus obtained, their values can sometimes be refined further by doing a 3-parameter curve fit to the above equation for *Δ*, using the values obtained above as starting values, and the same data set {(*Δ*, *b*, *A*)}.
This refinement, if successful, also gives an indication of the overall goodness of fit.
However, in practice (based on our trials using initially precise data sets rounded to either 2 or 3 significant figures), the 3-parameter curve fit either does not converge,
or converges to values that are much less precise than those obtained from the 2-step method above.

# Mathematical toolkit for quantitative analysis of cooperative binding of two or more ligands to a substrate

Jacob Peacock and James B. Jaynes

Dept. of Biochemistry and Molecular Biology, Thomas Jefferson University,
Philadelphia PA 19107 United States of America

james.jaynes@jefferson.edu

## Abstract

We derive mathematical expressions for quantitative analysis of cooperative binding covering the following cases: 1) a single ligand binds to either two non-equivalent sites, or an arbitrary number of equivalent sites, on a substrate (Fig. 1), and 2) two different ligands bind distinct sites on a substrate (Figs. 2, 3). We show how to analyze "competition experiments" using non-linear regression, where a ligand binds to a single site on a labeled substrate in the presence of increasing amounts of identical but unlabeled competitor substrate, to simultaneously determine the Kd and active ligand concentration (Fig. 4A). We compare the performance of this competition method with the commonly used saturation binding method (Fig. 4B). We also provide methods to analyze such experiments that include a second competitor substrate with non-specific binding sites. We show how to build on results from single-ligand competition experiments to fully characterize cooperative binding in systems with two distinct ligands and binding sites (Fig. 4C and Section 5). We generalize the methodology to more than two cooperating ligands, such as an array of DNA binding proteins (Fig. 6). See Peacock and Jaynes [1] for discussion of the various ways these tools can be used, and results using them.

- Visualize characteristics and limitations of Hill plots applied to more realistic binding models than those described by the Hill equation, which implies multiple simultaneous ligand binding.
- Efficiently find individual ligand-substrate Kd's and active ligand concentrations (Fig. 4A), and from this determine the cooperativity factor (Fig. 4C).
- Connect ternary complex formation with varying total concentration of one ligand, and find the cooperativity factor, even if only one individual Kd can be determined by the method of Fig. 4 (Section 5).

## Keywords

cooperative ligand binding, cooperative DNA binding, quantifying cooperativity, cooperative complex formation, Hill plots, yeast a1-alpha-2, Engrailed, Extradenticle, Homothorax, oligo competition, curve fitting, non-linear least-squares regression, finding Kd's, finding dissociation constants, EMSA complex analysis, cooperativity methods

### Specifications Table

| | |
|---|---|
| Subject area | *Biochemistry, Genetics, and Molecular Biology* |
| More specific subject area | *Cooperative ligand binding to multiple sites on a substrate (such as DNA)* |
| Method name | *Quantitation of cooperativity among multiple ligands for a substrate* |
| Name and reference of original method | *Does not modify a specific original method, but does reference and compare to several general methodologies (see Jaynes & Peacock [1])* |
| Resource availability | *N/A* |

# Method Details

**<u>Fig. 3.</u>** S3 – p. 1

Derives expressions:

- for graphing concentrations of ternary complex, free ligand, substrate, competitor, and other complexes as a function of added total [unlabeled competitor substrate] using simple equation-graphing software, in a system where two distinct ligands bind cooperatively to two distinct binding sites on a substrate (in **A**)

- for total occupancy of such a substrate by each of two ligands, as a function of total [ligand-1] (without unlabeled competitor) (in **B**)

- for the saturation value of each total [bound ligand] as the total [ligand-1] increases (in **B**)

**<u>details for A:</u>**

Derivation of an expression for total [specific unlabeled competitor substrate] (= *U*) as a function of [labeled ternary complex] (= *A*) and [free ligand1] (= *f*). S3 – p. 1-5

An expression for *f* as a function of *A* is given. (The original expression is quartic in *f*, and so it uses the general solution for a quartic equation.) This can be substituted into the expression for *U* (as a function of *A* and *f*) to get *U* as a function of *A* for graphing, using simple software. S3 – p. 5

The original expression involving *f* and *A* is cubic in *A*, and can be solved for *A* as a function of *f* using the general cubic solution, which is simpler and easier for software packages to handle than the quartic solution. So, this solution is given, along with expressions for the other variables in the system as a function of *A* and *f*. S3 – p. 7-8

Together, these allow graphing of each of the varying forms involved in the experiment as a function of any other one of these forms (most usefully, *A* as a function of *U*), using f as a parameter along with the cubic solution for *A* as a function of *f*. For example, we can graph *y* vs. *x* where $y = A(f)$ and $x = U[A(f), f]$. S3 – p. 7-8

Peacock and Jaynes, 2017

Fig. 1A. One ligand binding to two non-equivalent sites. Expressions are given for
1) fractional occupancy as a function of [free ligand], for a given Kd and cooperativity factor,
2) the condition where the binding curve is identical to that for two equivalent sites without cooperativity, and
3) relationships between macroscopic and microscopic Kd's.

Refer to Fig. 1 of Peacock and Jaynes [1] for illustrations of how these tools can be used.

CONTENTS:

Expressions for:

**1)** fractional occupancy as a function of [free ligand], Kd's for each site, and cooperativity factor;
**2)** condition when the binding curve is identical to that for two equivalent sites without cooperativity.
**3)** Relationships between macroscopic and microscopic Kd's.

Definitions of variables:

$h$ = [free probe DNA, "hot" probe]
$H$ = total ["hot" DNA]
$f$ = [free protein]
$a$ = [protein – site 1 complex] = [($fh$)]
$b$ = [site 2 – protein complex] = [($hf$)]
$A$ = [ternary complex] = [($fhf$)]

Dissociation (equilibrium) constants, including cooperativity factor $n$:

$k$ = dissociation constant of ($fh$), site 1 binary complex
$L$ = dissociation constant of ($hf$), site 2 binary complex
$k/n$ = dissociation constant of protein from site 1 of ternary complex ($fhf$)
$L/n$ = dissociation constant of protein from site 2 of ternary complex ($fhf$)

Equations governing equilibrium concentrations:

equation 1:

☒ $L = \frac{fh}{b}$

equation 2:

☒ $k = \frac{fh}{a}$

equation 3:

☒ $H = h + a + b + A$

equation 4:

☒ $\frac{L}{n} = \frac{fa}{A}$

The binding polynomial *P* is *H/h* (as in Freire, et al., 2009 [2]):

☒ $P = \frac{H}{h} = 1 + \frac{a}{h} + \frac{b}{h} + \frac{A}{h} = 1 + \frac{a}{h} + \frac{naf}{hL} + \frac{b}{h} = 1 + \left(1 + \frac{nf}{L}\right)\frac{a}{h} + \frac{f}{L} = 1 + \left(1 + \frac{nf}{L}\right)\frac{f}{k} + \frac{f}{L}$

☒ $P = 1 + \frac{f}{k} + \frac{f}{L} + \frac{nf^2}{kL}$

This form shows that *k* and *L* are 1/(microscopic association constants) while *n* is kappa, the "cooperativity constant" of Freire, et al., 2009 [2].

**1)** To get *a* in terms of *f* and constants, get *A*, *h*, and *b* in terms of *f* and constants, and substitute into eq'n 3, which gives:

☒ $a = \frac{HLf}{kL + Lf + kf + nf^2}$

Now, since *b* = *k/L * a* (above, from eq'ns 1 and 2):

☒ $$b = \frac{Hkf}{kL + Lf + kf + nf^2}$$

and since *A* = *nf/L * a* (above, from eq'n 4):

☒ $$A = \frac{Hnf^2}{kL + Lf + kf + nf^2}$$

This allows us to express fractional occupancy, ø, in terms of *f* and constants:

☒ $$\phi = \frac{b + a + 2A}{2H} = \frac{(k + L + 2nf)\,f}{2\,(kL + Lf + kf + nf^2)}$$

For a Hill plot, we need this over (1 – fractional occupancy) which is:

☒ $$1 - \phi = \frac{2kL + Lf + kf}{2\,(kL + Lf + kf + nf^2)}$$

and ø / (1 – ø) is:

$$\frac{\phi}{1 - \phi} = \frac{(k + L + 2nf)\,f}{2kL + \,(k + L)\,f} = \frac{(k + L)\,\left(1 + \frac{2n}{k + L}f\right) f}{2kL\left(1 + \frac{k + L}{2kL}f\right)}$$

For a Hill plot, we graph the natural logarithm (ln) of this vs. ln *f*.
Rather than using a double ln scale, it is often easier to substitute *x* = ln *f*, which means that

☒ $$f = e^x$$

and graph the ln of the above, with this substitution, as a function of *x*.

**2)** Next, specialize this to the case of two equivalent binding sites, for which *L* = *k*:

$$\frac{\phi}{1-\phi} = \frac{(k+nf)\,f}{kk+kf} = \frac{(k+nf)\,f}{k\,(k+f)}$$

Then graph it as a Hill plot (below).

Aside:
(Note that when *n* = 1, this is just *f* / *k*, the same as for a single site, where *f* / *k* = *a* / *h* = (*a*/*H*) / (*1* – *a*/*H*).)

When does this without cooperativity (*n* = 1) look the same as two non-equivalent sites with cooperativity?
When 2*n* / (*k*+*L*) = (*k*+*L*) / 2*kL*, because then ø / (1 – ø) becomes a constant times *f* (specifically, (*k*+*L*)/2*kL* = 1/*k'*),
as it is without cooperativity for equivalent sites (it then = *f* / *k*, as stated above, but call it *k'* to distinguish it):

$$\frac{f}{k'} = \left(\frac{k+L+2nf}{2kL+\,(k+L)\,f}\right)f = \frac{(k+L)\,\left(1+\dfrac{2n}{k+L}f\right)f}{2kL\left(1+\dfrac{k+L}{2kL}f\right)}$$

When 2*n* / (*k*+*L*) = (*k*+*L*) / 2*kL*:

$$\frac{f}{k'} = \frac{(k+L)\,f}{2kL}$$

or,

🟪 $$k' = \frac{2kL}{k+L}$$

and from the assumption that,

☒ $$\frac{2n}{k+L} = \frac{k+L}{2kL}$$

🟪 $$n = \frac{(k+L)^2}{4kL}$$

If *k* and *L* differ by a factor of *R*:

⊠ $$n = \frac{(k + Rk)^2}{4kRk} = \frac{(1 + R)^2}{4R} = \left( \frac{\sqrt{R} + \sqrt{\frac{1}{R}}}{2} \right)^2$$

and if *R* >> 1, this is approximated by *R*/4. Specifically, the value of *n* for which the Hill plot is linear is always > *R*/4, and approaches *R*/4 (if *R* is defined so that *R* > 1; if *R* < 1, this is 1/4*R* instead) as *R* goes to infinity.

This means that as long as *n* < *R*/4, positive cooperativity cannot produce a max. Hill plot slope that is > 1.

The Hill plot is linear when *R* = 100 and *n* is exactly:

⊠ $$\frac{\left(\sqrt{R} + \sqrt{\frac{1}{R}}\right)^2}{4} = \frac{\left(10 + \frac{1}{10}\right)^2}{4} = \frac{102.01}{4} = 25.5025$$

or when *R* = 4 and *n* is:

⊠ $$\frac{\left(2 + \frac{1}{2}\right)^2}{4} = \frac{6.25}{4} = 1.5625$$

**3)** It is interesting to relate the above to the macroscopic dissociation constants:
For the singly bound complex, the macroscopic dissociation constant is:

⊠ $$\frac{fh}{a + b} = \frac{1}{\frac{a}{fh} + \frac{b}{fh}} = \frac{1}{\frac{1}{k} + \frac{1}{L}} = \frac{kL}{k + L}$$

while for the ternary complex, it is:

⊠ $$\frac{(a + b) f}{A} = \frac{L}{n} + \frac{k}{n} = \frac{k + L}{n}$$

The ratio of these is:

⊠ $$\frac{nkL}{(k+L)^2}$$

We saw above that the Hill plot slope goes above 1 when *n* >

⊠ $$\frac{(k+L)^2}{4kL}$$

When this is the case, then the ratio of the macroscopic dissociation constants is >

□ $$\frac{\frac{(k+L)^2}{4kL}kL}{(k+L)^2} = \frac{1}{4}$$

This is the ratio of the association (or dissociation) constants required in order for two non-equivalent sites to always appear cooperative, as stated in the introductory summary in Bardsley, 1977 [3].
(Rho is defined as 4 * this ratio in equation 5.19 of Freire, et al., 2009 [2], and it must exceed 1 in order to get unambiguously cooperative behavior, which is equivalent to the ratio exceeding 1/4).

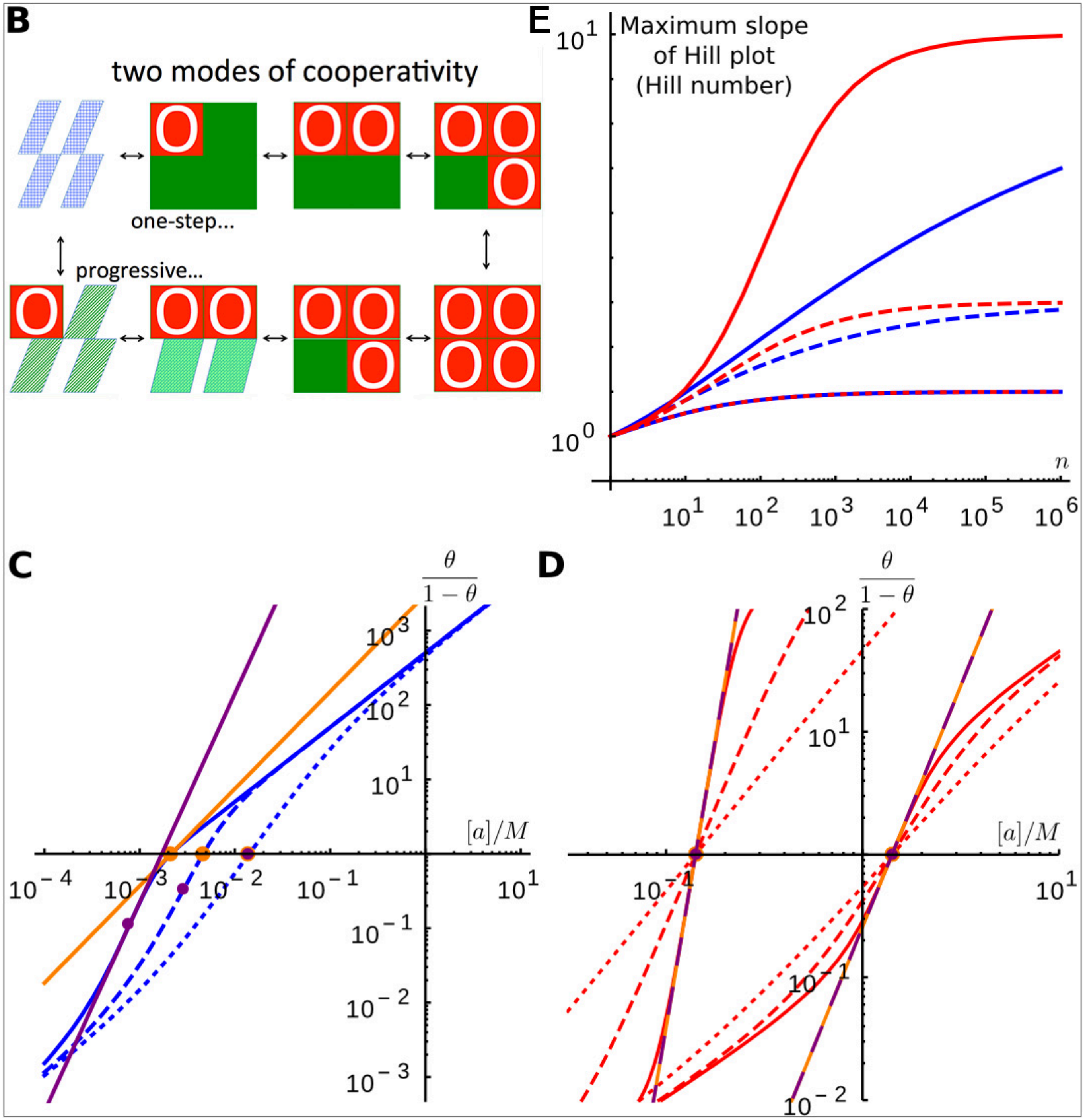


**Fig. 1B-E. Two simple modes of positive cooperativity and corresponding Hill plots.** **B:** models, two modes of cooperativity. One-step, upper path: When any one site is bound (of the Z equivalent sites), the Kd for all subsequent binding events is decreased to the initial Kd / *n* (*n* > 1). Progressive, lower path: the Kd decreases multiplicatively in (Z – 1) equal steps from the initial binding event's Kd down to 1/*n* times that starting value. **C:** Hill plots for one-step cooperativity, showing how the shape of the plot changes with the number of sites (2, 4 and 10 sites, fine-dashed, coarse-dashed, and solid, respectively, with *n* = 50, initial Kd = 0.1), where the maximum slope occurs (purple dots and tangent line), and what happens if slopes are measured at 50% occupancy instead (orange dots and tangent line). Note that the points of maximum slope shift to lower occupancies for increasing numbers of sites. Thus, the maximum slopes are 1.75 for 2 sites, 2.37 for 4 sites, and 2.81 for 10 sites, while the slope at 50% occupancy for 10 sites is only 1.31, less than it is for either 2 or 4 sites. See Fig. 1F,G for expressions used to generate these graphs, and their derivations. **D:** Hill plots for progressive cooperativity, showing how shapes change with the number of sites (2, 4 and 10 sites, fine-dashed, coarse-dashed, and solid curves, respectively, with initial Kd = 10) and with *n* (maximum cooperativity); the two sets of curves show two different values of *n* (5000 on the left, and 50 on the right). The

bicolored line is tangent to the 10-site curve at the point of maximum slope, which occurs at 50% occupancy for all curves. Note that the slopes at 50% occupancy for *n* = 5000 (1.97, 3.80, and 9.37, for 2, 4, and 10 sites, respectively) approximate the number of sites, while for *n* = 50, this it true only for the 2-site curve (slopes are 1.75, 2.56, and 3.93, respectively). See Table 1 for more examples; see Fig. 1H,I for derivations of expressions used here. **E:** graphs of maximum slopes of Hill plots as a function of cooperativity factor (*n*), for one-step cooperativity (blue curves) and progressive cooperativity (red curves), and for 2, 4 and 10 sites, (dotted, dashed, and solid curves, respectively). Note that the maximum slopes approach the number of sites as *n* increases, but that this approach is slower as the number of sites increases. See Fig. 1G,I for a derivation of expressions used here.

In the case of one-step cooperativity (illustrated in Fig. 1B, upper path), occupancy of any single site causes a decrease in the Kd of all other sites by a factor of *n* (assuming positive cooperativity, i.e., *n* > 1). As expected, the maximum slope (at very high cooperativity) of a Hill plot approaches the number of sites, Z (Table 1). However, surprisingly, this maximum slope occurs when the fractional occupancy is 1/Z (not ½, as might be expected, Fig. 1G). Hill plots for one-step cooperativity with 2, 4, and 10 equivalent sites are shown in Fig. 1C. It is noteworthy that in all cases, the slope at each extreme of ligand concentration approaches 1, not Z.

With one-step cooperativity, measuring the slope at 50% occupancy leads to an underestimate of the number of sites (in fact, the slope at this point can decrease with increasing number of binding sites, depending on *n*). Furthermore, very high cooperativity is required for the maximum slope to approach Z. For example, with 50-fold cooperativity, the maximum slope is between 2 and 3 for 3-10 binding sites (Table 1). Even with 5000-fold cooperativity, the slope at 50% occupancy peaks at 3.81 for 7 sites, and the maximum slope is only 5.1 for 10 sites (Table 1).

To round out the consideration of Hill plots as a way to quantify cooperativity, we consider the case of progressive cooperativity, where each additional bound ligand changes the affinity for subsequent ligands in equal increments. Here, the maximum slope occurs at 50% occupancy for any number of sites (Figs. 1D, 1I), and approaches Z as *n* increases (Table 1, Figs. 1E, 1I). However, with 50-fold cooperativity, the maximum slope is ~3 for 5 sites, and ~4 for 10 sites (Table 1).

One of the under-appreciated aspects of the Hill formalism is just how unrealistic the Hill equation can be relative to more plausible biophysical models. A Hill plot of the basic Hill equation is simply a straight line of slope Z. In particular, the slope does not approach 1 at very low and high occupancies. This gives the impression that to determine the number of sites, data can be obtained anywhere in the occupancy range, when in fact, precise data must be obtained bracketing the point of maximum slope. Furthermore, because this point can shift as the number of sites changes, it is not sufficient to obtain data near 50% occupancy. The difficulty of accurately estimating the number of cooperating sites from Hill plots has been noted previously, even for the classic case of hemoglobin. Our analysis shows some of the general reasons behind such observations.

Peacock and Jaynes, 2017

Fig. 1F-I. One ligand binding cooperatively to an arbitrary number of equivalent sites.

F: schematic of the binding equilibrium for one-step cooperativity, where cooperativity becomes maximum (max.) as soon as one ligand is bound to any site, and derivation of an expression for the concentration of the i^th complex $a_i$, which contains i ligand molecules, where:

Z = the total number of sites on the substrate
f = [free ligand],
r = forward rate constant for binding of a ligand molecule to the complex (or to the unliganded substrate h, free "hot" probe),
k = dissociation rate constant for the singly bound complex,
k/n = dissociation rate constant for each ligand molecule except the first from the complex;
to simplify some of the expressions (here and later on), s is defined as:
s = fr/k; free ligand concentration divided by the equilibrium dissociation constant without cooperativity (k/r).

G: one ligand binding cooperatively to an arbitrary number (*Z*) of equivalent sites (one-step cooperativity); contains expressions for fractional occupancy and Hill plot formulae, including conditions for points of maximum slope, and slopes of Hill plots at their points of max. slope (which is where the fractional occupancy = 1/*Z*).

H: schematic of the binding equilibrium for progressive cooperativity, where cooperativity increases by an equal factor as each ligand is added, and derivation of an expression for the concentration of the i^th complex $a_i$, which contains i ligand molecules, where:

Z = the total number of sites on the substrate
f = [free ligand],
r = forward rate constant for binding of a ligand molecule to the complex (or to the unliganded substrate h, free "hot" probe),
k = dissociation rate constant for the singly bound complex,
$k / n^{(i-1)/(Z-1)}$ = dissociation rate constant for a ligand molecule from the i^th complex
k / n = dissociation rate constant for a ligand molecule from the fully liganded (the Z^th) complex.
to simplify some of the expressions, s is defined as:
s = fr/k; free ligand concentration divided by the equilibrium dissociation constant without cooperativity (k/r).

I: one ligand binding cooperatively to an arbitrary number of equivalent sites (progressive cooperativity); contains expressions for fractional occupancy and Hill plot formulae, including the condition for the point of maximum slope, and slopes of Hill plots at their point of max. slope (which is always where the fractional occupancy = ½).

F: one-step cooperativity

$$Zf + h \underset{k}{\overset{r}{\rightleftarrows}} (Z-1)f + a_1 \underset{k/n}{\overset{r}{\rightleftarrows}} (Z-2)f + a_2 \underset{k/n}{\overset{r}{\rightleftarrows}} \cdots \underset{k/n}{\overset{r}{\rightleftarrows}} (Z-i)f + a_i \underset{k/n}{\overset{r}{\rightleftarrows}} \cdots \underset{k/n}{\overset{r}{\rightleftarrows}} a_Z$$

$Zfhr = a_1k$

$(Z-1)fa_1r = 2a_2k/n$

$a_2 = n(Z-1)fra_1 / 2k$

$a_2 = n(Z-1)fr(Zfhr/k) / 2k$

$a_2 = nZ(Z-1)(f^2r^2/k^2)h / 2$

$(Z-2)fa_2r = 3a_3k/n$

$a_3 = n(Z-2)fra_2 / 3k$

$(Z-i)fa_ir = (i+1)a_{i+1}k/n$

$a_{i+1} = n(Z-i)fra_i / (i+1)k$

$a_3 = n(Z-2)fr \{nZ(Z-1)(f^2r^2/k^2)h / 2\} / 3k$

$a_3 = n^2Z(Z-1)(Z-2)(f^3r^3/k^3)h / (3)2$

$$a_i = \frac{n^{i-1}(fr/k)^i\, h\, Z!}{(Z-i)!\ i!} = \frac{h}{n}\,\frac{(nfr/k)^i\, Z!}{(Z-i)!\ i!} = \frac{h}{n}\,\frac{(ns)^i\, Z!}{(Z-i)!\ i!}$$

G

Starting with the formula for the concentration of each complex containing *i* bound ligand molecules for 1-step cooperativity, derived in F above:

🟥 $$a_i = \frac{\left(\frac{h}{n}\right)(ns)^i Z!}{(Z-i)!i!}$$

The fractional occupancy is the total number of occupied sites divided by the total number of sites; therefore:

🟪 $$\theta = \frac{\sum\limits_{i=1}^{Z} ia_i}{ZH}$$

where

🟦 $$\sum_{i=1}^{Z} ia_i = \sum_{i=1}^{Z} \frac{i\left(\frac{h}{n}\right)(ns)^i Z!}{(Z-i)!i!} = \sum_{i=1}^{Z} \frac{hs(ns)^{i-1}Z(Z-1)!}{(Z-i)!(i-1)!} = Zhs\sum_{i=1}^{Z} \frac{(ns)^{i-1}(Z-1)!}{(Z-i)!(i-1)!}$$

This can be converted to an analytic form that is easier to work with by first changing the summation limits to start at 0; when we do this, *i*-1 becomes *i*:

$$\sum_{i=1}^{Z} i a_i = Zhs \sum_{i=0}^{Z-1} \frac{(ns)^i (Z-1)!}{(Z-1-i)!i!}$$

This summation is a standard form that can be recognized as the binomial expansion of

$$(1+ns)^{Z-1} = \sum_{i=0}^{Z-1} \frac{(ns)^i (Z-1)!}{(Z-1-i)!i!}$$

which makes the original summation:

$$\sum_{i=1}^{Z} i a_i = Zhs(1+ns)^{Z-1}$$

This, in turn, makes the fractional occupancy:

$$\theta = \frac{Zhs(1+ns)^{Z-1}}{ZH} = \frac{s(1+ns)^{Z-1}}{\frac{H}{h}}$$

This can be further modified to eliminate *h* by noticing that:

$$H = h + \sum_{i=1}^{Z} a_i$$

and then converting this summation to an analytic form by noticing that it contains a standard binomial expansion, except that it is missing the first term:

$$\sum_{i=1}^{Z} a_i = \sum_{i=1}^{Z} \frac{\left(\frac{h}{n}\right)(ns)^i Z!}{(Z-i)!i!} = \left(\frac{h}{n}\right) \sum_{i=1}^{Z} \frac{(ns)^i Z!}{(Z-i)!i!}$$

$$\sum_{i=1}^{Z} a_i = \left(\frac{h}{n}\right) \left( \sum_{i=0}^{Z} \frac{(ns)^i Z!}{(Z-i)!i!} - 1 \right)$$

and using the analytic form that this summation corresponds to:

$$\sum_{i=0}^{Z} \frac{(ns)^{i} Z!}{(Z-i)!i!} = (1+ns)^{Z}$$

This gives:

$$\sum_{i=1}^{Z} a_{i} = \frac{h}{n}\left(\left(1+ns\right)^{Z}-1\right)$$

Now we can see that

$$H = h + \sum_{i=1}^{Z} a_{i} = h + \frac{h}{n}\left(\left(1+ns\right)^{Z}-1\right)$$

so that

$$\frac{H}{h} = 1 + \frac{1}{n}\left(\left(1+ns\right)^{Z}-1\right)$$

which then gives us an analytic form for the fractional occupancy in terms of only *n*, *s*, and *Z*:

$$\theta = \frac{s\left(1+ns\right)^{Z-1}}{\frac{H}{h}} = \frac{s\left(1+ns\right)^{Z-1}}{1+\frac{1}{n}\left(\left(1+ns\right)^{Z}-1\right)} = \frac{ns\left(1+ns\right)^{Z-1}}{n+\left(1+ns\right)^{Z}-1}$$

For a Hill plot, we need

$$\frac{\theta}{1-\theta}$$

so we need

$$1-\theta = 1 - \frac{ns\left(1+ns\right)^{Z-1}}{n+\left(1+ns\right)^{Z}-1} = \frac{n+\left(1+ns\right)^{Z}-1-ns\left(1+ns\right)^{Z-1}}{n+\left(1+ns\right)^{Z}-1}$$

which simplifies to:

$$1-\theta = \frac{n-1+(1+ns)^{Z-1}}{n+(1+ns)^{Z}-1}$$

which gives

$$\frac{\theta}{1-\theta} = \frac{ns\,(1+ns)^{Z-1}}{n-1+(1+ns)^{Z-1}} = \frac{ns}{(n-1)\,(1+ns)^{1-Z}+1}$$

A Hill plot is the natural logarithm (ln) of this as a function of ln(*f*).
Since $s = fr/k$,

$$\frac{\theta}{1-\theta} = \frac{\frac{nfr}{k}}{(n-1)\left(1+\frac{nfr}{k}\right)^{1-Z}+1}$$

Taking the ln of both sides,

$$\ln\left(\frac{\theta}{1-\theta}\right) = \ln\frac{\frac{nfr}{k}}{(n-1)\left(1+\frac{nfr}{k}\right)^{1-Z}+1}$$

One way to generate a ln plot is to substitute *x* for ln(*f*), then graph as a function of *x*; *x* = ln(*f*) means that:

$$f = e^{x}$$

So, using these substitutions,

$$\ln\left(\frac{\theta}{1-\theta}\right) = \ln f + \ln\frac{nr}{k} - \ln\left((n-1)\left(1+\frac{nfr}{k}\right)^{1-Z}+1\right)$$

The Hill plot is then a plot of the following *y* vs. *x*:

$$y = x + \ln\frac{nr}{k} - \ln\left((n-1)\left(1+\frac{nr}{k}e^{x}\right)^{1-Z}+1\right)$$

In order to determine and plot the slope of this curve, we need the derivative of the above *y* with respect to *x* (which is the slope of the Hill plot). This is:

$$1 + \frac{\left((n-1)\,(Z-1)\left(1+\frac{nr}{k}e^{x}\right)^{-Z}\left(\frac{nr}{k}\right)\right)e^{x}}{(n-1)\left(1+\frac{nr}{k}e^{x}\right)^{1-Z}+1} = 1 + \frac{\left((n-1)\,(Z-1)\left(\frac{nr}{k}\right)\right)e^{x}}{(n-1)\left(1+\frac{nr}{k}e^{x}\right)+\left(1+\frac{nr}{k}e^{x}\right)^{Z}}$$

This is used below to find a simple expression for the max. slopes of Hill plots.
However, to do this, the point where the max. slope occurs must first be found.

To find where this is maximum, we need to find where its derivative = 0.
The derivative of the slope with respect to *x*, which = ln(*f*), is *N* / *D* where:

$$N = \frac{(n-1)\,(Z-1)\,nr}{k}e^{x}\left((n-1)+\left(1+\frac{nr}{k}e^{x}\right)^{Z-1}\left(1+\frac{nr}{k}e^{x}-\frac{Znr}{k}e^{x}\right)\right)$$

and

$$D = \left((n-1)\left(1+\frac{nr}{k}e^{x}\right)+\left(1+\frac{nr}{k}e^{x}\right)^{Z}\right)^{2}$$

*D* is not relevant for finding where the slope is max., except for ruling out where it = 0 as a possible solution.

This maximum occurs where $N = 0$, which occurs where the second factor in $N = 0$. So the condition for maximum slope is:

$$n - 1 = \left(1 + \frac{nr}{k} e^x\right)^{Z-1} \left((Z-1)\left(\frac{nr}{k}\right) e^x - 1\right)$$

To simplify the expressions, as above, use $s = fr/k$, the free ligand concentration divided by the equilibrium dissociation constant without cooperativity, $k/r$.

$$s = \frac{r}{k} e^x$$

This makes the condition for max. slope:

$$n - 1 = (1 + ns)^{Z-1} ((Z-1)\, ns - 1)$$

or

$$(n-1)\,(1 + ns)^{1-Z} + 1 = (Z-1)\, ns$$

This expression can be solved explicitly for the variable $s$ or $x$ only for $Z = 2, 3$, or $4$ (see below). However, it can be solved graphically, and the value thus obtained used to get the max. slope using the formula above.

This was done to draw lines of slope equal to the max. slope, at the points of max. slope, for the Hill plots in Fig. 1C. It was also used to obtain the values for max. slopes shown in Table 1 for one-step cooperativity.

Although finding the point of max. slope is somewhat complicated, it can be seen to have a simple relationship to the fractional occupancy, as follows.
From above:

$$\frac{\theta}{1-\theta} = \frac{ns}{(n-1)\,(1+ns)^{1-Z} + 1}$$

At the point of max. slope, this is:

$$\frac{\theta}{1-\theta} = \frac{ns}{(Z-1)\,ns} = \frac{1}{Z-1}$$

This is used to draw horizontal lines that cross the Hill plots at their points of maximum slope in Fig. 1C. Rearranging this to find the fractional occupancy at the point of maximum slope:

$$\theta\,(Z-1) = 1-\theta$$

$$\theta Z = 1$$

**So, the fractional occupancy at the point of max. slope is:**

$$\theta = \frac{1}{Z}$$

What is the slope at this point, where the slope is maximum?
The slope is, from above, incorporating the variable s:

$$1 + \frac{(n-1)\,(Z-1)\,ns}{(n-1)\,(1+ns) + (1+ns)^{Z}}$$

The condition for max. slope tells us that:

$$(n-1)\,(1+ns) + (1+ns)^{Z} = (Z-1)\,ns\,(1+ns)^{Z}$$

Substituting this for the denominator in the expression for the max. slope gives:

$$1 + \frac{(n-1)\,(Z-1)\,ns}{(Z-1)\,ns\,(1+ns)^{Z}} = 1 + \frac{(n-1)}{(1+ns)^{Z}}$$

Another form of the condition for max. slope is:

$$\frac{(n-1)}{(1+ns)^{Z}} = \frac{ns\,(Z-1)-1}{1+ns}$$

so that the expression for max. slope can also be written:

$$1+\frac{ns\,(Z-1)\;-1}{1+ns} \;=\; \frac{1+ns+Zns-ns-1}{1+ns}$$

which simplifies to the following;
**maximum slope of a Hill plot with 1-step cooperativity *n*:**

$$\frac{Zns}{1+ns} \;=\; \frac{Z}{1+\dfrac{1}{ns}}$$

where

$$s \;=\; \frac{fr}{k}$$

AND must satisfy the condition for maximum slope.
This is used to graph lines of max. slope at the points of max. slope in Fig. 1C, after determining the appropriate value of *s* for each value of *Z* from the condition for max. slope (see below for examples).

The limit of the maximum slope as *n* goes to infinity can be seen to be *Z* from this, because *ns* goes to infinity as *n* goes to infinity. This can be seen from the following argument.
If the limit of *ns* were < infinity, then the condition for maximum slope

$$n-1 \;=\; (1+ns)^{Z-1}\,(\,(Z-1)\,ns-1)$$

could not hold, since the left side would blow up while the right side did not.

For $Z = 2$, the condition for max. slope:

$$n - 1 = (1 + ns)^{Z - 1}((Z - 1)\,ns - 1)$$

becomes:

$$n - 1 = (1 + ns)^{2 - 1}((2 - 1)\,ns - 1)$$

$$n - 1 = (1 + ns)\,(ns - 1) = n^2 s^2 - 1$$

$$n = n^2 s^2$$

$$s = \frac{1}{\sqrt{n}} = \frac{fr}{k}$$

Using this, the slope at its maximum for $Z = 2$ is:

$$\frac{2ns}{1 + ns} = \frac{2\sqrt{n}}{1 + \sqrt{n}}$$

For $Z = 3$, the condition for max. slope is:

$$n - 1 = (1 + ns)^{3-1}((3-1)ns - 1)$$

$$n - 1 = (1 + ns)^{2}(2ns - 1)$$

This is a cubic equation in the variable $s$ that can be solved by standard methods to give:

$$s = \frac{1}{2n}\left((2n - 1 + 2\sqrt{n(n-1)})^{1/3} + (2n - 1 + 2\sqrt{n(n-1)})^{\frac{-1}{3}} - 1\right)$$

the slope at its max. is then:

$$\frac{3}{1 + \frac{1}{ns}} = \frac{3}{1 + \frac{2}{(2n - 1 + 2\sqrt{n(n-1)})^{1/3} + (2n - 1 + 2\sqrt{n(n-1)})^{\frac{-1}{3}} - 1}}$$

For $Z = 4$, the condition for max. slope is:

$$n - 1 = (1 + ns)^3 (3ns - 1)$$

This is a quartic equation in *s* that can be solved by standard methods. The solution is easier to give in parts. Defining:

$$R = \sqrt{\frac{2}{3}\left(\frac{2}{3} - (n-1)^{1/3}\left((1+\sqrt{n})^{1/3} + (1-\sqrt{n})^{1/3}\right)\right)}$$

The single positive, real root of the quartic in this case gives, for the value of *ns* where the max. slope occurs:

$$ns = \frac{-2}{3} + \frac{R + \sqrt{\frac{4}{3} - R^2 + \frac{16}{27R}}}{2}$$

and the corresponding slope is given by

$$\frac{4}{\left(1 + \frac{1}{ns}\right)} = \frac{4}{\left(1 + \frac{1}{\frac{-2}{3} + \frac{R + \sqrt{\frac{4}{3} - R^2 + \frac{16}{27R}}}{2}}\right)}$$

The above expressions for the max. slope for $Z = 2$, 3, and 4 are used in Fig. 1C and Table 1.

## H progressive cooperativity

schematic of the binding equilibrium for progressive cooperativity, where cooperativity increases by an equal factor as each ligand is added, and derivation of an expression for the concentration of the $i^{th}$ complex $a_i$, which contains i ligand molecules, where:

Z = the total number of sites on the substrate

f = [free ligand],

r = forward rate constant for binding of a ligand molecule to the complex (or to the unliganded substrate h, free "hot" probe),

k = dissociation rate constant for the singly bound complex,

$k / n^{(i-1)/(Z-1)}$ = dissociation rate constant for a ligand molecule from the $i^{th}$ complex

k / n = dissociation rate constant for a ligand molecule from the fully liganded (the $Z^{th}$) complex.

to simplify some of the expressions, s is defined as:

s = fr/k; free ligand concentration divided by the equilibrium dissociation constant without cooperativity (k/r).

$$Zf + h \underset{k}{\overset{r}{\rightleftarrows}} (Z-1)f + a_1 \underset{\frac{k}{n^{1/(Z-1)}}}{\overset{r}{\rightleftarrows}} (Z-2)f + a_2 \underset{\frac{k}{n^{2/(Z-1)}}}{\overset{r}{\rightleftarrows}} \cdots \underset{\frac{k}{n^{(i-1)/(Z-1)}}}{\overset{r}{\rightleftarrows}} (Z-i)f + a_i \underset{\frac{k}{n^{i/(Z-1)}}}{\overset{r}{\rightleftarrows}} \cdots \underset{k/n}{\overset{r}{\rightleftarrows}} a_Z$$

$$Zfhr = a_1k$$

$$(Z-1)fa_1r = 2a_2k/n^{1/(Z-1)}$$

$$a_2 = n^{1/(Z-1)}(Z-1)fra_1 / 2k$$

$$a_2 = n^{1/(Z-1)}(Z-1)fr(Zfhr/k) / 2k$$

$$a_2 = n^{1/(Z-1)}Z(Z-1)(f^2r^2/k^2)h / 2$$

$$(Z-2)fa_2r = 3a_3k/n^{2/(Z-1)}$$

$$a_3 = n^{2/(Z-1)}(Z-2)fra_2 / 3k$$

$$(Z-i)fa_ir = (i+1)a_{i+1}k/n^{i/(Z-1)}$$

$$a_{i+1} = n^{(i+...+1)/(Z-1)}(Z-i)fra_i / (i+1)k$$

$$a_3 = n^{2/(Z-1)}(Z-2)fr \{n^{1/(Z-1)}Z(Z-1)(f^2r^2/k^2)h / 2\} / 3k$$

$$a_3 = n^{(2+1)/(Z-1)}Z(Z-1)(Z-2)(f^3r^3/k^3)h / (3)2$$

$$a_i = \frac{n^{i(i-1)/2(Z-1)}(fr/k)^i\, hZ!}{(Z-i)!\; i!} = \frac{n^{i(i-1)/2(Z-1)}s^i\, hZ!}{(Z-i)!\; i!}$$

I one ligand binding cooperatively to an arbitrary number of equivalent sites (progressive cooperativity); expressions for fractional occupancy and Hill plot formulae, including the condition for the point of maximum slope, and slopes of Hill plots at their point of max. slope (which is always where the fractional occupancy = ½).

Starting with the formula for the concentration of each complex containing *i* bound ligand molecules for progressive cooperativity, as show in H above:

🟥 $$a_i = \frac{hZ!}{(Z-i)!i!} n^{\frac{i(i-1)}{2(Z-1)}} s^i$$

The fractional occupancy is the total number of occupied sites divided by the total number of sites; therefore:

🟪 $$\theta = \frac{\sum_{i=0}^{Z} ia_i}{ZH} = \frac{\sum_{i=0}^{Z} ia_i}{Z\sum_{i=0}^{Z} a_i} = \frac{\sum_{i=0}^{Z} ia_i}{\sum_{i=0}^{Z} Za_i}$$

We will also need:

🟥 $$1-\theta = \frac{\left(\sum_{i=0}^{Z} Za_i\right) - \sum_{i=0}^{Z} ia_i}{\sum_{i=0}^{Z} Za_i} = \frac{\sum_{i=0}^{Z} (Z-i)a_i}{\sum_{i=0}^{Z} Za_i}$$

where

🟦 $$\sum_{i=0}^{Z} ia_i = \sum_{i=0}^{Z} \frac{ihZ!}{(Z-i)!i!} n^{\frac{i(i-1)}{2(Z-1)}} s^i$$

and

🟩 $$\sum_{i=0}^{Z} (Z-i)a_i = \sum_{i=0}^{Z} \frac{(Z-i)hZ!}{(Z-i)!i!} n^{\frac{i(i-1)}{2(Z-1)}} s^i$$

For a Hill plot, we need

$$\frac{\theta}{1-\theta} = \frac{\sum_{i=0}^{Z} i a_i}{\sum_{i=0}^{Z} (Z-i)\, a_i} = \frac{\sum_{i=0}^{Z} \frac{ihZ!}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i}{\sum_{i=0}^{Z} \frac{(Z-i)\, hZ!}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i} = \frac{\sum_{i=0}^{Z} \frac{i}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i}{\sum_{i=0}^{Z} \frac{(Z-i)}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i}$$

(cancelling out the common factors *hZ*! in the last step)

A Hill plot is the logarithm (ln) of this as a function of ln(*f*).
Since *s* = *fr*/*k*, if we substitute *x* = ln(*f*), then

$$f = e^x$$

and

$$s = \frac{r}{k} e^x$$

The slope of our Hill plot is:

$$\frac{\sum_{i=0}^{Z} \frac{i^2}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i}{\sum_{i=0}^{Z} \frac{i}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i} - \frac{\sum_{i=0}^{Z} \frac{(Z-i)\, i}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i}{\sum_{i=0}^{Z} \frac{(Z-i)}{(Z-i)\,!i!} n^{\frac{i\,(i-1)}{2\,(Z-1)}} s^i}$$

The max. slope occurs at 50% occupancy for all *Z*, and this occurs where *s* = 1/√*n*:

$$s = n^{\frac{-1}{2}}$$

After making this substitution, the slope at maximum is:

$$\frac{\sum_{i=0}^{Z} \frac{i\,(i-Z+i)}{(Z-i)\,!i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}{\sum_{i=0}^{Z} \frac{i}{(Z-i)\,!i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}} = \frac{\sum_{i=0}^{Z} \frac{i\,(2i-Z)}{(Z-i)\,!i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}{\sum_{i=0}^{Z} \frac{i}{(Z-i)\,!i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}$$

Each power of *n* in both summations occurs twice, since it is the same for *i* and *Z* – *i*, except for the middle term if *Z* is even, which = 0 in the numerator, because *i* = *Z*/2, so 2*i* – *Z* = 0; however, in the denominator, this term is non-zero and occurs only once, so, if we want to combine equal powers in the summation, we have to separate it out.

Combining them gives a form that is easier to use to write out the terms for a given *Z*.

For *Z* odd, the upper limit in the summations is (*Z* – 1)/2, giving a total of (*Z*+1)/2 terms (including *i* = 0).

For *Z* even, the upper limit in the summations is (*Z*/2 – 1), giving a total of *Z*/2 terms.

So, the maximum slope of a Hill plot with progressive cooperativity is,

for *Z* odd:

$$\frac{\sum_{i=0}^{\frac{Z-1}{2}} \frac{(Z-2i)^2}{(Z-i)\,!i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}{\sum_{i=0}^{\frac{Z-1}{2}} \frac{Z}{(Z-i)\,!i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}$$

For *Z* even, the same manipulations within the summation apply, except that the middle term in the denominator, where *i* = *Z*/2, occurs only once, and so we have to separate it out. The corresponding term in the numerator equals zero, giving the sums in the numerator and denominator the same summation limits.

So, the maximum slope of a Hill plot with progressive cooperativity is,
for *Z* even:

🟥 $$\frac{\displaystyle\sum_{i=0}^{\frac{Z}{2}-1} \frac{(Z-2i)^2}{(Z-i)!\,i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}}}{\displaystyle\sum_{i=0}^{\frac{Z}{2}-1} \frac{Z}{(Z-i)!\,i!}\, n^{\frac{(-i)\,(Z-i)}{2\,(Z-1)}} + \frac{\frac{Z}{2}}{\left(\frac{Z}{2}\right)!^{2}}\, n^{\frac{-Z^2}{8\,(Z-1)}}}$$

To see that the limit, as *n* goes to infinity, of the maximum slope is *Z*, we can note that all the powers of *n* are negative, except when $i = 0$. This gives a constant term, which in the numerator is:

🟩 $$\frac{Z^2}{Z!} = \frac{Z}{(Z-1)!}$$

while in the denominator it is:

⬜ $$\frac{Z}{Z!} = \frac{1}{(Z-1)!}$$

So, as *n* goes to infinity, all the powers of *n* go to 0, except for these constant terms, so that the slope in this limit is *Z*.

In fact, we can write the expression for the maximum slope to give terms that have coefficients of the powers of *n* that are all > 1, by multiplying both numerator and denominator by $(Z-1)!$.

**Table 1. Slopes of Hill plots at points of maximum slope and at 50% occupancy**, for one ligand binding to two or more equivalent sites

| | | | **1-step cooperativity** | | **progressive cooperativity** |
|---|---|---|---|---|---|
| cooperativity factor | # of sites | % occupancy at max. slope | slope at max. | slope at 50% occu. | slope at max. (50% occupancy) |
| 5 | **2** | 50% | 1.38 | 1.38 | 1.38 |
| 5 | **3** | 33% | 1.50 | 1.46 | 1.47 |
| 5 | **4** | 25% | 1.56 | 1.41 | 1.52 |
| 5 | **5** | 20% | 1.59 | 1.33 | 1.55 |
| 5 | **10** | 10% | 1.66 | 1.03 | 1.60 |
| 50 | **2** | 50% | 1.75 | 1.75 | 1.75 |
| 50 | **3** | 33% | 2.14 | 2.10 | 2.21 |
| 50 | **4** | 25% | 2.37 | 2.20 | 2.56 |
| 50 | **5** | 20% | 2.51 | 2.14 | 2.85 |
| 50 | **10** | 10% | 2.81 | 1.31 | 3.93 |
| 500 | **2** | 50% | 1.91 | 1.91 | 1.91 |
| 500 | **3** | 33% | 2.56 | 2.54 | 2.68 |
| 500 | **4** | 25% | 2.99 | 2.87 | 3.41 |
| 500 | **5** | 20% | 3.30 | 2.99 | 4.11 |
| 500 | **10** | 10% | 4.00 | 2.14 | 7.65 |
| 5000 | **2** | 50% | 1.97 | 1.97 | 1.97 |
| 5000 | **3** | 33% | 2.79 | 2.78 | 2.89 |
| 5000 | **4** | 25% | 3.41 | 3.33 | 3.80 |
| 5000 | **5** | 20% | 3.88 | 3.66 | 4.72 |
| 5000 | **10** | 10% | 5.09 | 3.30 | 9.37 |
| 5.00E+08 | **2** | 50% | 2.00 | 2.00 | 2.00 |
| 5.00E+08 | **3** | 33% | 3.00 | 2.99 | 3.00 |
| 5.00E+08 | **4** | 25% | 3.96 | 3.96 | 4.00 |
| 5.00E+08 | **5** | 20% | 4.88 | 4.86 | 5.00 |
| 5.00E+08 | **10** | 10% | 8.35 | 7.64 | 10.00 |
| 5.00E+13 | **2** | 50% | 2.00 | 2.00 | 2.00 |
| 5.00E+13 | **3** | 33% | 3.00 | 3.00 | 3.00 |
| 5.00E+13 | **4** | 25% | 4.00 | 4.00 | 4.00 |
| 5.00E+13 | **5** | 20% | 4.99 | 4.99 | 5.00 |
| 5.00E+13 | **10** | 10% | 9.47 | 9.24 | 10.00 |

Peacock and Jaynes, 2017

Fig. 2. Two ligands binding cooperatively to two distinct sites. **A**: Expression connecting [ternary complex] to the total [ligand-1] added to the mixture, which includes fixed total [ligand-2] and total [substrate], given the single-ligand binding site dissociation constants and the cooperativity factor. Expressions are also given for each [single-ligand complex], each free [ligand], and the free [substrate] as functions of the total [ligand-2], the total [substrate], the [ternary complex], the single-ligand binding site dissociation constants, and the cooperativity factor. Formulae are also given for the case of a single ligand binding to two distinct sites on a substrate. **B**: Modeling the graph of [ternary complex] as a function of the total concentration of one ligand (while both the total concentrations of the other ligand and of the substrate are held constant) as a single ligand binding to a single site. This model is based on the saturation concentration of ternary complex and the apparent Kd. **C**: An expression for where two such binding curves cross, as a way of determining the range of values where they cross and where they don't.

key for single-character notation used here:

| below | Peacock & Jaynes [1] | description |
|---|---|---|
| $H$ | $[AB]_T$ | total concentration of labeled substrate |
| $h$ | $[AB]$ | free concentration of labeled substrate |
| $p$ | $[a]_T$ | total concentration of protein1 |
| $f$ | $[a]$ | free concentration of protein1 |
| $q$ | $[b]_T$ | total concentration of protein2 |
| $g$ | $[b]$ | free concentration of protein2 |
| $k$ | $K_A$ | dissociation constant of protein1 from its single-protein complex |
| $L$ | $K_B$ | dissociation constant of protein2 from its single-protein complex |
| $n$ | $n$ | cooperativity factor |
| $a$ | $[AaB]$ | concentration of single-protein1 complex |
| $b$ | $[ABb]$ | concentration of single-protein2 complex |
| $A$ | $[AaBb]$ | concentration of ternary complex |

## A

**Contents:**

Expression connecting the [ternary complex] ($A$) and the total [ligand-1] ($p$), with fixed total [ligand-2] ($q$) and total [substrate] ($H$), given the single-ligand dissociation constants ($k$ and $L$) and the cooperativity factor ($n$).

Included are expressions for $a, b, f, g,$ and $h$ in terms of $q, A, k, L, n,$ and $H$.
The binding polynomial in terms of $f, g, k, L$ and $n$ is given at the end.

Definitions of variables:
$h$ = [free probe DNA, "hot" probe]
$a$ = [protein1 – site1 complex] = [($fh$)]
$b$ = [site2 – protein2 complex] = [($hg$)]
$A$ = [ternary complex] = [($fhg$)]
$H$ = total ["hot" DNA] = $h + a + b + A$
$f$ = [free protein1]
$p$ = total [protein1] = $f + a + A$
$g$ = [free protein2]
$q$ = total [protein2] = $g + b + A$

Dissociation (equilibrium) constants, including cooperativity factor $n$:
$k$ = dissociation constant of ($fh$), protein1 – site1 binary complex
$L$ = dissociation constant of ($hg$), site2 – protein2 binary complex
$k/n$ = dissociation constant of protein1 from site1 of ternary complex ($fhg$)
$L/n$ = dissociation constant of protein2 from site2 of ternary complex ($fhg$)

Equations governing equilibrium concentrations:
equation 1:

☒ $$L = \frac{gh}{b}$$

equation 2:

☒ $$k = \frac{fh}{a}$$

equation 3:

☒ $$H = h + a + b + A$$

equation 4:

☒ $$q = g + b + A$$

equation 5 (from either $L/n = ga / A$ combined with eq'n 1, OR, $k/n = fb / A$ combined with eq'n 2):

☒ $$nab = Ah$$

In order to get an expression connecting $A$ and $p$, with $q$ constant, given $H, k, L$, and $n$, we can use the fact that $p = f + a + A$. If we first get expressions for $f$ and $a$ that involve only the known quantities, this will give the desired connection.

**First, to get *a* in terms of *A* and constants, do the following.**

Summary of derivation:
Eliminate $g$ using 1 and 4.
Solve this for $h$ to get eq'n 14.
Eliminate $h$ using 5 and 14 and solve this for $b$ to get eq'n 145.
Eliminate $h$ using 3 and 5 and solve this for $b$ to get eq'n 35.
Eliminate $b$ using 145 and 35 to get eq'n 1345.
Manipulate this to get a quadratic in $a$ in terms of only the starting amounts of total protein2 ($q$) and DNA ($H$), the ternary complex $A$,
and the constants $L$ and $n$.

Detailed derivation available on request.

Applying the quadratic formula gives this expression for $a$:

☒ $$\frac{\left(\frac{A}{n}\right)(H + L - q) \pm \sqrt{\left(\frac{A}{n}\right)^2 (H + L - q)^2 + 4\left(\frac{A}{n}\right)^2 L \left(q - A + \frac{A}{n}\right)}}{2\left(q - A + \frac{A}{n}\right)}$$

from which $A/n$ can be factored out. For $a > 0$, the + sign applies when the denominator is $> 0$, and the – sign applies when the denominator is $< 0$;
however, from equation 4, $q–A = g+b$, which is always $> 0$, so the latter is never true, and

$a =$:

🟦 $$a = \frac{A}{n}\left(\frac{H+L-q+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left(q-A+\frac{A}{n}\right)}\right)$$

**To get *f* in terms of *A* and constants:**

Summary of derivation:
eliminate *h* using 3 and 5 and solve this for *b* to get eq'n 35;
eliminate both *a* and *h* (*h*/*a*) using 2 and 5; solve this for *f* to get eq'n 25;
substitute for *b* in 25 using 35 to get eq'n 235;
eliminate both *b* and *h* (*h*/*b*) using 1 and 5; solve this for *a* to get eq'n 15;
substitute this for *a* in 235 to get eq'n 1235;
eliminate *b* using 25 and 4 and solve this for *g* to get eq'n 245;
use this to substitute into 1235 to get a quadratic in *f* in terms of only the starting amounts of total protein2 (*q*) and DNA (*H*), the ternary complex *A*, and the constants *k*, *L*, and *n*.

Detailed derivation available on request.
Applying the quadratic formula gives:

☒ $$f = \frac{kA}{n}\left(\frac{H+L+q-2A\pm\sqrt{(H+L+q-2A)^2-4\left((H-A)\,(q-A)-\frac{LA}{n}\right)}}{2\left((H-A)\,(q-A)-\frac{LA}{n}\right)}\right)$$

For very small *p*, *A* is very small, and the denominator of *f* is > 0. As *p* becomes very large, it occupies all of the free DNA, and both *b* and *h* go to 0. The only forms of the DNA are then *a* and *A*, so $a = H - A$. Similarly, since $b = 0$, $q = g + A$, so $g = q - A$. Now, since $L/n = ga/A$, $AL/n = ga = (q - A)(H - A)$. This means that the denominator of *f* goes to 0. So, for all positive values of *p*, the denominator of $f > 0$, and the + sign in front of the radical applies.

**Because the denominator of *f* goes to zero as *p* goes to infinity, this gives us an implicit expression for *A* as *p* goes to infinity, in terms of *H, q, L*, and *n*:**
***(H – A)(q – A) = LA/n.***

The radicands in the formulae for *a* and *f* are equal (derivation available on request), and substituting the simpler version from the expression for *a* gives the following form for *f:*

🟥 $$f = \frac{kA}{n}\left(\frac{H+L+q-2A+\sqrt{(H+L-q)^2+4L\left(q-A+\frac{A}{n}\right)}}{2\left((H-A)\,(q-A)-\frac{LA}{n}\right)}\right)$$

**Now, because $p = A + a + f$, we can combine the above expressions for $a$ and $f$ to get:**

$$■ \quad p = A + \frac{A}{n}\left(\frac{H + L - q + \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}}{2\left(q - A + \frac{A}{n}\right)}\right) + \frac{kA}{n}\left(\frac{H + L + q - 2A + \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}}{2\left((H - A)\,(q - A) - \frac{LA}{n}\right)}\right)$$

**<u>Graphing *A* vs. *p* at constant *q* and *H* based on the above:</u>**

By setting $x = p$ and $y = A$ in the expression above, we obtain an expression that can be used in a variety of graphing applications to graph $A$ as a function of $p$.

For a given $p$ and $A$, it also gives an implicit expression for $n$ (see below for others).

**<u>Expressions for the other variables</u> in terms of the same constants and *A* are as follows:**

$$☒ \quad g = \frac{2L\left(q - A + \frac{A}{n}\right)}{H + L - q + \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}}$$

The + sign in front of the radical is appropriate for $g > 0$, because the numerator is always $> 0$, and the magnitude of the square root is greater than that of the terms that precede it in the denominator.

***h***:

$$☒ \quad h = \frac{(q - A)\left(H - L - q + \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}\right) - 2\frac{LA}{n}}{2\left(q - A + \frac{A}{n}\right)}$$

The + sign in front of the radical is always appropriate for $h > 0$, because the magnitude of the square root is greater than that of the preceding terms in the numerator, and the denominator is always $> 0$.
So, with the – sign, the numerator would be $< 0$, making $h < 0$.

***b***:

☒ $$b = \frac{1}{2}\left(H + L + q - 2A - \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}\right)$$

The – sign in front of the radical is always appropriate for $b > 0$, because with a + sign, $b + a + h$ would be greater than $H - A$, rather than $= H - A$, as it should be. This is consistent with the signs that are appropriate for $a$ and $h$, as explained above.

**We can also get the following two independent expression for *h* (derivations available on request):**

🟥 $$h = \frac{(p - A)\left(H - k - p + \sqrt{(H + k - p)^2 + 4k\left(p - A + \frac{A}{n}\right)}\right) - \frac{2kA}{n}}{2\left(p - A + \frac{A}{n}\right)}$$

🟥 $$h = \frac{1}{2A}\left(n\,(p - A)\,(q - A) - A\,(k + L) - \sqrt{\left(n\,(p - A)\,(q - A) - A\,(k + L)\right)^2 - 4A^2 kL}\right)$$

**<u>The binding polynomial</u>**, after e.g., Haiech et al, 2014 [4], is
P(x,y) = 1 + k1*x + k2*y + c1,1*k1*k2*x*y

🟪 $$P = 1 + \frac{f}{k} + \frac{g}{L} + \frac{n}{kL} fg$$

Is $P = H/h$?
Using expressions for $f$, $g$, and $n$ from equations 1, 2, and 5:

🟦 $$P = 1 + \frac{k\frac{a}{h}}{k} + \frac{L\frac{b}{h}}{L} + \frac{n}{kL} k\frac{a}{h} L\frac{b}{h} = 1 + \frac{a}{h} + \frac{b}{h} + \frac{Ah}{ab}\frac{a}{h}\frac{b}{h} = 1 + \frac{a}{h} + \frac{b}{h} + \frac{A}{h} = \frac{H}{h} + \frac{a}{h} + \frac{b}{h} + \frac{A}{h}$$

Yes, $P = H/h$ since, from equation 3:

⬛ $$\frac{H}{h} = \frac{h + a + b + A}{h}$$

**<u>Formulae for a SINGLE protein binding to 2 distinct binding sites on a substrate.</u>**
The formulae are somewhat simpler here, and we can obtain an explicit formula for $n$ in terms of constants and a measured amount of ternary complex as the total [ligand] is varied.

The variables used are similar to those used above, except that now there is no second ligand:
[2-ligand complex] = $A$
total [substrate] = $H$
free [substrate] = $h$
total [ligand] = $p$
free [ligand] = $f$

dissociation constants:

*k*, *L*, for ligand + each site on the substrate (the corresponding complexes are *a* and *b*, respectively).

🟦 $p = f + a + b + 2A$

Derivations of the following are available on request:

🟦 $$a = \frac{A}{2n}\left(\sqrt{\left(1+\frac{L}{k}\right)^2 + \frac{4nL}{Ak}(H-A)} - 1 - \frac{L}{k}\right) = \frac{p + H - 3A + \frac{kL}{k+L} - \sqrt{\left(H + A - p - \frac{kL}{k+L}\right)^2 + \frac{4(H-A)kL}{k+L}}}{2\left(1+\frac{k}{L}\right)}$$

⬛ $$b = \frac{k}{L}a = \frac{A}{2n}\left(\sqrt{\left(1+\frac{k}{L}\right)^2 + \frac{4nk}{AL}(H-A)} - 1 - \frac{k}{L}\right) = \frac{p + H - 3A + \frac{kL}{k+L} - \sqrt{\left(H + A - p - \frac{kL}{k+L}\right)^2 + \frac{4(H-A)kL}{k+L}}}{2\left(1+\frac{L}{k}\right)}$$

🟩 $$h = H - A - \frac{A}{2n}\left(\sqrt{\left(1+\frac{L}{k}\right)^2 + \frac{4nL}{Ak}(H-A)} - 1 - \frac{L}{k}\right)\left(1+\frac{k}{L}\right) = \frac{H - p + A - \frac{kL}{k+L} + \sqrt{\left(H + A - p - \frac{kL}{k+L}\right)^2 + \frac{4(H-A)kL}{k+L}}}{2}$$

🟨 $$f = \frac{k}{\frac{H-A}{\frac{A}{2n}\left(\sqrt{\left(1+\frac{L}{k}\right)^2 + \frac{4nL}{Ak}(H-A)} - 1 - \frac{L}{k}\right)} - 1 - \frac{k}{L}} = \frac{\frac{kL}{k+L}}{\frac{2(H-A)}{p + H - 3A + \frac{kL}{k+L} - \sqrt{\left(H + A - p - \frac{kL}{k+L}\right)^2 + \frac{4(H-A)kL}{k+L}}} - 1}$$

◻ $$n = \frac{2A\left(2 + \frac{k}{L} + \frac{L}{k}\right)\left(\frac{2(H-A)}{p + H - 3A + \frac{kL}{k+L} - \sqrt{\left(H + A - p - \frac{kL}{k+L}\right)^2 + \frac{4(H-A)kL}{k+L}}} - 1\right)}{p + H - 3A + \frac{kL}{k+L} - \sqrt{\left(H + A - p - \frac{kL}{k+L}\right)^2 + \frac{4(H-A)kL}{k+L}}}$$

B Modeling the graph of [ternary complex] as a function of the total concentration of one ligand (while both the total concentrations of the other ligand and of the substrate are held constant) as a single ligand binding to a single site. This model is based on the saturation concentration of ternary complex and the apparent Kd.

**Contents:**

"Modeling" the graph of *A* vs. *p* as a single protein binding to a single site based on *A-max* and
the apparent Kd, which is the concentration of "unbound" protein1 where *A* = 1/2 *A-max*.

Note: For a binary complex, when the concentration of complex is at half-maximum, the concentration of free ligand equals the Kd. Also, for a binary complex, the binding curve saturates at 100% of probe bound, whereas here, this is not the case. However, the apparent Kd and saturation value are typically found by curve fitting to a theoretical curve that best approximates the data, so that the saturation value found is not necessarily at 100% of probe bound. Thus, what is derived below is a theoretical binary binding curve that gives an apparent Kd that is similar to what is typically found when this procedure is applied to binding data of this kind.

Definitions of variables:
$h$ = [free probe DNA, "hot" probe]
$a$ = [protein1 – site1 complex] = [($fh$)]
$b$ = [site2 – protein2 complex] = [($hg$)]
$A$ = [ternary complex] = [($fhg$)]
$H$ = total ["hot" DNA] = $h + a + b + A$
$f$ = [free protein1]
$p$ = total [protein1] = $f + a + A$
$g$ = [free protein2]
$q$ = total [protein2] = $g + b + A$

Dissociation (equilibrium) constants, including cooperativity factor $n$:
$k$ = dissociation constant of ($fh$), protein1 – site1 binary complex
$L$ = dissociation constant of ($hg$), site2 – protein2 binary complex
$k/n$ = dissociation constant of protein1 from site1 of ternary complex ($fhg$)
$L/n$ = dissociation constant of protein2 from site2 of ternary complex ($fhg$)

From Fig. 2A, $f$ is:

🟥
$$f = \frac{kA}{n}\left(\frac{H + L + q - 2A + \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}}{2\left((H - A)\,(q - A) - \frac{LA}{n}\right)}\right)$$

Maximum $A$ ($A$-$max$) as $p$ approaches infinity (and therefore $f$ approaches infinity) occurs when the denominator of $f$ goes to 0:

☒
$$(H - A)\,(q - A) - \frac{LA}{n} = 0$$

☒
$$Hq - AH - Aq + A^2 - \frac{LA}{n} = 0 = A^2 - \left(H + q + \frac{L}{n}\right)A + Hq$$

So, *A-max* is:

☒ $$A\,(max) \;=\; \frac{1}{2}\left(H + q + \frac{L}{n} - \sqrt{\left(H + q + \frac{L}{n}\right)^{2} - 4Hq}\right)$$

Define *B* as *A-max* / 2:

☒ $$B \;=\; \frac{1}{4}\left(H + q + \frac{L}{n} - \sqrt{\left(H + q + \frac{L}{n}\right)^{2} - 4Hq}\right)$$

The apparent Kd in a single protein – site model, based on the [free ligand] where *A* is at 50% of its maximum, is found by substituting *A-max*/2 for *A* in the formula for *p* derived in Fig. 2A, then subtracting *B* (= *A-max*/2) from this to get the expected [free ligand] in a binary model.

In the binary model, there is no complex that corresponds to *a*. So, in the binary model, total ligand is simply the sum of bound ligand and free ligand, which means that free ligand is the difference between total ligand (*p*) and bound ligand (*A*).

Therefore, the [free ligand] in the binary model corresponds to *a*+*f* here, which = *p*–*A*.

From Fig. 2A, *p*–*A* = *a*+*f*:

☒ $$p - A \;=\; \frac{A}{n}\left(\frac{H + L - q + \sqrt{(H + L - q)^{2} + 4L\left(q - A + \frac{A}{n}\right)}}{2\left(q - A + \frac{A}{n}\right)}\right) + \frac{kA}{n}\left(\frac{H + L + q - 2A + \sqrt{(H + L - q)^{2} + 4L\left(q - A + \frac{A}{n}\right)}}{2\left(\,(H - A)\;(q - A) - \frac{LA}{n}\right)}\right)$$

The value of *p*–*A* = *f*+*a*, with *B* substituted for *A*, is the apparent Kd, which we call *S*:

$$S \;=\; \frac{B}{n}\left(\frac{(H - q + L) + \sqrt{(H - q + L)^{2} + 4L\left(q - B + \frac{B}{n}\right)}}{2\left(q - B + \frac{B}{n}\right)}\right) + \frac{kB}{n}\left(\frac{(H + q + L - 2B) + \sqrt{(H - q + L)^{2} + 4L\left(q - B + \frac{B}{n}\right)}}{2\left(\,(H - B)\;(q - B) - \frac{BL}{n}\right)}\right)$$

Now, a graph of the expected saturation curve with this Kd, in a single-protein – site model, is obtained by putting *k* = *S* into the formula *p* = *A* + *f* = *A* + *kA* / *h*, or *p* / *A* = 1 + *k* / (*H* – *A*); here, *H* is replaced by *A-max*, which = 2*B*:

☒ $$\frac{p}{A} \;=\; 1 + \frac{S}{2B - A}$$

C

**Contents:**

An expression for where two binding curves cross (the concentration of one ligand is varied, while the total concentration of the other, and that of the substrate, are held constant), as a way of determining the range of values where they cross and where they don't:

Binding reaction:
two proteins (or complexes) bind to 2 sites on the "probe" (labeled) substrate (e.g., DNA);

The following are the variables:

[ternary complex] = *A*
total [labeled substrate] = *H*
total [each ligand] = *p, q*
dissociation constants are *k, L*, for each ligand + probe (= labeled substrate; single ligand-substrate complexes are *a* and *b*);
*n* = cooperativity factor;

*H* = total [labeled DNA], which includes both bound (*a*+*b*+*A*) and free (*h*;
*a* and *b* are the labeled single-protein complexes, *A* is the ternary complex);

the dissociation constants of the labeled DNA – protein complexes are *k* and *L*;

dissociation constants of 2-protein complexes with probe (ternary complex) are *k*/*n* and *L*/*n*.

$H = h + a + b + A$

concentrations of free proteins in terms of concentrations of single-protein complexes and free probe:
$f = ka\,/\,h$; $g = Lb\,/\,h$; the two total proteins:
$p = f + a + A$ (graphed as a function of *A*, with axes reversed, so *A* as a function of *p*)
$q = g + b + A$ (here given, so *g* and *b* are not explicit, but have been eliminated as variables)

We can graph total protein1 concentration (*x*), in 2 forms for comparison of how it varies with *A* (*y*)
at two different values of *k* ($\mathbb{k}$), *L* ($\mathbb{L}$), and *n* ($\mathbb{m}$):

🟦 $$x = \left( y + \frac{y}{n} \left( \frac{H - q + L + \sqrt{(H - q + L)^2 + 4L\left(q - y + \frac{y}{n}\right)}}{2\left(q - y + \frac{y}{n}\right)} \right) + \frac{ky}{n} \left( \frac{H + q + L - 2y + \sqrt{(H - q + L)^2 + 4L\left(q - y + \frac{y}{n}\right)}}{2\left( (H - y)\ (q - y)\ - \frac{yL}{n}\right)} \right) \text{ if } y < q \right)$$

🟥 $$x = \left( y + \frac{y}{\mathbb{m}} \left( \frac{H - q + \mathbb{L} + \sqrt{(H - q + \mathbb{L})^2 + 4\mathbb{L}\left(q - y + \frac{y}{\mathbb{m}}\right)}}{2\left(q - y + \frac{y}{\mathbb{m}}\right)} \right) + \frac{\mathbb{k}y}{\mathbb{m}} \left( \frac{H + q + \mathbb{L} - 2y + \sqrt{(H - q + \mathbb{L})^2 + 4\mathbb{L}\left(q - y + \frac{y}{\mathbb{m}}\right)}}{2\left( (H - y)\ (q - y)\ - \frac{y\mathbb{L}}{\mathbb{m}}\right)} \right) \text{ if } y < q \right)$$

To find where these curves cross (and to see if it occurs for any relevant values of $A$=$y$ and $H$=$x$ below), solve these simultaneously, by setting them = to each other and cancelling terms, after substituting $x$ for $H$ and $y$ for $A$. Since we want only values of $A$ that are $< H$ and also $< q$, we can use the conditions $y < x$ and $y < q$ to limit the search for solutions.

This graph then tells us the range of values of $H$ (and $A$) where the curves cross, given the assigned variables.

If we also assign $H$, then the values of $A = y$ where the curves cross is the intersection between the above curve and $x = H$.
This intersection may consist of 0, 1, or 2 points.
If there is no point of intersection, then, of course, the curves do not cross.
If there are two points of intersection, then the curves generally are close together throughout their range.
If there is one point of intersection, then there is a range of possible behaviors, as illustrated in Fig. 2 of Peacock and Jaynes [1].

🟩
$$\frac{1}{n}\left(\frac{x-q+L+\sqrt{(x-q+L)^2+4L\left(q-y+\frac{y}{n}\right)}}{q-y+\frac{y}{n}}\right)+\frac{k}{n}\left(\frac{x+q+L-2y+\sqrt{(x-q+L)^2+4L\left(q-y+\frac{y}{n}\right)}}{(x-y)\,(q-y)-\frac{yL}{n}}\right)$$
$$=\left(\left(\frac{1}{\mathbb{m}}\left(\frac{x-q+\mathbb{L}+\sqrt{(x-q+\mathbb{L})^2+4\mathbb{L}\left(q-y+\frac{y}{\mathbb{m}}\right)}}{q-y+\frac{y}{\mathbb{m}}}\right)+\frac{\mathbb{k}}{\mathbb{m}}\left(\frac{x+q+\mathbb{L}-2y+\sqrt{(x-q+\mathbb{L})^2+4\mathbb{L}\left(q-y+\frac{y}{\mathbb{m}}\right)}}{(x-y)\,(q-y)-\frac{y\mathbb{L}}{\mathbb{m}}}\right)\text{ if } x>y\right)\text{ if } y<q\right)$$

The curve with the higher value of (probe-protein2 Kd / cooperativity factor) (here, $L/n$ or $\mathbb{L}/\mathbb{m}$) saturates at a lower value,
which means that the curves will necessarily cross if its initial slope is greater.
We can use this to obtain an estimate for when two curves will cross.

Initial slopes (reciprocals of the derivatives of the above expressions w.r.t. $y$ when $A = y = 0$):

⊠
$$\frac{2qn}{2qn+\left(H-q+L+\sqrt{(H-q+L)^2+4qL}\right)+\frac{k}{H}\left(H+q+L+\sqrt{(H-q+L)^2+4qL}\right)}$$

When $H$ and $q$ are both much smaller than $k$ and $L$, this can be approximated by:

⊠
$$\frac{1}{1+\frac{kL}{nHq}}$$

The initial slope in the equivalent single-site experiment (the plot of $a$ vs. $p$) is ($K$ = single-ligand Kd):

⊠
$$\frac{1}{1+\frac{K}{H}}$$

so the apparent Kd at very low [ligand] for this type of 2-site experiment is (approximately, when *H* and *q* are both much smaller than *k* and *L*):

☒ $$\frac{kL}{nq}$$

and the ratio of apparent Kd's for the two sites is, therefore (approximately, when *H* and *q* are both much smaller than *k* and *L*):

☒ $$\frac{\mathbb{m} k L}{n \mathbb{k} \cdot \mathbb{L}}$$

This would likely approximate the "relative affinity" found in a SELEX-seq experiment of the type described in Riley et al., 2014 [5], where the protein concentrations are assumed to be much less than the "Kd", and the concentrations of the oligonucleotides containing the binding sites are also very low. Such determinations made at very low concentrations generally underestimate the affinity of sites for pairs of ligands with high cooperativity relative to sites with lower individual Kd's (and lower cooperativity). Such sites that support high cooperativity are also those that have the highest occupancy when one ligand concentration (e.g., [En]) is high while the other (e.g., [Exd/Hth]) is limiting, as illustrated in Figs. 2B and 3C of Peacock & Jaynes [1].

The initial slope with the second set of Kd's and cooperativity factor is:

☒ $$\frac{2q\mathbb{m}}{2q\mathbb{m} + H - q + \mathbb{L} + \sqrt{(H - q + \mathbb{L})^2 + 4q\mathbb{L}} + \frac{\mathbb{k}}{H}\left(H + q + \mathbb{L} + \sqrt{(H - q + \mathbb{L})^2 + 4q\mathbb{L}}\right)}$$

when initial slopes are equal:

■ $$\frac{\left(x + \mathbb{L} + \sqrt{(x - y + \mathbb{L})^2 + 4y\mathbb{L}}\right)(\mathbb{k} + x) + y(\mathbb{k} - x)}{\mathbb{m}} = \frac{\left(x + L + \sqrt{(x - y + L)^2 + 4yL}\right)(k + x) + y(k - x)}{n}$$

To graph as *q* (= *y*) vs. *H* (= *x*) where this condition is met:

☒ $$\frac{\left(H + \mathbb{L} + \sqrt{(H - q + \mathbb{L})^2 + 4q\mathbb{L}}\right)(\mathbb{k} + H) + q(\mathbb{k} - H)}{\mathbb{m}} = \frac{\left(H + L + \sqrt{(H - q + L)^2 + 4qL}\right)(k + H) + q(k - H)}{n}$$

Values of *x* (= *H*) and *y* (= *q*) that are to the left of this curve are in the range where the curves (of *A* vs. *p*) will cross.
However, the converse is not always true, because if the curves cross twice,
the initial slope of the curve with the higher saturation value of *A* will be higher, not lower, than the initial slope of the other curve.

If *H* and *q* are both negligibly small relative to *k*, $\mathbb{k}$, *L*, and $\mathbb{L}$, the above simplifies to:

☒ $$\frac{\left(\mathbb{L} + \sqrt{\mathbb{L}^2}\right)\mathbb{k}}{\mathbb{m}} = \frac{\left(L + \sqrt{L^2}\right)k}{n}$$

So:

☒ $$\frac{\mathbb{k}}{\mathbb{m}}\mathbb{L} = \frac{k}{n}L$$

Fig. 3. **A:** expressions that allow graphing of [ternary complex] (as well as free ligand, substrate, competitor, and other complexes) as a function of added total [unlabeled competitor substrate] using simple equation-graphing software. **B:** total occupancy of substrate by each of two ligands, as a function of total [ligand-1] (without unlabeled competitor), and the maximum value of each total [bound ligand] as the total [ligand-1] goes to infinity.

key for single-character notation used here:

| below | Peacock & Jaynes [1] | description |
|---|---|---|
| $H$ | $[AB]_T$ | total concentration of labeled substrate |
| $h$ | $[AB]$ | free concentration of labeled substrate |
| $p$ | $[a]_T$ | total concentration of protein1 |
| $f$ | $[a]$ | free concentration of protein1 |
| $q$ | $[b]_T$ | total concentration of protein2 |
| $g$ | $[b]$ | free concentration of protein2 |
| $k$ | $K_A$ | dissociation constant of protein1 from its single-protein complex |
| $L$ | $K_B$ | dissociation constant of protein2 from its single-protein complex |
| $n$ | $n$ | cooperativity factor |
| $a$ | $[AaB]$ | concentration of single-protein1 complex |
| $b$ | $[ABb]$ | concentration of single-protein2 complex |
| $A$ | $[AaBb]$ | concentration of ternary complex |

Other variables not used in Peacock and Jaynes [1] are described below.

# A.

**Contents:** An expression for $U$ (= total [specific unlabeled competitor substrate]) as a function of
$A$ (= [labeled ternary complex]) and $f$ (= [free ligand1]).

An expression for $f$ as a function of $A$.
(The original expression is quartic in $f$, and so it uses the general solution for a quartic equation.)
This can be substituted into the expression for $U$ (as a function of $A$ and $f$) to get $U$ as a function of $A$ for graphing, using simple software.

The original expression involving $f$ and $A$ is cubic in $A$, and so it can be solved for $A$ as a function of $f$ using the general cubic solution,
which is simpler and easier for software packages to handle than the quartic solution.
So, this solution is given, along with expressions for the other variables in the system as a function of $A$ and $f$.
Together, these allow graphing of each of the varying forms involved in the experiment
as a function of any other one of these forms (most usefully, $A$ as a function of $U$), using $f$ as a parameter
along with the cubic solution for $A$ as a function of $f$. For example, we can graph $y$ vs. $x$ where $y = A(f)$ and $x = U[A(f), f]$.

Definitions of variables:
$h$ = free [labeled substrate = "hot" DNA]
$a$ = [protein1 – site1 complex] = [$(fh)$]
$b$ = [site2 – protein2 complex] = [$(hg)$]
$A$ = ["hot" ternary complex] = [$(fhg)$]
$H$ = total [labeled substrate = "hot" DNA] = $h + a + b + A$

$u$ = free [competitor substrate, "unlabeled"]
$c$ = [protein1 – site1 complex] = [$(fu)$]
$d$ = [site2 – protein2 complex] = [$(ug)$]
$B$ = [unlabeled ternary complex] = [$(fug)$]
$U$ = total [unlabeled substrate = competitor DNA] = $u + c + d + B$

$f$ = free [protein1]
$p$ = total [protein1] = $f + a + c + A + B$
$g$ = free [protein2]
$q$ = total [protein2] = $g + b + d + A + B$

<u>Dissociation (equilibrium) constants, including cooperativity factor *n*:</u>
for labeled complexes:
$k$ = dissociation constant of ($fh$), protein1 – site1 binary complex
$L$ = dissociation constant of ($hg$), site2 – protein2 binary complex
$k/n$ = dissociation constant of protein1 from site1 of ternary complex ($fhg$)
$L/n$ = dissociation constant of protein2 from site2 of ternary complex ($fhg$)

for unlabeled complexes:
$Q$ = dissociation constant of ($fu$), protein1 – site1 binary complex
$R$ = dissociation constant of ($ug$), site2 – protein2 binary complex
$Q/m$ = dissociation constant of protein1 from site1 of ternary complex ($fug$)
$R/m$ = dissociation constant of protein2 from site2 of ternary complex ($fug$)

equation 0:

🟨 $p = f + a + A + c + B$

equation 1:

🟦 $L = \frac{gh}{b}$

equation 2:

🟩 $k = \frac{fh}{a}$

equation 3:

🟦 $H = h + a + b + A$

equation 4:

🟨 $q = g + b + A + d + B$

equation 5:

⬜ $\frac{k}{n} = \frac{fb}{A}$

equation 6:

⬛ $\frac{L}{n} = \frac{ga}{A}$

equation 7:

🟪 $Q = \frac{fu}{c}$

equation 8:

$$R = \frac{g u}{d}$$

equation 9:

$$U = u + c + d + B$$

equation 10:

$$\frac{Q}{m} = \frac{f d}{B}$$

equation 11:

$$\frac{R}{m} = \frac{g c}{B}$$

To get $U$ as a function of $A, f, g, p, H, k, L, n, Q, R$, and $m$:
starting with equation 9,
eliminate $u$ using equation 7,
then $d$ using equation 10,
then $B$ using equation 11,
then $c$ using equation 0', which is 0 with $B$ eliminated using equation 11, to yield equation 9';
eliminate $a$ using equation 6 to yield equation 9''.
Eliminate $g$ from this (see below) to get equation 9'''.

equation 9':

$$U = \frac{p - f - a - A}{1 + \frac{m g}{R}} \left(1 + \frac{Q}{f} + \frac{m g}{R} + \frac{Q g}{R f}\right)$$

equation 9''':

$$U = \frac{p - f - \frac{L A}{n g} - A}{\frac{R}{g} + m} \left(\frac{R}{g} + m + \frac{Q R}{f g} + \frac{Q}{f}\right)$$

To get $g$ in terms of $f$ and only $A, k, L, n, and\ H$:
starting with equation 3,
eliminate $h$ using equation 2 to yield equation 3';
eliminate $b$ using equation 5,
then $a$ using 6 to yield equation 3'';
solve for $g$.

equation 3'':

🟦 $$H = \frac{LA}{ng}\left(1+\frac{k}{f}\right) + A\left(1+\frac{k}{nf}\right)$$

to be used in the final equation for *U*:

⬜ $$\frac{H - A\left(1+\frac{k}{nf}\right)}{\left(1+\frac{k}{f}\right)} = \frac{LA}{ng} = c$$

🟨 $$g = \frac{\frac{L}{n}(k+f)}{\left(\frac{H}{A}-1\right)f - \frac{k}{n}} = \frac{L(k+f)}{\left(\frac{H}{A}-1\right)nf - k}$$

Substituting these for *LA*/*ng* (= *c*) and *g* in *U* from above gives:

🟪 $$U = \left(p - f - A - \frac{H - A\left(1+\frac{k}{nf}\right)}{\left(1+\frac{k}{f}\right)}\right)\left(1 + \frac{Q\left(R + \frac{L(k+f)}{\left(\frac{H}{A}-1\right)nf-k}\right)}{f\left(R + \frac{mL(k+f)}{\left(\frac{H}{A}-1\right)nf-k}\right)}\right)$$

After rearranging:

⬜ $$U = \left(\left(p - A + \frac{A}{n}\right)k + (p-k-H)f - f^2\right)\left(\frac{((H-A)nR + mLA)f^2 + ((H-A)nQR + (mL-R)kA + QLA)f + (L-R)QkA}{((Rn(H-A) + mLA)f^2 + (Rnk(H-A) + (2mL-R)kA)f + (mL-R)k^2A)f}\right)$$

This is the same form as that used for graphing with the quartic solution (below) for *f* as a function of *A*.

Rearranging terms to get a more consistent-looking form:

⬛ $$U = \left(k\left(p - A + \frac{A}{n}\right) + (p-k-H)f - f^2\right)\left(\frac{kQ(L-R)A + (nQRH + (k(mL-R) + Q(L-nR))A)f + (nRH + (mL-nR)A)f^2}{(k^2(mL-R)A + (nkRH + k(2mL-(n+1)R)A)f + (nRH + (mL-nR)A)f^2)f}\right)$$

To get what will become a quartic in *f* after *g* and *U* are eliminated (and *q* is reintroduced in the process)
containing only the other variables *A*, *f*, *g*, *p*, *q*, *k*, *L*, *n*, *Q*, *R*, and *m*:
starting with equation 4,
eliminate *d* using equation 10,
then *B* using equation 11,
then *c* using equation 0' to yield equation 4';
eliminate *b* using equation 5 and *a* using equation 6 to yield equation 4'':

equation 4'':

$$\frac{q-g-A-\frac{kA}{nf}}{\frac{Q}{f}+m} = \frac{p-f-A-\frac{LA}{ng}}{\frac{R}{g}+m} = \frac{U}{\left(\frac{QR}{fg}+\frac{Q}{f}+\frac{R}{g}+m\right)}$$

where the last equality uses the expression for *U* from above, equation 9'''.

The first two of these equalities give a quartic in *f* (containing only constants and *A*)
when *g* is expressed in terms of *f* using the above formula, along with:

$$\frac{1}{g} = \frac{(H-A)\,nf-kA}{LA\,(k+f)}$$

and

$$\frac{LA}{ng} = \frac{(H-A)\,f-k\frac{A}{n}}{(k+f)}$$

Making these substitutions in the first two equalities of the above:

$$\frac{q-\frac{LA\,(k+f)}{(H-A)\,nf-kA}-A-\frac{kA}{nf}}{\frac{Q}{f}+m} = \frac{p-f-A-\frac{(H-A)\,f-k\frac{A}{n}}{(k+f)}}{R\frac{(H-A)\,nf-kA}{LA\,(k+f)}+m}$$

and rearranging to eliminate negative powers of *f*, then collecting powers of *f*, resolving the fractions, and expanding and collecting powers of *f* gives:

$$f^4+\left(q-p+k+H-A+\frac{Q-R}{m}+\frac{nR}{mLA}\,(q-A)\,(H-A)-\frac{A\,(k+L)}{n\,(H-A)}\right)f^3$$
$$+\left(k\left(q-p-3\frac{A}{n}+(3A-H-2q)\,\frac{R}{mL}+\frac{Q-R}{m}+\frac{A}{n}\,\frac{p-q-k-2L+\frac{R-Q}{m}}{H-A}\right)+\frac{Q}{m}\,(H-p)\right)f^2$$
$$+k\left(\frac{kA}{n}\left(\frac{2R}{mL}-1\right)-\frac{Q}{m}\left(p-A+2\frac{A}{n}\right)+\frac{A}{n}\,\frac{\frac{Q}{m}\,(p-A-k)+k\left(p-q-L+2\frac{A}{n}+\frac{R}{m}\left(1+\frac{q-A}{L}\right)\right)}{H-A}\right)f+\frac{k^2A}{n}\,\frac{\frac{Q}{m}\,(p-A)+\frac{A}{n}\left(\frac{Q}{m}+k-\frac{kR}{mL}\right)}{H-A} = 0$$

Applying the general solution of a quartic equation using these coefficients gives *f* as a function of *A* and
other quantities that are all constant in an experiment where *U* is varied and *A* is measured.

Start with the expression for *U* that is quartic in *f*, given above, and simplify it to get fewer explicit *A*'s (because *A* will be substituted for using the solution to the cubic given below):

☒ $$U = \left(\left(p - A\left(1 - \frac{1}{n}\right)\right)k + (p - k - H)f - f^2\right)\left(\frac{nHRf(Q+f) + ((mL - nR)f^2 + ((mL - R)k + (L - nR)Q)f + (L - R)Qk)A}{(nHRf(k+f) + ((mL - nR)f^2 + ((2mL - (n+1)R)f + (mL - R)k)k)A)f}\right)$$

This can be used to graph *A* as a function of *U* (or vice versa), using the cubic solution for *A*(*f*), and *f* as the parameter, starting with the same expression as above, and solving it for *A* as a function of *f*, instead of vice versa. This expression is:

🟩 $$\frac{q - \dfrac{LA(k+f)}{(H-A)nf - kA} - A - \dfrac{kA}{nf}}{\dfrac{Q}{f} + m} = \frac{p - f - A - \dfrac{(H-A)f - k\dfrac{A}{n}}{(k+f)}}{R\dfrac{(H-A)nf - kA}{LA(k+f)} + m}$$

First, before giving the general solution for *A* as a function of *f* and constants:

<u>For self competition</u>, where *m* = *n*, *Q* = *k*, and *R* = *L*, the solution is simpler (and the cubic solution given below fails).
So, here is the correct expression in that special case;
substituting *m* = *n*, *Q* = *q*, and *R* = *L*:

☒ $$\frac{q - \dfrac{LA(k+f)}{(H-A)nf - kA} - A - \dfrac{kA}{nf}}{\dfrac{k}{f} + n} = \frac{p - f - A - \dfrac{(H-A)f - k\dfrac{A}{n}}{k+f}}{\dfrac{(H-A)nf - kA}{A(k+f)} + n}$$

so,

☒ $$0 = \left(q - \frac{LA(k+f)}{(H-A)nf - kA} - A - \frac{kA}{nf}\right)\left(\frac{(H-A)nf - kA}{A(k+f)} + n\right) - \left(\frac{k}{f} + n\right)\left(p - f - A - \frac{(H-A)f - k\dfrac{A}{n}}{k+f}\right)$$

which gives a quadratic in *A*:

☒ $$\left(q\frac{(k+nf)^2}{k+f} - nL(k+f) - (k+nf)\left((k+nf)\left(1 - \frac{p}{f}\right) + nq - L\right)\right)A^2 + \left(nf^2 + (k - L + n(q-p))f - pk - \frac{2qf(k+nf)}{k+f}\right)nHA + \frac{q(Hnf)^2}{k+f}$$
= 0.

This expression can be used to model self-competition (by graphing any of the variables as a function of any other variable) using the quadratic solution to this for *A* as a function of *f*, and *f* as a parameter, along with expressions for each of the variables as a function of *A* and *f* given below, as described below for the cubic solution.

**For non-self competition**, rearranging the above to give the cubic in *A*
0 = :

☒ $$\left(\left(m-\frac{R}{L}\right)(k+nf)^2-(n-1)\,(mf(2k+nf)+kQ)\right)\frac{k+nf}{nf(k+f)}A^3$$
$$+\left(\left(\frac{(q+2H)\,R}{L\,(k+f)}\right)(k+nf)^2-\left(m\,(f-p+q)+Q-R-p\frac{Q}{f}+2H\frac{Q+mf}{k+f}\right)(k+nf)+(nQ-mk)\,H-mL\,(k+f)\right)A^2$$
$$+\left(\frac{f}{k+f}\left(H\,(Q+mf)-\frac{R\,(H+2q)}{L}\,(k+nf)\right)+mf^2+(Q-R+m\,(q-p))\,f-pQ\right)nHA+\frac{qR\,(Hnf)^2}{L\,(k+f)}$$

The 4 parts are:

☒ $$\left(\left(\left(m-\frac{R}{L}\right)(k+nf)^2-(n-1)\,(mf(2k+nf)+kQ)\right)\frac{k+nf}{nf(k+f)}\right)A^3$$

☒ $$\left(\left(\frac{(q+2H)\,R}{L\,(k+f)}\right)(k+nf)^2-\left(m\,(f-p+q)+Q-R-p\frac{Q}{f}+2H\frac{Q+mf}{k+f}\right)(k+nf)+(nQ-mk)\,H-mL\,(k+f)\right)A^2$$

☒ $$\left(\left(\frac{f}{k+f}\left(H\,(Q+mf)-\frac{R\,(H+2q)}{L}\,(k+nf)\right)+mf^2+(Q-R+m\,(q-p))\,f-pQ\right)nH\right)A$$

☒ $$\frac{qR\,(Hnf)^2}{L\,(k+f)}$$

From this, construct the cubic solution for *A* as a function of *f*.
From the general solution to the cubic equation:

☒ $$Vz^3+Wz^2+Xz+Y\;=\;0$$

The 3 solutions are:

☒ $$\frac{-W}{3V}+S+T$$

☒ $$\frac{-W}{3V}-\left(\frac{S+T}{2}\right)+i\left(\frac{S-T}{2}\right)\sqrt{3}$$

☒ $$\frac{-W}{3V}-\left(\frac{S+T}{2}\right)-i\left(\frac{S-T}{2}\right)\sqrt{3}$$

where

☒ $$S\;=\;\left(\frac{WX}{6V^2}-\frac{Y}{2V}-\left(\frac{W}{3V}\right)^3+\sqrt{\frac{X^3}{27V^3}-\frac{W^2X^2}{108V^4}-\frac{WXY}{6V^3}+\frac{Y^2}{4V^2}+\frac{YW^3}{27V^4}}\right)^{1/3}$$

and

☒ $$T = \left(\frac{WX}{6V^2} - \frac{Y}{2V} - \left(\frac{W}{3V}\right)^3 - \sqrt{\frac{X^3}{27V^3} - \frac{W^2X^2}{108V^4} - \frac{WXY}{6V^3} + \frac{Y^2}{4V^2} + \frac{YW^3}{27V^4}}\right)^{1/3}$$

Assigning coefficients from above (with *f* changed to *t*):

☒ $$V = \left(\left(m - \frac{R}{L}\right)(k+nt)^2 - (n-1)\,(mt\,(2k+nt) + kQ)\right)\frac{k+nt}{nt\,(k+t)}$$

☒ $$W = \left(\frac{(q+2H)\,R}{L\,(k+t)}\right)(k+nt)^2 - \left(m\,(t-p+q) + Q - R - p\frac{Q}{t} + 2H\frac{Q+mt}{k+t}\right)(k+nt) + (nQ-mk)\,H - mL\,(k+t)$$

☒ $$X = \left(\frac{t}{k+t}\left(H\,(Q+mt) - \frac{R\,(H+2q)}{L}\,(k+nt)\right) + mt^2 + (Q - R + m\,(q-p))\,t - pQ\right)nH$$

☒ $$Y = \frac{qR\,(Hnt)^2}{L\,(k+t)}$$

At least for the values tested, the first solution is the correct one.

The formulae for the variables are as follows:
Note: to graph these parametrically (the parameter is *t*) as a function of *A*, substitute the cubic solution above for *A*, and *t* for *f*.
For example, to graph *A* as a function of *U*, graph *A*(*t*) vs. *U*(*t*), as *t* ranges from 0 to its maximum value (when *U* = 0),
where *A*(*t*) is the cubic solution above, and *U* is the following
(replace *A* with the cubic solution above and replace *f* with *t*):

🟥 $$U = \left(pk - kA\left(1 - \frac{1}{n}\right) + (p-k-H)\,f - f^2\right)\left(\frac{nHRf\,(Q+f) + ((mL-nR)\,f^2 + ((mL-R)\,k + (L-nR)\,Q)\,f + (L-R)\,Qk)\,A}{(nHRf\,(k+f) + ((mL-nR)\,f^2 + ((2mL - (n+1)\,R)\,f + (mL-R)\,k)\,k)\,A)\,f}\right)$$

☒ $$u = R\,\frac{\dfrac{q - \left(\frac{k}{nf} + 1\right)A}{L\,(k+f)}\left(\left(\frac{H}{A} - 1\right)nf - k\right) - 1}{1 + \dfrac{mf}{Q}}$$

☒ $$a = \frac{Hf - \left(\frac{k}{n} + f\right)A}{k+f}$$

☒ $$b = \frac{kA}{nf}$$

☒ $$c = \frac{p - f - \dfrac{kA\left(1 - \dfrac{1}{n}\right) + Hf}{k + f}}{1 + \dfrac{m}{R}\dfrac{L\,(k + f)}{\left(\dfrac{H}{A} - 1\right)nf - k}}$$

☒ $$d = \frac{q - \dfrac{L\,(k + f)}{\left(\dfrac{H}{A} - 1\right)nf - k} - \left(\dfrac{k}{nf} + 1\right)A}{1 + \dfrac{mf}{Q}}$$

☒ $$B = \frac{q - \dfrac{L\,(k + f)}{\left(\dfrac{H}{A} - 1\right)nf - k} - \left(\dfrac{k}{nf} + 1\right)A}{\dfrac{Q}{mf} + 1}$$

☒ $$h = \frac{H - \left(\dfrac{k}{nf} + 1\right)A}{1 + \dfrac{f}{k}}$$

## B. Contents:

An expression for total occupancy (= $A + a$, and $A + b$) as a function of $p$.
Graphing of total occupancy vs. $p$ based on that.

The maximum value of each total [bound protein] as $p$ goes to infinity:
for protein1, this is $H$, because as $f$ goes to infinity, all of the DNA becomes saturated with protein1 ($h$ goes to 0, $a$ becomes $H - A$);
for protein2, this is the same as the max. value of $A$,
since all single protein2–DNA complex is driven into the ternary complex (that is, $b$ goes to zero).

Variables used:
$h$ = free [labeled DNA] ("hot" probe)
$f$ = [free protein1]
$g$ = [free protein2]
$a$ = [DNA – protein1 complex]
$b$ = [DNA – protein2 complex]
$A$ = [ternary complex]
$H$ = total [labeled DNA], which includes both bound ($a$+$b$+$A$) and free ($h$)
$p$ = total [protein1], which includes both bound ($a + A$) and free ($f$)
$q$ = total [protein1], which includes both bound ($b + A$) and free ($g$)
$k$ = dissociation constant of $a$ into $h$ and $f$
$L$ = dissociation constant of $b$ into $h$ and $g$
$n$ = cooperativity factor
$r$ = total [bound protein1] = $a + A$
$s$ = total [bound protein2] = $b + A$

Governing equations:

equation 0:
$H = h + a + b + A$

equation 1:
☒ $L = \frac{gh}{b} = \frac{gh}{s - A}$

equation 2:
☒ $k = \frac{fh}{a}$

equation 3:
☒ $H = h + a + s$

equation 4:
☒ $q = g + s$

equation 5:

☒ $$n = \frac{Ah}{ab} = \frac{Ah}{a(s-A)}$$

Summary of derivation of total [bound protein2] ($s = b + A$) as a function of total [protein1] ($p$):

Eliminate $g$ using equations 1 and 4 to give equation 14;
eliminate $h$ using equations 14 and 3 to give equation 134;
eliminate $h$ using equations 14 and 5 to give equation 145;
eliminate $A$ using equations 134 and 145 to give equation 1345, and solve this for $a$.
This is $a$ in terms of $L, H, q, n$, and $s$.

Solve equation 145 for $A$ in terms of $a$;
use this to get an expression for $A$ from equation 1345.
This is $A$ in terms of $L, H, q, n$, and $s$.

Eliminate $h$ using equations 14 and 2 to give equation 124;
solve this for $f$;
substitute the expressions derived above for $a$ and $A$ into this to get equation 12345.
This is $f$ in terms of $k, L, H, q, n,$ and $s$.

Adding the above expressions for $f, A,$ and $a$
gives $p$ in terms of $k, L, H, q, n$, and $s$.

$s$ can then be graphed as a function of $p$ by
substituting $x$ for $p$ and $y$ for $s$ in that expression.

Total [bound protein2] ($s = b + A$) as a function of total [protein1] ($p$):

☒ $$p = (Ls - (H-s)(q-s)) \left( \frac{k}{n(H-s)(q-s) - Ls} + \frac{\frac{n}{L} + \frac{1}{q-s}}{n-1} \right)$$

where $n \neq 1$.
(Note: if $n = 1$, protein1 and protein2 bind independently, so each of their total occupancies is independent of the concentration of the other.)

Substitute $y$ for $s$, and $x$ for $p$ to graph
**total [bound protein2] as a function of total [protein1]**:
$p$ (= $x$) vs. $s$ (= $b+A = y$):

🟥 $$x = \left( (Ly - (H-y)(q-y)) \left( \frac{k}{n(H-y)(q-y) - Ly} + \frac{\frac{n}{L} + \frac{1}{q-y}}{n-1} \right) \text{ if } Ly < n(H-y)(q-y) \right)$$

The constraint prevents drawing a solution where the first denominator is negative; this occurs where $n(H–s)(q–s)$, the first term in the first denominator, which = $n(h+a)g$, < $L(b+A) = Ls$. Since $hg=Lb$ and $nga=LA$, this means that $(n–1)Lb < 0$, which means that $n < 1$. So for $n > 1$, both denominators are positive, and the first factor is also positive, as can be seen from the following:
As above, $Ls = L(b+A)$ and $(H–s)(q–s) = (h+a)g = Lb+LA/n$. Therefore, $Ls–(H–s)(q–s) = LA–LA/n$, which is positive as long as $n > 1$.

Total [bound protein1] can be obtained implicitly by exchanging the roles of $p$ and $q$, which means $x$ becomes $q$, ($p$ becomes $q$), $q$ becomes $x$ ($q$ becomes $p$), and $L$ becomes $k$, $k$ becomes $L$.

☒ $$(kr - ((H-r)\ (p-r)))\left(\frac{L}{n\,(H-r)\ (p-r)\ -kr} + \frac{\frac{n}{k} + \frac{1}{p-r}}{n-1}\right) = q$$

**Total [bound protein1] as a function of total [protein1]**:

To graph $p$ (= $x$) vs. ($r = A{+}a = y$):

🟩 $$(ky - ((H-y)\ (x-y)))\left(\frac{L}{n\,(H-y)\ (x-y)\ -ky} + \frac{\frac{n}{k} + \frac{1}{x-y}}{n-1}\right) = (q \text{ if } n\,(H-y)\ (x-y) > ky)$$

Peacock and Jaynes, 2017

Fig. 4A. Using competition to simultaneously determine the single-ligand dissociation constant and [ligand]. Formulae are given for using curve fitting to find both of these as parameters from experimental data in which known amounts of specific, unlabeled substrate (such as a DNA oligo) compete with a known amount of specific, labeled substrate (such as a labeled oligonucleotide) for binding to a fixed amount of ligand (such as a DNA binding protein), and the amount of bound, labeled substrate is measured as competitor substrate is varied.

key for single-character notation used here:

| below | Peacock & Jaynes [1]* | description |
|---|---|---|
| $H$ | $[A]_\mathrm{T}$ | total concentration of labeled substrate |
| $h$ | $[A]$ | free concentration of labeled substrate |
| $U$ | $[U]_\mathrm{T}$ | total concentration of specific unlabeled substrate |
| $V$ | | total concentration of specific unlabeled substrate with same dissociation constant as that of labeled substrate |
| $u$ | $[U]$ | free concentration of specific unlabeled substrate |
| $p$ | $[a]_\mathrm{T}$ | total concentration of ligand |
| $f$ | $[a]$ | free concentration of ligand |
| $k$ | $K_A$ | dissociation constant of ligand from labeled complex |
| $q$ | | dissociation constant of ligand from unlabeled complex |
| $Q$ | | dissociation constant of ligand from non-specific unlabeled complex |
| $a$ | $[Aa]$ | concentration of labeled complex |
| $b$ | $[Ua]$ | concentration of specific unlabeled complex |
| $c$ | | concentration of non-specific unlabeled complex |
| $n$ | $n$ | cooperativity factor |

* also used in Fig. 4B below

Contents:

For a single ligand (e.g., protein) binding to a specific site on a labeled substrate (e.g., DNA) either with or without unlabeled substrate present:

1) the equation describing the relationship between total [unlabeled competitor substrate] (= $U$) and the [labeled complex] (= $a$), in terms of the total [ligand] (= $p$), the total [labeled substrate] (= $H$), and the dissociation constants ($q$ and $k$);

2) the equation describing the relationship between the total [ligand] (= $p$) and [labeled complex] (= $a$), in terms of the total [labeled substrate] (= $H$), and the dissociation constant $k$, and the expression which can be used for curve fitting to find both $p$ and $k$ from data points ($\Delta$, $a$), where $\Delta$ is the dilution factor for a stock solution;

3) the above expression for $U$, specialized to the case where unlabeled substrate has the same Kd as that of labeled substrate, in which case $U$ is changed to $V$;

4) Then, in order to get an initial estimate of $p$ and $k$ as a starting point for curve fitting, an equation for each of them independent of the other is given based on the initial value of $a$ without competitor ($Z = a$ when $U = 0$), along with a second value of $a$ for any $U \neq 0$;

**The main equations are summarized at the end.**

1) the equation describing the relationship between total [unlabeled competitor substrate] (= $U$) and the [labeled complex] (= $a$), in terms of the total [ligand] (= $p$), the total [labeled substrate] (= $H$), and the dissociation constants ($q$ and $k$):

Competitive Equilibrium binding reaction;
a single protein (or complex) binds to a single site on each DNA (labeled and unlabeled);

forms of labeled DNA:

| | |
|---|---|
| total, | $H$ |
| bound, | $a$ |
| free, | $h$ |

forms of unlabeled DNA:

| | | |
|---|---|---|
| total, | $U$ | ($U$ becomes $V$ when $q = k$.) |
| bound, | $b$ | |
| free, | $u$ | |

forms of protein:

| | |
|---|---|
| total, | $p$ |
| free, | $f$ |
| labeled complex, | $a$ |
| unlabeled complex, | b |

dissociation constants:

| | |
|---|---|
| labeled DNA, | $k = fh / a$ |
| unlabeled DNA, | $q = fu / b$ |

$p$ = total [protein], which includes both bound ($a + b$) and free ($f$);
$H$ = total [labeled DNA], which includes both bound ($a$) and free ($h$);
$U$ = total [unlabeled DNA], which includes both bound ($b$) and free ($u$);

$Z$ = ($a$ when $U = 0$) = [labeled complex] without unlabeled competitor DNA;

$V$ = ($U$ when $q = k$), the [unlabeled competitor] when it has the same dissociation constant as the labeled DNA

🟦 $h + a = H$

Labeled DNA equilibrium equation:

🟪 $$k = \frac{fh}{a} = \frac{f(H-a)}{a} = f\left(\frac{H}{a} - 1\right)$$

☒ $$f = \frac{k}{\frac{H}{a} - 1}$$

🟨 $$p = f + a + b = \frac{k}{\frac{H}{a} - 1} + a + b$$

☒ $$b = p - a - \frac{k}{\frac{H}{a} - 1}$$

Total protein in the absence of competitor ($U = 0 = b$ , so $a$ becomes $Z$) as a function of $k$ and measurable quantities:

🟨 $$p = f + Z = \frac{k}{\frac{H}{Z} - 1} + Z = Z\left(\frac{k}{H - Z} + 1\right)$$

Solving for $k$ in terms of $p$, $H$, and $Z$ (however, since the fraction of $p$ that is active is unknown, we should use competition data to determine $p$ and $k$, see below):

🟪 $$k = (p - Z)\left(\frac{H}{Z} - 1\right) = \left(\frac{p}{Z} - 1\right)(H - Z)$$

Unlabeled DNA equilibrium equation:

☒ $$q = \frac{fu}{b}$$

Total unlabeled DNA = unlabeled complex ($b$) + free unlabeled DNA ($u = bq / f$);  substituting for $f$ from above in the second step:

☒ $$U = b + u = b\left(1 + \frac{q}{f}\right) = b\left(1 + \frac{q}{k}\left(\frac{H}{a} - 1\right)\right)$$

Substituting for $b$ from above:

☒ $$U = \left(p - a - \frac{k}{\frac{H}{a} - 1}\right)\left(1 + \frac{q}{k}\left(\frac{H}{a} - 1\right)\right)$$

"Transferring" the term ($H/a$ – 1) from the right factor to the left factor gives:
total [competitor DNA] as a function of total protein ($p$), total [labeled DNA] ($H$), the
[labeled DNA bound] ($a$), and the dissociation constants of the unlabeled ($q$) and labeled ($k$) complexes:

🟥 $$U = \left((p - a)\left(\frac{H}{a} - 1\right) - k\right)\left(\frac{1}{\frac{H}{a} - 1} + \frac{q}{k}\right)$$

2) the equation describing the relationship between the total [ligand] (= $p$) and [labeled complex] (= $a$), in terms of the total [labeled substrate] (= $H$), and the dissociation constant $k$, and the expression which can be used for curve fitting to find both $p$ and $k$ from data points ($\Delta$, $a$), where $\Delta$ is the dilution factor for a stock solution:
When $U = 0$ above, the first factor = 0, and we have the basic relationship between $p$, $a$, $H$, and $k$ without unlabeled competitor:

🟦 $$0 = (p - a)\left(\frac{H}{a} - 1\right) - k$$

🟥 $$p = a\left(\frac{k}{H - a} + 1\right)$$

We can, in principle, use this to find both $p$ and $k$ from data points ($\Delta$, $a$), where $\Delta$ is the dilution factor for a stock solution of concentration $P$, so:

⬛ $$\frac{P}{\Delta} = a\left(\frac{k}{H - a} + 1\right)$$

For curve fitting, we can thus use:

⬜ $$\Delta = \frac{P}{a\left(\frac{k}{H - a} + 1\right)}$$

3) The above expression for *U*, specialized to the case where unlabeled substrate has the same Kd as that of labeled substrate, in which case *U* is changed to *V*:

To find the dissociation constant *k*, we can use unlabeled competitor oligo of the same sequence as the labeled probe.
When *q* = *k* (competitor DNA has same dissociation constant as labeled DNA), *U* becomes *V*, and *q*/*k* becomes 1, so the above equation for *U* becomes:

☒ $$V = \left( (p-a)\left(\frac{H}{a} - 1\right) - k \right) \left( \frac{1}{\frac{H}{a} - 1} + 1 \right)$$

which simplifies to:

🟦 $$V = p\frac{H}{a} - H - \frac{k}{1 - \frac{a}{H}}$$

or, more compactly,

🟩 $$\frac{V}{H} = \frac{p}{a} - 1 - \frac{k}{H - a}$$

The above can be used in a high-throughput analysis to determine the relative affinities of related binding sites, where the highest affinity site is tagged, and measurements are made using a panel of untagged sites of how well they compete for binding to a fixed amount of protein, as described in Hallikas, et al., 2006 [6]. In that work, an approximate expression is given for relative affinity that does not have the correct limit behavior when the affinity approaches that of the tagged oligo. One remedy for this is to use the exact expression given below for *q*/*k*. Another is to use the approximation for *q*/*k* given below, which is both simpler than that in Hallikas et al. [6] and has the correct limit behaviors.
Let *Z* = [labeled complex] without competitor (or in the presence of a reference amount of a non-specific competitor, in which case the final expression, given below, for the ratio of dissociation constants is the same; derivation available on request),
*å* = [labeled complex] in the presence of a reference concentration (*V*) of untagged site oligo with the same sequence as the tagged site, and
*a* = [labeled complex] in the presence of the reference concentration (*U*) of untagged site oligo with the experimental sequence.
From above, the expression for *p* in terms of *Z*, which we will use to eliminate *p* from the equation, is:

🟦 $$p = \frac{k}{\frac{H}{Z} - 1} + Z$$

Equating the above expressions for *U* and *V* (because the same amount of the two competitors is used in parallel experiments):

🟩 $$U = \left( p - a - \frac{k}{\frac{H}{a} - 1} \right) \left( 1 + \frac{q}{k}\left(\frac{H}{a} - 1\right) \right) = V = p\frac{H}{\mathring{a}} - H - \frac{k}{1 - \frac{\mathring{a}}{H}} = \left( p - \mathring{a} - \frac{k}{\frac{H}{\mathring{a}} - 1} \right) \frac{H}{\mathring{a}}$$

And substituting for *p* in terms of *Z, k*, and *H*:

$$\left(\frac{k}{\frac{H}{Z}-1}+Z-a-\frac{k}{\frac{H}{a}-1}\right)\left(1+\frac{q}{k}\left(\frac{H}{a}-1\right)\right) = \left(\frac{k}{\frac{H}{Z}-1}+Z-\mathring{a}-\frac{k}{\frac{H}{\mathring{a}}-1}\right)\frac{H}{\mathring{a}}$$

Solving this for *q*/*k* gives:

$$\frac{q}{k} = \frac{\dfrac{\left(\frac{k}{\frac{H}{Z}-1}+Z-\mathring{a}-\frac{k}{\frac{H}{\mathring{a}}-1}\right)\frac{H}{\mathring{a}}}{\frac{k}{\frac{H}{Z}-1}+Z-a-\frac{k}{\frac{H}{a}-1}}-1}{\frac{H}{a}-1}$$

If the experiment is done under conditions where *Z* << *H*, that is, where only a small fraction of the tagged oligo is bound without competitor, then we can approximate this expression by the following (dropping the 1 from *H*/*a*–1, *H*/*å*–1, and *H*/*Z*–1, and replacing 1–*å*/*H* with 1);
*q*/*k* ~ :

$$\frac{\frac{Z}{\mathring{a}}(k+H)-k-H}{\frac{Z}{a}(k+H)-k-H}-\frac{a}{H} = \frac{\left(\frac{Z}{\mathring{a}}-1\right)(k+H)}{\left(\frac{Z}{a}-1\right)(k+H)}$$

where the *a*/*H* was dropped in the last step because it is assumed to be << 1 < *q*/*k* (because *k* is assumed to be the Kd of our highest affinity site).
So, *q*/*k* ~ :

$$\frac{\frac{Z}{\mathring{a}}-1}{\frac{Z}{a}-1}$$

4) In order to get an initial estimate of $p$ and $k$ as a starting point for curve fitting, an equation for each of them independent of the other is derived based on the initial value of $a$ without competitor ($Z = a$ when $U = 0$), along with a second value of $a$ for any $U \neq 0$:

Substituting for $p$ in terms of $k$, $H$, and $Z$ from above, we get an expression that can be solved for $k$ in terms of measurable quantities:

$$V = \left(\frac{k}{\frac{H}{Z} - 1} + Z\right)\frac{H}{a} - H - \frac{k}{1 - \frac{a}{H}}$$

Rearranging:

$$V = \left(\frac{Z}{a} - 1\right)\left(\frac{k}{\left(1 - \frac{Z}{H}\right)\left(1 - \frac{a}{H}\right)} + H\right)$$

solving this for **k in terms of measurable quantities**:

$$k = \left(\frac{V}{\frac{Z}{a} - 1} - H\right)\left(1 - \frac{Z}{H}\right)\left(1 - \frac{a}{H}\right)$$

Now, we can get an expression for $p$ (the total [active protein]) by substituting this expression for $k$ into the expression for $p$ in terms of $k$, $H$, and $Z$ from above:

$$p = \frac{k}{\frac{H}{Z} - 1} + Z$$

$$p = \frac{V\frac{Z}{H}\left(1 - \frac{a}{H}\right)}{\frac{Z}{a} - 1} - Z\left(1 - \frac{a}{H}\right) + Z = \frac{V\frac{Za}{H}\left(1 - \frac{a}{H}\right)}{Z - a} + \frac{Za}{H}$$

This is ***p* (total active protein concentration)** in terms of the [labeled DNA] bound in the absence of competitor ($Z$), the total concentration of labeled DNA ($H$), and the concentration of competitor with the same dissociation constant as the labeled DNA ($V$) that is required to reduce the [labeled DNA] bound from its initial value ($Z$) to $a$:

$$p = \frac{Za}{H}\left(\frac{V\left(1 - \frac{a}{H}\right)}{Z - a} + 1\right)$$

Summary:

For curve fitting, the above equations for $p$ and $k$ can be used to estimate their values from the two data points $(a, V)$ and $(Z, 0)$, and these then used to optimize $p$ and $k$ for all data points ($a$ and $V$) in:

🟩 $$\frac{V}{H} = \frac{p}{a} - 1 - \frac{k}{H - a}$$

In terms of $a$, $Z$, and $H$, the formulae for $k$ and $p$ take the forms:

🟪 $$k = \left(\frac{V}{\frac{Z}{a} - 1} - H\right)\left(1 - \frac{Z}{H}\right)\left(1 - \frac{a}{H}\right)$$

🟨 $$p = \frac{Za}{H}\left(\frac{V\left(1 - \frac{a}{H}\right)}{Z - a} + 1\right)$$

The forms that are easiest to compare to those derived for the case where non-specific competitor is also present (Fig. 4C, below) are as follows:

🟩 $$V = p\left(\frac{H}{a}\right) - a\left(\frac{H}{a}\right) - \frac{k\left(\frac{H}{a}\right)}{\frac{H}{a} - 1} = \left(p - a - \frac{k}{\frac{H}{a} - 1}\right)\frac{H}{a}$$

With $p$ substituted for its value that is necessary to get any given $Z$:

🟨 $$p = \frac{k}{\frac{H}{Z} - 1} + Z$$

we get $V$ in terms of $Z$, $a$, $H$, and $k$:

🟦 $$V = \left(\frac{k}{\frac{H}{Z} - 1} + Z - a - \frac{k}{\frac{H}{a} - 1}\right)\frac{H}{a}$$

The expression which can be used for curve fitting to find both $p$ (initial value = $P$) and $k$ from data points $(\Delta, a)$, where $\Delta$ is the dilution factor for a stock solution of concentration $P$:

🟥 $$\Delta = \frac{P}{a\left(\frac{k}{H - a} + 1\right)}$$

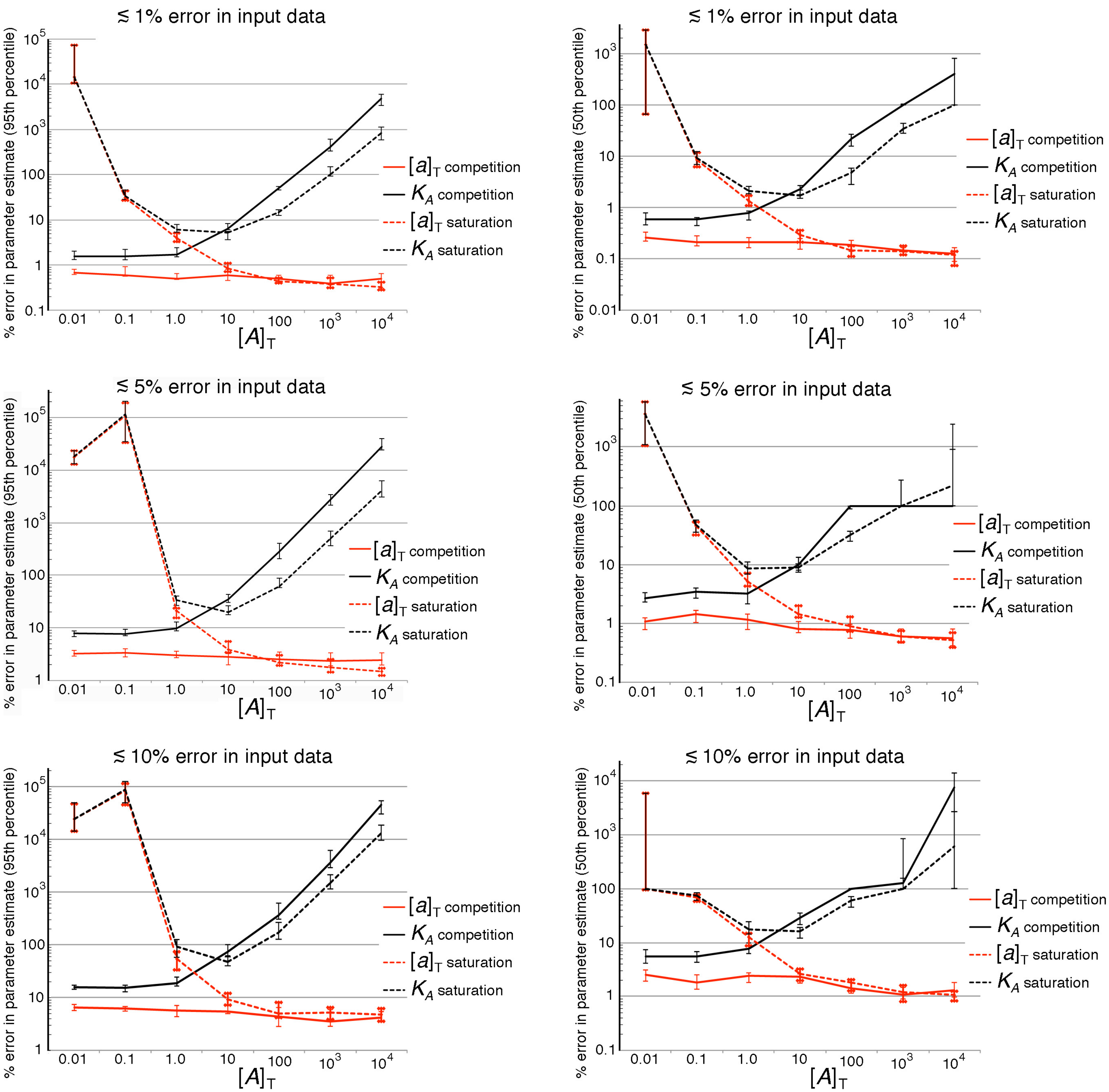


**Fig. 4B:** Performance of competition and saturation binding methods for simultaneously finding $[a]_T$ and $K_A$ with different input errors. Monte Carlo analysis (100 runs for each data point) of the effectiveness of curve fitting to find $[a]_T$ and $K_A$ as parameters was run with 15-point data sets at 7 different $[A]_T$ using either competition (fixed $[a]_T$, varying $[U_A]_T$) or standard saturation binding (varying $[a]_T$, no competitor). Data sets were generated by introducing random errors into calculated values of $[Aa]$ (ligand-substrate complex). These input values were randomly chosen from a normal distribution such that 95% of the errors were within the specified % of the actual value: 1% for the top two graphs (standard deviation = 0.5%, mean error = 0.40%, median error = 0.34%), 5% for the middle two (standard deviation = 2.5%, mean error = 2.0%, median error = 1.7%), and 10% for the bottom two (standard deviation = 5%, mean error = 4.0%, median error = 3.4%). Percent errors are shown in the values found for each parameter ($[a]_T$ and $K_A$) using least-squares curve fitting. These errors were ranked by increasing absolute value and plotted: either the 95th largest (left column, with errors bars extending from the 90th to the 99th largest) or the 50th largest (right column, with error bars extending from the 40th to the 60th largest). In each case, the error bars represent a 95% confidence interval for the true value of the specified error percentile, based on standard statistical analysis. Note that the best estimate for $K_A$ is provided by the competition method at low $[A]_T$, which simultaneously provides a precise estimate for $[a]_T$.

# Detailed methods for Fig. 4B

## Competition Binding

Competition binding simulations used a system of labeled substrate $A$ (e.g., a DNA oligo), identical (but unlabelled) competitor substrate $U$, and ligand $a$ (e.g., protein) forming complexes $Aa$ and $Ua$:

$a + U + A \leftrightarrow Ua + Aa$

Experiments were simulated using a statistical model of the form,

$\mathbf{Aa} = F(\mathbf{U_T};\ K_A, a_T, A_T) + \boldsymbol{\epsilon}$

where

- $\mathbf{Aa}$ is the list (i.e., vector) of concentrations of complex $Aa$, measured as the dependent variable for each value of $U_T$.
- $\mathbf{U_T}$ is the list of total concentrations of $U$, and is the independent variable in the experiment. $\mathbf{U_T}$ consists of 15 points evenly distributed from 0 to a value of $U_T$ such that $Aa$ is reduced to 1/3 its initial concentration, when $U_T = 0$.
- $A_T$ is the total concentration of substrate $A$. Seven values (0.01, 0.1, 1, 10, 100, 1000 and 10000) were tested. For each value of $A_T$, three values of $e$ (see below) were tested, yielding a total of 21 simulated experiment conditions.
- $K_A$ is the equilibrium constant of ligand $a$ binding either $A$ or $U$. Since all parameters have units of concentration, $K_A$ can always be normalized to 1.
- $a_T$ is the total concentration of ligand $a$, and is set such that $Aa = A_T/2$ when $U_T = 0$.
- F is a function derived from the governing equations of the system which relates each component of $\mathbf{Aa}$ to the corresponding component of the independent variable $\mathbf{U_T}$ and the parameters $K_A, a_T, A_T$:

$$F = \frac{a_T A_T + A_T^2 + A_T K_A + A_T U_T - \sqrt{4 a_T A_T^2(-A_T - U_T) + (a_T A_T + A_T^2 + A_T K_A + A_T U_T)^2}}{2(A_T + U_T)}.$$

- $\boldsymbol{\epsilon}$ is the list of simulated measurement errors of $\mathbf{Aa}$. Errors were drawn from a normal distribution with variance proportional to the concentration of $Aa$,

$\epsilon \sim N(0, e \cdot Aa/2)$

- where $e$ is a percent error, tested at values of 1%, 5% and 10%. In this way, the interval $Aa \pm (e \cdot Aa)$ will have a radius of two standard deviations centered at $Aa$, and therefore contains 95% of observations. Thus, we interpret “10 with a 10% error” to mean 95% of measurements fall between 9.9 and 10.1.

For each of the 21 simulated experiment conditions, defined as combinations of $e$ and $A_T$, 100 such experiments were simulated. Nonlinear least-squares regression, weighted by the inverse of the variance at each point $(4/(e \cdot \mathbf{Aa})^2)$, was used to produce estimates $\hat{K}_A$ and $\hat{a}_T$ from the 15 points $(U_T, Aa)$, with $A_T$ assumed as a given. The weighting is necessitated by the non-constant variance of the error at each data point; i.e., heteroskedasticity. With 100 simulated experiments for each set of conditions, we compared the performance of the simulated experiments in determining $K_A$ and $a_T$, as detailed in Analysis.

For complete details of this methodology, see `Fig_4B_MC_competition_model.nb` within the Supplemental_Mathematica_notebooks of Peacock and Jaynes [1].

## Saturation Binding

Saturation binding simulations used the same system as the competition binding experiments, but without the competitor $U$,

$a + A \leftrightarrow Aa$

Experiments were simulated using a similar model,

$\mathbf{Aa} = G(\mathbf{V};\ K_A, a_T, A_T) + \boldsymbol{\epsilon}$

However, rather than an independent variable of $\mathbf{U_T}$, we use $\mathbf{V}$, a list of unitless scalars representing a dilution/concentration series of $a_T$. We treat $a_T$ as an unknown initial concentration, to be found in the experiment. The unknown concentration stock is then diluted or concentrated according to $\mathbf{V}$ to perform the experiment. As in the competition experiment, $\mathbf{V}$ consists of 15 points. However, these points are evenly distributed from $V_{\max}/15$ to $V_{\max}$, where $V_{\max}$ represents a concentration of $a$ such that $Aa = 0.9A_T$ (i.e., 90% saturation is achieved). Rather than beginning at zero, which would give zero complex and any signal would be subtracted as noise, we begin at the next increment, $V_{\max}/15$.

The following function, which is derived from the governing equations of the system, relates each componenet of $\mathbf{Aa}$ to the corresponding component of the independent variable $\mathbf{V}$ and the parameters $K_A, a_T, A_T$:

$$G = \left(Va_T + A_T + K_A - \sqrt{(-Va_T - A_T - K_A)^2 - 4Va_TA_T}\right)/2$$

The remaining parameters $a_T, K_A$ and $A_T$ were set to the same values as those used in the competition binding experiments. The simulated errors were produced in the same way as well, using the same values of $e$.

This design allows direct comparison of the two techniques in determining the unknown parameters $K_A$ and $a_T$. As with the competition binding experiments, nonlinear least-squares regression, with the same weighting scheme, was used to estimate $K_A$ and $a_T$ from the 15 data points, here of the form $(V, Aa)$. Again, we assume $A_T$ is a known parameter of the experiment, and we produce 100 simulated experiments for each of the 21 experiment conditions, yielding 100 estimates $\hat{K}_A$ and $\hat{a}_T$ for each condition.

For complete details of this experiment see `Fig_4B_MC_saturation_model.nb` within the Supplemental_Mathematica_notebooks of Peacock and Jaynes [1].

## Analysis

The competition binding and saturation binding experiments each generated 100 estimates of $K_A$ and $a_T$ for the 21 experiment conditions. In each case, the 100 estimates are treated as an empirical distribution of the parameter estimates. To relate this to common experimental metrics, this distribution is normalized to represent an absolute percent error. For example, we normalize each estimate $\hat{K}_A$ of the actual value $K_A$ as follows,

$$100\% \left| \frac{K_A - \hat{K}_A}{k_A} \right|$$

Note that this distribution is strictly positive, starting at zero with a tail to the right. To compare the normalized parameter estimate distributions between models and conditions, we consider the 95th percentile of the distribution. This value is easily interpreted as the maximum percent error in 95% of experiments, and is easily estimated by the 95th order statistic of the 100 estimates.

A 95% confidence interval for the 95th percentile is produced by considering a symmetric interval centered on the 95th order statistic. This interval is expanded until the expected probability of the 95th percentile lying in the interval is greater than 0.95. In this case, the interval spans from the 91st to the 99th order statistic,

$$0.9659 = \sum_{j=91}^{99} \binom{100}{j} (0.95)^j (1 - 0.95)^{n-j}$$

Thus, the 91st and 99th order statistics conservatively bound a 95% confidence interval (96.6% more exactly) for the 95th percentile.

An analogous procedure was followed for the 50th percentile analysis, illustrated in the right column of Fig. 4B.

Fig. 4C. Using competition to simultaneously determine the single-ligand dissociation constant and [ligand] in the presence of non-specific competitor substrate. Formulae are given for using curve fitting to find both of these as parameters from experimental data in which known amounts of specific, unlabeled substrate (such as a DNA oligo) compete with a known amount of specific, labeled substrate (such as a labeled DNA oligo) and a known amount of non-specific, unlabeled substrate (such as poly-dA/dT) for binding to a fixed amount of ligand (such as a DNA binding protein), and the amount of bound, labeled substrate is measured as competitor substrate is varied.

key for single-character notation used here:

| below | Peacock & Jaynes [1] | description |
|---|---|---|
| $H$ | $[A]_T$ | total concentration of labeled substrate |
| $h$ | $[A]$ | free concentration of labeled substrate |
| $U$ | $[U]_T$ | total concentration of specific unlabeled substrate |
| $V$ | | total concentration of specific unlabeled substrate with same dissociation constant as that of labeled substrate |
| $u$ | $[U]$ | free concentration of specific unlabeled substrate |
| $p$ | $[a]_T$ | total concentration of ligand |
| $f$ | $[a]$ | free concentration of ligand |
| $k$ | $K_A$ | dissociation constant of ligand from labeled complex |
| $q$ | | dissociation constant of ligand from unlabeled complex |
| $Q$ | | dissociation constant of ligand from non-specific unlabeled complex |
| $a$ | $[Aa]$ | concentration of labeled complex |
| $b$ | $[Ua]$ | concentration of specific unlabeled complex |
| $c$ | | concentration of non-specific unlabeled complex |
| $n$ | $n$ | cooperativity factor |

Contents:
an expression useful for curve fitting to find $p$ and $k$ as parameters, by adding known amounts of unlabeled substrate having the same Kd as that of labeled substrate, when non-specific competitor substrate is included in the reaction, along with a description of how to analyze the data;
the formula to find $n$ from curve fitting, after the individual ligand concentrations ($p$ and $q$) and Kd's ($k$ and $L$) have been determined from single-ligand experiments, derived in Fig. 2A, is given at the end:

Competitive equilibrium binding reaction; Assumptions:
a single protein (or complex) binds to a single site on each "specific" DNA;
non-specific competitor DNA is also present: its concentration ($D$) is based on the total # of
potential binding sites, approximately one per bp if [protein] is far below saturation, ignoring end effects;

$H$ = total [labeled DNA], which includes both bound ($a$) and free ($h$);
$U$ = total [unlabeled DNA, specific], which includes both bound ($b$) and free ($u$);
$D$ = total [unlabeled DNA, non-specific], which includes both bound ($c$) and free ($d$);
$k, q, Q$ = the dissociation constants of the protein with probe, specific competitor, and non-specific competitor, respectively;
$p$ = total [protein], which includes both bound ($a + b + c$) and free ($f$);

$D$, $H$, and $U$ (which becomes $V$ when it has the same Kd for ligand as does $H$) are assumed to be known (or directly measurable);
$p, k, q$, and $Q$ are to be determined.
The dissociation constant of the labeled DNA – protein complex is:

☒ $$\frac{fh}{a} = k$$

From which we can get [free protein], since $h = H - a$:

$$\frac{k}{\frac{H}{a} - 1} = f$$

total protein:

$$\boxtimes\quad f + a + b + c \;=\; p \;=\; \frac{k}{\frac{H}{a} - 1} + a + b + c$$

unlabeled complex (specific):

$$\boxtimes\quad p - a - \frac{k}{\frac{H}{a} - 1} - c \;=\; b$$

dissociation constant of specific competitor (since $u = U - b$):

$$\boxtimes\quad \frac{fu}{b} \;=\; q \;=\; f\left(\frac{U}{b} - 1\right)$$

dissociation constant of non-specific competitor (since $d = D - c$):

$$\boxtimes\quad \frac{fd}{c} \;=\; Q \;=\; f\left(\frac{D}{c} - 1\right)$$

non-specific competitor complex as a function of measurable quantities
($k/Q$ becomes "measurable" below), substituting for $f$ from above:

$$\boxtimes\quad \frac{D}{\frac{Q}{f} + 1} \;=\; c \;=\; \frac{D\left(\frac{k}{Q}\right)}{\left(\frac{H}{a} - 1\right) + \frac{k}{Q}} \;=\; \frac{D}{\left(\frac{H}{a} - 1\right)\frac{Q}{k} + 1}$$

specific competitor complex as a function of "measurable" quantities,
substituting for $c$ in the expression for $p$ from above:

$$\boxtimes\quad p - a - \frac{k}{\frac{H}{a} - 1} - \frac{D}{\left(\frac{H}{a} - 1\right)\frac{Q}{k} + 1} \;=\; b$$

Note that the above expression for $b$ is independent of its dissociation constant $q$.
(The amount of specific competitor bound when it has reduced the amount of probe bound
from its initial value without competitor to $a$ is independent of its dissociation constant.
This makes sense because this corresponds to how much protein it has removed from solution,
which is how it affects the amount of probe bound.)
Therefore, in a competition experiment, for a given value of $a$, $b$ is fixed (independent of $q$),
and $U/b$ varies rather simply with $q/k$:

$$\boxtimes\quad b\left(\frac{q}{f} + 1\right) \;=\; U \;=\; b\left(\frac{q}{k}\left(\frac{H}{a} - 1\right) + 1\right) \;=\; \left(p - a - \frac{k}{\frac{H}{a} - 1} - \frac{D}{\left(\frac{H}{a} - 1\right)\frac{Q}{k} + 1}\right)\left(\left(\frac{H}{a} - 1\right)\frac{q}{k} + 1\right)$$

Therefore, at the same value of $a$ for different competitors, $b$ is the same, while $U$ and $q$ are different.

The ratio of the amounts of competitors (*U* and *X*) of dissociation constants *q* and $\underset{,}{k}$ that are required to reduce the bound labeled DNA (*a*) to a given value is independent of *p, H, D* and *Q*:

☒ $$b\left(\frac{\underset{,}{k}}{k}\left(\frac{H}{a}-1\right)+1\right) = X$$

☒ $$b\left(\frac{q}{k}\left(\frac{H}{a}-1\right)+1\right) = U$$

So, the ratio of these amounts of competitor is

☒ $$\frac{U}{X} = \frac{\frac{q}{k}\left(\frac{H}{a}-1\right)+1}{\frac{\underset{,}{k}}{k}\left(\frac{H}{a}-1\right)+1}$$

This ratio has this form at all points along the two competition curves (*X* and *U* as a function of *a*).
It varies from its initial value to $q/\underset{,}{k}$ as a goes to 0.

The ratio depends on *k* still, which is unknown, but if $\underset{,}{k} = k$ (competitor oligo *V* is the same as probe, *H*), then the ratio is simpler:

☒ $$\frac{U}{V} = \frac{\frac{q}{k}\left(\frac{H}{a}-1\right)+1}{\left(\frac{H}{a}-1\right)+1} = \frac{\frac{q}{k}\left(\frac{H}{a}-1\right)+1}{\frac{H}{a}} = \frac{q}{k}\left(1-\frac{a}{H}\right)+\frac{a}{H}$$

Solving this for *q*/*k*:

☒ $$\frac{\frac{U}{V}-\frac{a}{H}}{1-\frac{a}{H}}$$

However, we can't fit the two curves independently to get parameters for comparison, because *b* varies with *a* in a complex way. We have to compare the two curves simultaneously.
From above,

☒ $$U = \left(p-a-\frac{k}{\frac{H}{a}-1}-\frac{D}{\left(\frac{H}{a}-1\right)\frac{Q}{k}+1}\right)\left(\left(\frac{H}{a}-1\right)\frac{q}{k}+1\right)$$

Now,
substituting *q* = *k* in the expression for *U* from above (*U* becomes *V* when *q* = *k*):

☒ $$V = b\left(\frac{k}{k}\left(\frac{H}{a}-1\right)+1\right) = b\frac{H}{a} = \left(p-a-\frac{k}{\frac{H}{a}-1}-\frac{D}{\left(\frac{H}{a}-1\right)\frac{Q}{k}+1}\right)\frac{H}{a}$$

If we use additional non-specific competitor as our competitor, then $q = Q$, and $U$ becomes $W$ ($W$ = [added competitor] when $q = Q$):

☒ $$W = b\left(\frac{Q}{k}\left(\frac{H}{a}-1\right)+1\right) = \left(p - a - \frac{k}{\frac{H}{a}-1} - \frac{D}{\left(\frac{H}{a}-1\right)\frac{Q}{k}+1}\right)\left(\left(\frac{H}{a}-1\right)\frac{Q}{k}+1\right)$$

☒ $$W = \left(p - a - \frac{k}{\frac{H}{a}-1}\right)\left(\left(\frac{H}{a}-1\right)\frac{Q}{k}+1\right) - D$$

**<u>The following works well:</u>**

If we use the $V$ equation to eliminate $k$ (changing $a$ to $\mathring{a}$ to distinguish the data sets), then we can get an expression for curve fitting using a combination of each data point ($\mathring{a}$,$V$) with each data point ($a$,$W$), and use these composite data points ($\mathring{a}$,$V$, $a$,$W$) to get the 2 parameters $p$ and $Q/k$:

☒ $$\left(p - \mathring{a} - \frac{k}{\frac{H}{\mathring{a}}-1} - \frac{D}{\left(\frac{H}{\mathring{a}}-1\right)\frac{Q}{k}+1}\right)\frac{H}{\mathring{a}} = V = \frac{pH}{\mathring{a}} - H - \frac{k}{1-\frac{\mathring{a}}{H}} - \frac{DH}{(H-\mathring{a})\frac{Q}{k}+\mathring{a}}$$

☒ $$\frac{k}{1-\frac{\mathring{a}}{H}} = \frac{pH}{\mathring{a}} - H - V - \frac{DH}{(H-\mathring{a})\frac{Q}{k}+\mathring{a}}$$

So, solving for $k$:

☒ $$k = \left(\frac{p}{\mathring{a}} - 1 - \frac{V}{H} - \frac{D}{(H-\mathring{a})\frac{Q}{k}+\mathring{a}}\right)(H-\mathring{a}) = \left(\frac{p}{\mathring{a}} - 1 - \frac{V}{H}\right)(H-\mathring{a}) - \frac{D}{\frac{Q}{k}+\frac{1}{\frac{H}{\mathring{a}}-1}}$$

Substituting this for $k$ in $W$ from above, we get the final curve fitting equation:

☒ $$W = \left(p - a - \frac{\left(\frac{p}{\mathring{a}} - 1 - \frac{V}{H}\right)(H-\mathring{a}) - \frac{D}{\frac{Q}{k}+\frac{1}{\frac{H}{\mathring{a}}-1}}}{\frac{H}{a}-1}\right)\left(\left(\frac{H}{a}-1\right)\frac{Q}{k}+1\right) - D$$

**<u>Once we have found *Q*/*k* from curve fitting with the above</u>**, we can use the following to get *k* and *p* (again) from the same data, (*å, V*); it generally works better to take only *Q*/*k* from the first curve fitting, and, using it, find *k* and *p* using the second curve fitting:

☒ $$\left(p - å - \frac{k}{\frac{H}{å} - 1} - \frac{D}{\left(\frac{H}{å} - 1\right)\frac{Q}{k} + 1}\right)\frac{H}{å} = V$$

**Once each of the individual ligand concentrations (*p* and *q*) and Kd's (*k* and *L*) have been determined**, we can use data on ternary complex formation as a function of one of the [ligand] (= *p*), at a fixed [the other ligand] (= *q*) and a fixed [labeled substrate] (= *H*) to find the cooperativity factor as a parameter, using the formula for *p* vs. *A* (= [ternary complex]) derived in Fig. 2A;
curve fit data points (*p, A*), given *H, L, k*, and *q*, to find *n* as the parameter:

☒ $$p = A + \frac{A}{n}\left(\frac{H + L - q + \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}}{2\left(q - A + \frac{A}{n}\right)}\right) + \frac{kA}{n}\left(\frac{H + L + q - 2A + \sqrt{(H + L - q)^2 + 4L\left(q - A + \frac{A}{n}\right)}}{2\left((H - A)\,(q - A) - \frac{LA}{n}\right)}\right)$$

## Section 5. Accurate Kd's and the cooperativity factor can be determined even when one Kd is too high to measure directly

**A.** Note: Section 5A and Fig. 5 use the terminology of Peacock and Jaynes [1] for substate, ligands, Kd's, and complexes (see Section 5B for a list). The methods described above and used in Peacock and Jaynes [1] require that each individual Kd be measurable. However, for some ternary complexes, one of the ligands is not observed to bind alone, even at concentrations much higher than those required for strong, cooperative binding. A classic example of this was described by Jin et al., 1999 [7], involving cooperative binding by the yeast transcription factors a1 and α2. Binding by α2 alone was seen, and addition of a1 gave much greater complex formation, suggestive of highly cooperative binding. Even at the highest concentrations tested, no binding by a1 alone was observed. We modeled binding by these proteins based on a rough quantitation of the published data. Our purpose was not to derive precise parameters for this particular case, but to illustrate and test our ability to obtain quantitative information in such cases more generally. Specifically, we show here that it is possible to obtain from binding data on the two visible complexes, $ABb$ and $AaBb$, via curve fitting, the Kd of the weakly binding protein ($K_A$), its concentration ($[a]_\mathrm{T}$), and $n$, knowing only the concentration and Kd of the other protein ($[b]_\mathrm{T}$ and $K_B$). The approach involves first determining $K_B$ and $[b]_\mathrm{T}$ (in this case for α2) by the method described above (or by any other method). Armed with these constants, we can conduct an experiment where $[b]_\mathrm{T}$ and $[AB]_\mathrm{T}$ are fixed, under conditions where $ABb$ is clearly visible and quantifiable without added ligand $a$. Ligand $a$ is then added at increasing concentrations, and the formation of $AaBb$, along with the simultaneous decrease in $[ABb]$, is quantified. The situation is modeled in Fig. 5A. The procedure "starts", as just described, where the curves all intersect the horizontal axis. This point represents the starting $[ABb]$ without added ligand $a$ (there is no $AaBb$). As $[a]_\mathrm{T}$ increases, we move up the curve of $[AaBb]$ vs. $[ABb]$ from right to left. The intersection of the curve with the vertical axis represents the theoretical maximum $[AaBb]$ when $[a]_\mathrm{T}$ becomes infinite, chasing all of the $ABb$ into $AaBb$, and so the $[ABb]$ goes to zero. This maximum $[AaBb]$ is sensitive to $n$, and the graph shows a family of curves with different $n$ values.

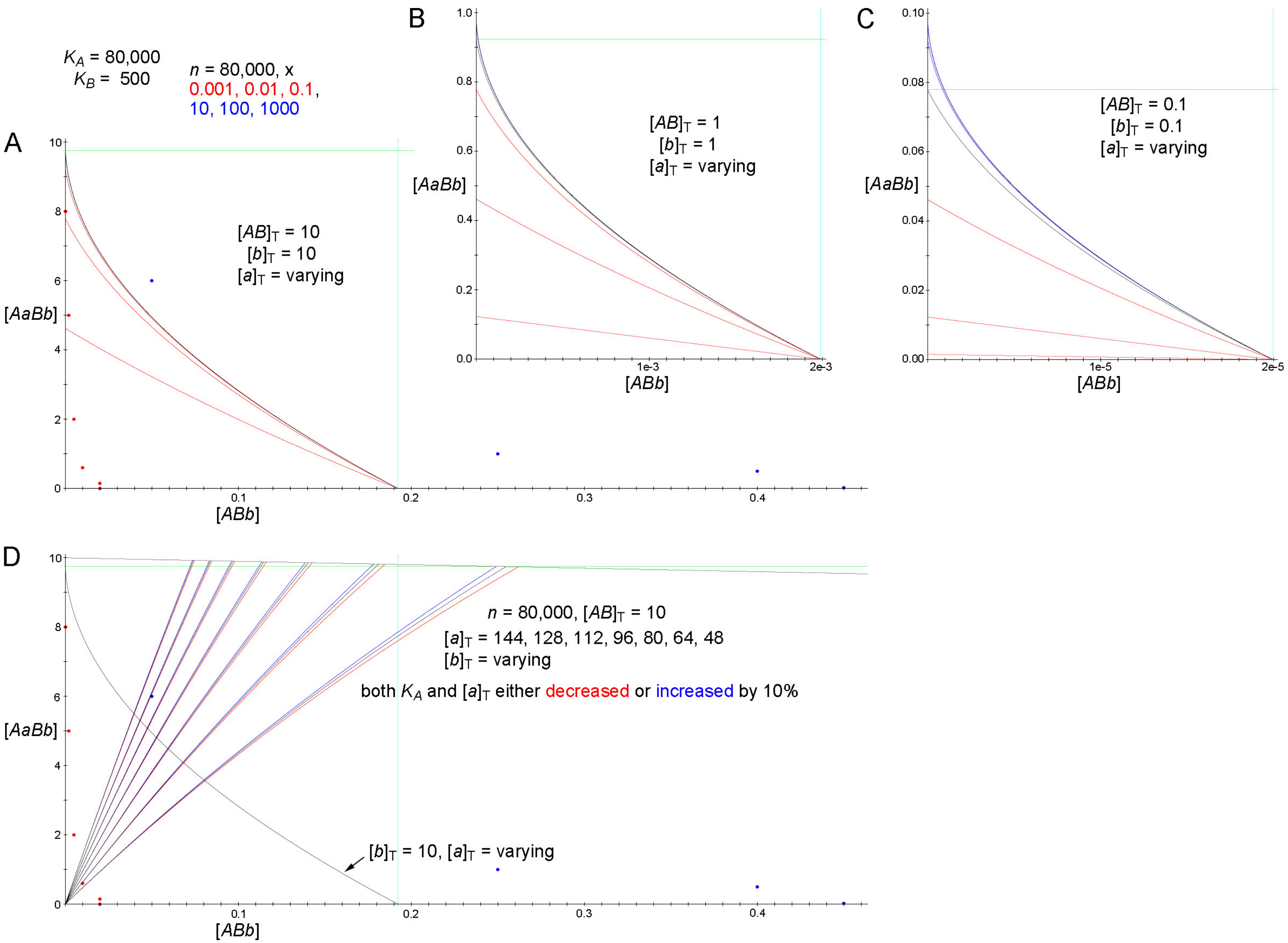


**Fig. 5. Graphs of equations used for curve fitting to find cooperativity factor and 2nd Kd when one protein binds weakly alone. A:** Family of curves of [*AaBb*] vs. [*ABb*] with different cooperativity factors. The concentration of ternary complex, [*AaBb*], is graphed as a function of increasing concentrations of the binary complex [*ABb*] containing the ligand that binds detectably on its own (*b*), with constant total amounts of both labeled binding site, $[AB]_T$, and *b*, $[b]_T$. Although not displayed on the graph, $[a]_T$ is the experimentally varying quantity, and it decreases from infinity at the vertical axis, to zero at the horizontal axis, which is the point of maximum [*ABb*]. Constants used are $[AB]_T$ = 10, $[b]_T$ = 10, $K_A$ = 80,000, and $K_B$ = 500, estimated from data in Jin et al., 1999 [7], as described in the Section 5A text. Curves are graphed for 7 different values of *n*, but the four with *n* >= 80,000 are indistinguishable (black) with these values of $[AB]_T$ and $[b]_T$. Only those where *n* is reduced by 1000x or 100x from 80,000 are clearly separated (red; the curve where *n* is reduced by 10x shows barely detectable separation). Filled circles, one set red, the other set blue, mark our estimated bracketing values for the published data points: the parameters were adjusted so that the black curve lies between these two sets of points. **B:** Family of curves of [*AaBb*] vs. [*ABb*] with the same set of cooperativity factors as in A, but with lower amounts of total binding site and total ligand *b*, allowing a minimum estimate for *n*. Here, $K_A$ and $K_B$, as well as the set of values for *n*, are the same as in A, but both $[AB]_T$ and $[b]_T$ are

reduced by a factor of 10 (now $[AB]_T = [b]_T = 1$). Note that the 3 lower curves (red, $n$ = 80, 800, and 8000) are better separated from the others, while the black curve is barely distinguishable from the 3 blue curves, which still lie on top of each other ($n$ >= 800,000). Thus, data represented by this black curve can provide a good minimum estimate for $n$ (within a factor of 10 or less, see Section 5A text).
**C:** Family of curves of $[AaBb]$ vs. $[ABb]$ with the same set of cooperativity factors as in A and B, but even lower amounts of total binding site and total ligand $b$, allowing a better estimate for $n$. Here, $K_A$ and $K_B$, as well as the values for $n$, are the same as in A, but both $[AB]_T$ and $[b]_T$ are now reduced by a factor of 100 ($[AB]_T = [b]_T = 0.1$). Note that the three lower curves (red), as well as the blue curves as a group ($n = 8 \times 10^5$, $8 \times 10^6$, $8 \times 10^7$), are well separated from the black curve ($n = 8 \times 10^4$). Thus, data represented by the black curve can now provide a reliable estimate for $n$, using the curve fitting method described in the Section 5A text.
**D:** Families of curves of $[AaBb]$ vs. $[ABb]$, now with varying $[b]_T$, and the indicated (constant) values of $[a]_T$ (black), plus (for reference) the black curve from A (which has constant $[b]_T$ and varying $[a]_T$). Each black curve is accompanied by two curves in which both $K_A$ and $[a]_T$ are either decreased (red) or increased (blue) by 10%. Although $[a]_T$ is as yet unknown, the method involves keeping track of dilution factors of it, relative to a reference concentration, which will be determined from the data. In those with the lower values of $[a]_T$ (toward the right), the red and blue curves are visibly separated from the bracketed black curve. This illustrates that curve fitting using such data sets can, in principle, give good estimates for both of the unknowns, $[a]_T$ and $K_A$, once a good estimate for $n$ is obtained from experiments with varying $[a]_T$, as illustrated in C. The dark gray curve across the top is $x + y = [AB]_T$, which is the limit of each curve at infinite $[b]_T$ (in that limit, all the $AaB$ is chased into $AaBb$, so $[ABb] + [AaBb] = [AB]_T$).

The colored dots on the graph (Fig. 5A,D) are our crude estimates of "bracketing" values for the relative $[AaBb]$ and $[ABb]$ in the published data (columns 2-7 of Fig. 5B of [7]). The red dots represent one extreme estimate and the blue dots the opposite extreme estimate, one of each color for each data point in the published figure (not all of the blue dots are in the range of this graph). The black curve is obtained by choosing values for the various constants so that it falls between these maximum and minimum estimates for all the data points. However, the value of $n$ chosen can only be a minimum estimate for $n$, because, as the graph shows, higher values for $n$ than that used for the black curve ($n = 8 \times 10^4$) also give curves that fit this criterion (these curves are actually indistinguishable from the black curve on this graph).

What can we do in such a case of high cooperativity, in order to get a better estimate of the actual value of $n$, and then go on to determine $K_A$ and $[a]_T$? The answer is illustrated in Fig. 5B and C. By reducing both the $[AB]_T$ and $[b]_T$ used in the experiment, the curves become sensitive to differences between larger $n$ values. Fig. 5B shows that when these are reduced by 10-fold, we now start to see separation between the curves for the chosen $n$ value and 10 times this value ($n = 8 \times 10^5$). We take this process further in Fig. 5C, which shows that reducing both $[AB]_T$ and $[b]_T$ by 100-fold allows us to distinguish $n = 8 \times 10^4$ from $n = 8 \times 10^5$, and we can also now just discern separation between the curves for the latter and for $n = 8 \times 10^6$.

We tested the ability of curve fitting to give accurate values for $n$ under these conditions, by modeling such experiments with these different values for $[AB]_{\mathrm{T}}$ and $[b]_{\mathrm{T}}$. We constructed data sets from these curves, and rounded the data to introduce random errors, then tested how well these data sets allowed curve fitting to recover the value for $n$. The effectiveness of the method is illustrated in section B (below). We found that for the situation graphed in Fig. 5C (100-fold reduced values for $[AB]_{\mathrm{T}}$ and $[b]_{\mathrm{T}}$), we could recover $n$ to within 10% using 28 data points rounded to 2 significant figures (and so containing random errors between 0.5% and 5%). These 28 data points were for 4 different values of $[b]_{\mathrm{T}}$ in combination with 7 different values of $[a]_{\mathrm{T}}$. A second such data set taken at the higher values for $[AB]_{\mathrm{T}}$ and $[b]_{\mathrm{T}}$ can then be used to find both $K_A$ and $[a]_{\mathrm{T}}$, as described below. The expressions used both to generate the graphs shown in Fig. 5 and for curve fitting are given in section 5B (below).

There are potential limitations in implementing this method to achieve such accuracy. One is the need to use high concentrations of ligand $a$ in order to reach the part of the binding curve that is most sensitive to $n$, and another is the requirement that low amounts of complexes be quantifiable. The former may be limited by the ability to purify enough ligand $a$ and/or to maintain its solubility under actual experimental conditions. We limited our data to concentrations of transcription factors that are within 10-fold of those commonly used in such DNA binding experiments. The latter limitation depends (for DNA binding studies) on both the specific activity of labeling of the oligo and the sensitivity of the detection methodology. Where these prove to be restrictive, it may not be possible to accurately measure very high $n$ values. In such extreme cases, it may be easier to define a lower limit for $n$ using its definition, given in Fig. 2A of Peacock and Jaynes [1]: $n = [AB]\,[AaBb] / [AaB]\,[ABb]$. For such cases, $[AaB]$ may be too low to quantify, but we can estimate its upper limit. This upper limit, combined with the other 3 quantities, gives a lower limit for $n$. However it may be found, a lower limit for $n$ can be useful, in combination with the value of $K_A$ still to be determined, in predicting cooperative binding behavior *in vivo*. This is particularly true when comparing related ligand combinations and binding sites, where relative behaviors are often key to understanding biological phenomena.

Once an estimate for $n$ has been found, we can determine $K_A$ and $[a]_{\mathrm{T}}$ using the following method. In the experimental approach described above, we vary $[a]_{\mathrm{T}}$, but do not yet know its actual value. For finding $n$, the expression used for curve fitting does not contain either $[a]_{\mathrm{T}}$ or $K_A$. However, we can use the same data to find these quantities, provided we keep track of how a reference $[a]_{\mathrm{T}}$ is diluted for each data point. We then use a different expression, containing all 3 of the original unknowns, $n$, $[a]_{\mathrm{T}}$, and $K_A$, for curve fitting to find the latter two. While it is theoretically possible to find all 3 unknowns simultaneously using this expression, we have found, through testing with a number of data sets, that the method works much more reliably with realistic data if we use the first expression to find $n$, then plug the $n$ value into the latter expression to find $K_A$ and $[a]_T$. Fig. 5D illustrates how the method works. The data used above to find $n$ consists of data points for $[AaBb]$ and $[ABb]$ at different values of both $[a]_T$ and $[b]_T$. Fig. 5A-C show only curves with varying $[a]_T$. Fig. 5D shows a family of 7 curves (plus variants) with varying $[b]_T$, each with a different, but constant, $[a]_T$. For reference, the black curve from Fig. 5A is also shown. The expression used to generate these curves is that used for curve fitting to find $K_A$ and $[a]_T$. Its dominant term contains the ratio $[a]_T / K_A$, which determines the approximate slope of these near straight-line curves. This ratio is relatively easy to determine, while separately determining $K_A$ and $[a]_T$ depends on the deviations of these curves from a straight line. The red and blue curves represent a simultaneous 10% change in $K_A$ and $[a]_T$ from their accompanying black curve. Where these are distinct, realistically precise data can be expected to provide the necessary distinction. We tested this, and found that with 28 data points, taken from these curves and then rounded to 2 significant figures, curve fitting gave $K_A$ and $[a]_T$ individually within 10%. However, this was assuming the exact value for $n$. In order to use curve fitting with data rounded to 2 significant digits to first recover $n$, and then $K_A$ and $[a]_T$, each within 10%, it was necessary to use data like that illustrated in Fig. 5C (with $[AB]_T = 0.1$ and $[b]_T$ in the same range) to find $n$ (within 4%), then a separate data set like that in Fig. 5D (with $[AB]_T = 10$ and $[b]_T$ in the same range) to recover $K_A$ and $[a]_T$. This is because there is a trade-off in determining $n$ and then $K_A$ and $[a]_T$: while data sets using lower $[AB]_T$ and $[b]_T$ give $n$ more accurately, higher $[AB]_T$ and $[b]_T$ give $K_A$ and $[a]_T$ more accurately. The reasons for this can be seen by examining the relevant equations, as described in section B (below). There, a general procedure is outlined for maximizing the precision obtainable for all 3 parameters, within experimental limitations. In section C (below), a summary of our curve fitting trials is given that illustrates the general principles involved.

In cases where the accuracy of $n$ is low, this limits the accuracy of $K_A$ to about the same level of precision. Specifically, the methodology gives an accurate value for $K_A / (n * [a]_T)$, while $[a]_T$ is typically determined quite accurately, along with the ratio of $K_A / n$. If all that can be obtained is a lower limit for $n$ (as discussed above), this then provides a lower limit for $K_A$.

## Section 5B.

Given here is
1) how to find the cooperativity factor (= $n$) as a parameter in curve fitting, using data for how [protein2-substrate complex] (= $b$) and [ternary complex] (= $A$) co-vary as [protein1] (= $p$) is changed, given the (constant) [protein2] (= $q$), [total probe] (= $H$), and the dissociation constant of the protein2-substrate complex (= $L$), and
2) how to get $p$ and $k$ as parameters in curve fitting using the data set $\{(b, A, \Delta)\}$, knowing $n$ and the dilution factors (= $\Delta$) for a stock solution of protein1, of unknown concentration.

key for single-character notation used here:

| below | Peacock & Jaynes [1] | description |
|---|---|---|
| $H$ | $[AB]_T$ | total concentration of labeled substrate |
| $h$ | $[AB]$ | free concentration of labeled substrate |
| $p$ | $[a]_T$ | total concentration of protein1 |
| $f$ | $[a]$ | free concentration of protein1 |
| $q$ | $[b]_T$ | total concentration of protein2 |
| $g$ | $[b]$ | free concentration of protein2 |
| $k$ | $K_A$ | equilibrium dissociation constant of protein1 from its single-protein complex |
| $L$ | $K_B$ | equilibrium dissociation constant of protein2 from its single-protein complex |
| $n$ | $n$ | cooperativity factor |
| $a$ | $[AaB]$ | concentration of single-protein1 complex |
| $b$ | $[ABb]$ | concentration of single-protein2 complex |
| $A$ | $[AaBb]$ | concentration of ternary complex |

**Contents:**
Given the (constant) [protein2] (= $q$), [total probe] (= $H$), and the dissociation constant of the protein2-substrate complex (= $L$), we first find the cooperativity factor (= $n$) as a parameter in curve fitting, using data for how [protein2-substrate complex] (= $b$) and [ternary complex] (= $A$) co-vary as [protein1] (= $p$) is changed. We keep track of the dilution factors used for $p$, for use below.
(The Kd of protein2 dissociating from the ternary complex, $L/n$, is then known).

Knowing $n$ and the dilution factors (= $\Delta$) for a stock solution of protein1, of unknown concentration $p$, we can then get $p$ and $k$ as parameters in curve fitting using the data set $\{(b, A, \Delta)\}$ using a different expression, derived below.

Here, the [strongly binding protein] (= $q$) is kept constant, as the weakly binding one ($p$) is varied.
We assume $q$ and $L$ are known from single-protein binding experiments **Note that here, *q* is a protein concentration, whereas elsewhere, it was a Kd.**

Equilibrium binding reaction: two proteins (or complexes) bind to two distinct sites on the "probe" (labeled) substrate (e.g., DNA);
(No non-specific competitor is present);

The following are known:
total [protein2] = $q$
dissociation constant, single-protein2 complex = $L$
total [labeled DNA] = $H$
[2-protein complex] = $A$ (measured, as [protein1] is varied)
[protein2 complexed with labeled substrate] = $b$ (this is also measured, as [protein1] is varied)
$H$ = total [labeled DNA], which includes both bound ($a$+$b$+$A$) and free ($h$);
$a$ and $b$ are the [labeled single-protein complexes], $A$ is the [ternary complex], and
$h$ is the free [labeled DNA];

Dissociation constants of the labeled-DNA – single-protein complexes are $k$ and $L$.
Only $L$ is known initially.

Dissociation (equilibrium) constants of the various complexes with labeled substrate;
protein 1:

■ $k = \frac{fh}{a}$

■ $\frac{fb}{A} = \frac{k}{n}$

protein 2:

🟨 $L = \frac{gh}{b}$

🟥 $\frac{ga}{A} = \frac{L}{n}$

We have the following governing equations with 12 variables ($H, h, p, q, f, g, a, b, A, k, L, n$),
5 of them known or measured during the experiment ($H, L, q, b, A$).
Initially, we want to determine $n$ in terms of the known quantities (then later, $p$ and $k$).
The 6 governing equations are:
#1:

☒ $p = f + a + A$

#2:

☒ $q = g + b + A$

#3:

☒ $H = h + a + b + A$

#4:

☒ $k = \frac{fh}{a}$

#5:

☒ $L = \frac{gh}{b}$

#6:

☒ $n = \frac{Ah}{ab}$

Equations 2, 3, 5, and 6 can be considered a simpler system without $p, f$, and $k$ (and so containing 9 variables, allowing $n$ to be solved for in terms of the 5 known quantities, using the 4 equations).
Rearranging #2 and using #5, and solving for $h$:

☒ $h = \frac{Lb}{q - b - A}$

gives an expression for $a$ (using #3):

☒ $$a = H - A - b - h = H - A - b - \frac{Lb}{q - b - A}$$

Now, using #6 and the expression for $h$ above:

☒ $$n = \frac{\frac{A}{b}h}{a} = \frac{\frac{\frac{A}{b}Lb}{q - b - A}}{H - A - b - \frac{Lb}{q - b - A}} = \frac{\frac{LA}{q - b - A}}{H - A - b - \frac{Lb}{q - b - A}} = \frac{LA}{(H - A - b)\ (q - A - b) - Lb}$$

gives:

🟪 $$n = \frac{LA}{(H - A - b)\ (q - A - b) - Lb}$$

Given $q$ and $L$ from single-protein experiments, this can be used for curve fitting to find $n$, taking as data points $b$ and $A$, which vary with $p$. The actual value of $p$ is unknown at this point, even if it can be measured directly, because the fraction of it that is active is unknown.

To use this in standard curve fitting software, solve this for $A$:

☒ $$A = \frac{1}{2}\left(H + q - 2b + \frac{L}{n} - \sqrt{\left(H + q - 2b + \frac{L}{n}\right)^2 - 4\,(\,(H - b)\ (q - b) - Lb)}\right)$$

Use this to find $n$ from data points $(b, A)$, curve fitting for the parameter $n$, given $H$, $q$, and $L$.

This expression involves $L/n$ only in relation to the term $H + q - 2b$. Therefore, to maximize the effectiveness of the curve fitting, we should minimize the values of $H$ and $q$ used in the experiment, consistent with maintaining accurately quantifiable levels of $b$ and $A$. The opposite is true for finding $p$ and $k$ (see below).

We can get $p$ and $k$ from the same data set used to find $n$, by "keeping track" of how much $p$ is varied (defining the factor $\Delta$), and curve fit to get $p$ and $k$, from the data set $\{(b, A, \Delta)\}$.
To do this, we will need $a$ and $h$ in terms of $b$, $A$, $H$, and $n$.
Using #3 and #6:

🟦 $$H - b - A = a + h = a\left(1 + \frac{nb}{A}\right) = h\left(\frac{A}{nb} + 1\right)$$

so we have:

🟨 $$a = \frac{H - b - A}{1 + \frac{nb}{A}}$$

and:

$$h = \frac{H - b - A}{1 + \frac{A}{nb}}$$

From #4 and #6:

$$k = \frac{fh}{a} = \frac{nfb}{A}$$

and using #1:

$$p = f + a + A$$

$$f = p - A - a = p - A - \frac{H - b - A}{1 + \frac{nb}{A}}$$

So,

$$k = \frac{nfb}{A} = \frac{nb}{A}\left(p - A - \frac{H - b - A}{1 + \frac{nb}{A}}\right)$$

$$\frac{kA}{nb} = p - A\left(\frac{1 + \frac{nb}{A}}{1 + \frac{nb}{A}}\right) - \frac{H - b - A}{1 + \frac{nb}{A}} = p - \frac{A + nb}{1 + \frac{nb}{A}} - \frac{H - b - A}{1 + \frac{nb}{A}} = p - \frac{H + b\,(n - 1)}{1 + \frac{nb}{A}}$$

$$p = \frac{kA}{nb} + \frac{H + (n - 1)\,b}{1 + \frac{nb}{A}} = A\left(\frac{k}{nb} + \frac{H + (n - 1)\,b}{A + nb}\right)$$

We modify this by substituting $\Delta$*p0 for $p$, and now $\Delta$ varies (and is known) but p0 is constant (and is as yet unknown). We solve the expression for $\Delta$, and using the data set $\{(b, A, \Delta)\}$, curve fit to find p0 and $k$ as parameters.
p0 can be chosen as any one value, and $\Delta = 1$ for that data point.
The others $p$ values that correspond to each data point $(b, A)$ are then $\Delta$*'p0'.
The final curve fitting expression is thus:

$$\Delta = \frac{A}{\text{p0}}\left(\frac{k}{nb} + \frac{H + (n - 1)\,b}{A + nb}\right)$$

Here, small values of $H$ tend to make the term containing it small, relative to $k/nb$. This makes it relatively difficult to individually determine $k$ and p0, because when the term containing $H$ vanishes, the curve depends only on the ratio $k$ / p0.
So, higher values of $H$ are generally better for accurately determining $p$ and $k$, while, as described above, lower values of $H$ are generally better for accurately determining $n$. In practice, therefore, it is optimal to first determine $n$ using the minimum $H$ that is consistent with good quantitation of $b$ and $A$, then redo the experiment at higher values of $H$ (and enough $q$ to get accurately quantifiable levels of $b$ and $A$), in order to get precise values for p0 and $k$ (rather than only a precise value for their ratio).

**Section 5C.** The following is a summary of our curve fitting trials to find first *n*, then *p* and *k* using the equations derived in Section 5B.

key for single-character notation used here:

| below | Peacock & Jaynes [1] | description |
|---|---|---|
| $H$ | $[AB]_T$ | total concentration of labeled substrate |
| $k$ | $K_A$ | equilibrium dissociation constant of protein1 from its single-protein complex |
| $p$ | $[a]_T$ | total concentration of protein1 |
| $q$ | $[b]_T$ | total concentration of protein2 |
| $n$ | $n$ | cooperativity factor |

First, we summarize our trials to recover the values which we gleaned from the Vershon lab paper Jin et al., 1999 [7], namely $\{n, k\} = \{80{,}000, 80{,}000\}$, along with the values for *p* which we used in each trial to obtain the data for curve fitting (listed below).

Then, we do the same for the three En binding sites that we focused on in Peacock and Jaynes [1], as follows:
B1a, where $\{n, k\} = \{\ 500,\ \ 3{,}000\}$;
B1b, where $\{n, k\} = \{\ 113,\ \ 5{,}975\}$;
A2a, where $\{n, k\} = \{7{,}000, 90{,}000\}$.

All of the following refer to trials with 28 data points (taken from curves drawn using the exact parameter values) rounded to 2 significant figures (thereby incorporating errors in the range of 0.5 – 5%):

**Note: %** values in **boldface** indicate cases where the procedure worked well enough to determine a parameter within about 10%.

With ***H* = 20**, *q* = 10, 20, 40, 80}, *p* = {48, 90, 144}, curve fitting to find *n* didn't converge;
assuming the correct *n*, then finding *p* and *k*:
***p***: error 3.0%, curve fit error **5.5%**;
***k***: error 4.4%, curve fit error **6.2%**. This low curve fitting error suggests that, once *n* is determined with reasonable accuracy using a lower value of *H*, using this higher value of *H* will yield accurate values for *p* and *k* (confirmed below).

With ***H* = 10**, *q* = {4, 8, 12, 16}, *p* = {48, 64, 80, 96, 112, 128, 144}:
***n***: error 58%, curve fit error 111%.
Using the *n* value obtained from this curve fitting, then finding *p* and *k*:
***p***: error 5.2%, curve fit error 12%;
***k***: error 56%, curve fit error 12%.

With ***H* = 1**, *q* = {0.4, 0.8, 1.6, 3.2}, *p* = {48, 144, 288, 480, 800, 1440, 2880}:
***n***: error 5335%, curve fit error 1069%.
Using the *n* value obtained from this curve fitting, then finding *p* and *k*:
***p***: error 88%, curve fit error 99%;
***k***: error 541%, curve fit error 100%.

With ***H* = 0.1**, *q* = {0.04, 0.08, 0.16, 0.32}, *p* = {1000, 1200, 1440, 1720, 2000, 2400, 2880}:
***n***: error 7.6%, curve fit error **8.0**%.
Using the *n* value obtained from this curve fitting, then finding *p* and *k*:
***p***: error 100%, curve fit error 147%;
***k***: error 100%, curve fit error 148%.

With ***H* = 0.1**, *q* = {0.04, 0.08, 0.16, 0.32}, *p* = {480, 1440, 2880, 4800, 8000, 14400, 28800}:
actual error is 2.9%, curve fitting error is **3.6%**.
***n*:** error 2.9%, curve fit error **3.6%**.
Using the *n* value obtained from this curve fitting, then finding *p* and *k*:
***p*:** curve fitting doesn't converge;
***k*:** curve fitting doesn't converge.

Combining the *H*=0.1 data sets with the others does not help either convergence or % errors in the curve fitting.

**Note** that only with *H* = 0.1 (or less, not listed) is *n* determined within 10%.
Using the value for *n* so obtained, then using a data set taken with *H* = 20 in curve fitting to find *p* and *k*, gives *p* and *k* within ~10%:

**Using the value for *n* obtained with *H* = 0.1, then finding *p* and *k*** (% in parentheses is for 1st set of *q* values above with *H* = 0.1;
curve fitting error is the same for both sets of *q* values with *H* = 0.1)**:**
Using the ***H* = 10** data,
***p*:** error 5.4% (5.4%), curve fit error 12%;
***k*:** error 8.6% (14%), curve fit error 12%.

Using the ***H* = 20** data,
***p*:** error 4.4% (4.4%), curve fit error **5.5%**;
***k*:** error 6.7% (12%), curve fit error **6.2%**.

From these same data sets, we can get an accurate value for the ratio *k*/*p* much more easily than we can obtain their individual values accurately. Assuming a high value for *p*, then using curve fitting to find the ratio *k*/*p*, given a previously determined *n*, gives a very accurate value for the ratio (within 0.4% curve fitting error).

## Fujioka et al., 2012 [8], En binding sites

B1a, where {*n*, *k*} = {500, 3000};
B1b, where {*n*, *k*} = {113, 5975};
A2a, where {*n*, *k*} = {7000, 90000}:

**with *H* = 1**, *q* = {0.4, 0.8, 1.2, 1.6}, *p* = {1200, 1440, 1720, 2000, 2400, 2880}:

B1a:
First find ***n***: 1.2% error, curve fitting error **0.65%**; then *p* and *k*:
***p*** error 99%, curve fit error 38%;
***k*** error 100%, curve fit error 40%; then *p* and *k*

**from *H* = 40 data, using the above value for *n* (*H* = 1)**:
***p*** error 2.9%, curve fit error **5.4%**;
***k*** error 1.7%, curve fit error **6.2%**.

For this site, we needed to use *H* = 1 (or less) to accurately determine *n*, then *H* = 40 (or greater) to accurately determine *p* and *k*.

B1b:
First find ***n***: 0.17% error, curve fitting error **0.48%**; then *p* and *k*:
***p*** error 100%, curve fit error 100%;
***k*** wouldn't converge; then *p* and *k*

**from *H* = 40 data, using the value for *n* found with *H* = 1**:
***p*** error 16%, curve fit error 24%;
***k*** error 16%, curve fit error 24%; then *p* and *k*

**from *H* = 160 data, using the value for *n* found with *H* = 1**
***p*** error 6.8%, curve fit error **5.6%**;
***k*** error 8.1%, curve fit error **6.7%**.

For this site, we needed to use *H* = 10 (or less, see *H* = 10 below) to accurately determine *n*, then *H* = 160 (or greater) to accurately determine *p* and *k*.

A2a:
First find ***n***: 2.7% error, curve fitting error **8.4%**; then *p* and *k*:
***p*** error 100%, curve fit error 31%;
***k*** error 100%, curve fit error 33%. then *p* and *k*

**from *H* = 40 data, using the value for *n* found with *H* = 1**
***p*** error 4.9%, curve fit error 24%;
***k*** error 2.6%, curve fit error 24%; then *p* and *k*

**from *H* = 160 data, using the value for *n* found with *H* = 1**
***p*** error 5.9%, curve fit error **1.4%**;
***k*** error 11%, curve fit error **1.9%**.

For this site, we needed to use *H* = 1 (or less) to accurately determine *n*, then *H* = 160 (or greater) to accurately determine *p* and *k*.

**with *H* = 10**:

B1a, *q* = {4, 8, 12, 16}, *p* = {48, 64, 80, 96, 112, 128, 144}:
First find ***n***: 7.6% error, curve fitting error 14%; then *p* and *k*:
***p*** error 217%, curve fit error 185%;
***k*** error 246%, curve fit error 186%.

B1b, *q* = {4, 8, 12, 16}, *p* = {1200, 1440, 1720, 2000, 2400, 2880}:
First find ***n***: 1.2% error, curve fitting error **0.94%**; then *p* and *k*:
***p*** error 97%, curve fit error 43%;
***k*** error 97%, curve fit error 46%.

B1b, *q* = {4, 8, 12, 16}, *p* = {48, 64, 80, 96, 112, 128, 144}:
First find ***n***: 146% error, curve fitting error 35%; then *p* and *k*:
***p*** error 93%, curve fit error 43%;
***k*** error 82%, curve fit error 44%.

A2a, *q* = {4, 8, 12, 16}, *p* = {48, 64, 80, 96, 112, 128, 144}:
First find ***n***: 62% error, curve fitting error 53%; then *p* and *k*:
***p*** error 60%, curve fit error 145%;
***k*** error 85%, curve fit error 140%.

**with *H* = 40**, *q* = {10, 20, 40, 80}, *p* = {16, 32, 48, 64, 88, 112, 144}:

B1a:
First find ***n***: 47% error, curve fitting error 30%;
then *p* and *k*:
***p*** error 3.7%, curve fit error **5.5%**;
***k*** error 53%, curve fit error **6.2%**.

B1b:
First find ***n***: 52% error, curve fitting error 41%;
then *p* and *k*:
***p*** error 15%, curve fit error 23%;
***k*** error 30%, curve fit error 24%.

A2a:
First find ***n***: 72% error, curve fitting error 105%;
then *p* and *k*:
***p*** error 4.6%, curve fit error 24%;
***k*** error 70%, curve fit error 24%.

**with *H* = 160**, *q* = {80, 160, 240, 320}, *p* = {32, 64, 96, 144, 196, 288}:

B1b:
First find ***n***: 21% error, curve fitting error 15%;
then *p* and *k*:
***p*** error 6.2%, curve fit error **5.5%**;
***k*** error 15%, curve fit error **6.7%**.

A2a:
First find ***n***: 43% error, curve fitting error 252%;
then *p* and *k*:
***p*** error 5.9%, curve fit error **1.4%**;
***k*** error 47%, curve fit error **1.9%**.

## Fig. 6, Peacock and Jaynes 2017

## Generalizing the methods to more than two binding sites and cooperating ligands

The binding polynomial expresses the relative concentrations of all species in a system of ligands interacting with a substrate. It is written as a sum of terms, one for each species, normalized to the concentration of free substrate. First, for comparison to the more complex case of a 3-ligand complex, the definition of the binding polynomial for a 2-ligand complex is, after, e.g., Haiech et al, 2014 [8]:

$$\mathrm{P}(a, b) = ([AB] + [AaB] + [ABb] + [AaBb]) / [AB],$$

which can be solved in terms of dissociation constants and free concentrations to give:

$$\mathrm{P}(a, b) = 1 + [a]/K_A + [b]/K_B + n * [a]/K_A * [b]/K_B.$$

Introducing another ligand *c* then requires additional terms in the binding polynomial, one for each species. Now our substrate (e.g., DNA containing 3 individual binding sites in positions that allow 3 distinct pairwise complexes to form) is *ABC*, and there are 3 different 2-ligand complexes, *AaBbC*, *AaBCc*, and ABbCc, as well as the 3-ligand complex *AaBbCc*. The binding polynomial is now:

$$\mathrm{P}(a, b, c) = ([ABC] + [AaBC] + [ABbC] + [ABCc] + [AaBbC] + [AaBCc] + [ABbCc] + [AaBbCc]) / [ABC],$$

which becomes, in terms of measurable constants and free ligand concentrations:

$$\mathrm{P}(a,b,c) = 1 + [a]/K_A + [b]/K_B + [c]/K_C + n_{AB} * [a]/K_A * [b]/K_B + n_{AC} * [a]/K_A * [c]/K_C + n_{BC} * [b]/K_B * [c]/K_C + n_{ABC} * [a]/K_A * [b]/K_B * [c]/K_C.$$

We now need subscripts for each pairwise cooperativity factor to distinguish which pair of ligands it is associated with, as well as another cooperativity factor associated with the 3-ligand complex. This factor is an independent quantity in the general case where additional free energy (positive or negative) may be associated with the formation of the 3-ligand complex beyond that associated with the formation of each 2-ligand complex.

As derived for a 2-ligand complex containing *a* and *b* in Fig. 2A of Peacock and Jaynes [1]:

$$n_{AB} = [ABC] * [AaBbC] / ([AaBC] * [ABbC])$$

and, by analogy,

$n_{AC} = [ABC] * [AaBCc] / ([AaBC] * [ABCc])$

$n_{BC} = [ABC] * [ABbCc] / ([ABbC] * [ABCc])$.

We can find $n_{ABC}$ to be:

$n_{ABC} = [ABC]^2 * [AaBbCc] / ([AaBC] * [ABbC] * [ABCc])$

by equating the term containing it in the binding polynomial with the definition of the corresponding term from above:

$n_{ABC} * [a]/K_A * [b]/K_B * [c]/K_C = [AaBbCc] / [ABC]$

$n_{ABC} = [AaBbCc] * K_A * K_B * K_C / ([ABC] * [a] * [b] * [c])$

where

$K_A = [ABC] * [a] / [AaBC]$

$K_B = [ABC] * [b] / [ABbC]$

$K_C = [ABC] * [c] / [ABCc]$.

In order to relate the "new" cooperativity factor for the 3-ligand complex to those of the 2-ligand complexes, we note that complete dissociation of the 3-ligand complex involves the sum of 3 free energies. For one possible dissociation route, these are: the free energy change when ligand *a* dissociates, that when ligand *b* dissociates from the 2-ligand complex, and that when ligand *c* dissociates from the single-ligand complex.

Because the standard Gibbs free energy and the Kd are related by:

$\Delta G^0 = - R * T * \ln(K_A)$ (where R is the gas constant and T is the absolute temperature), adding these 3 free energies is equivalent to multiplying the 3 dissociation constants that govern dissociation of ligand *a* from the 3-ligand complex, dissociation of ligand *b* from the 2-ligand complex containing ligands *b* and *c*, and dissociation of ligand *c* to release the free DNA.

This product is:

$([ABbCc] * [a] / [AaBbCc]) * ([ABCc] * [b] / [ABbCc]) * ([ABC] * [c] / [ABCc])$

$= [ABC] * [a] * [b] * [c] / [AaBbCc]$

$= K_A * K_B * K_C / n_{ABC}$,

where the last equality comes from the last expression above for $n_{ABC}$.

So, $1 / n_{ABC}$ represents the free energy in the complex that is due to cooperative binding; i.e., this "extra" $\Delta G^0 = R * T * \ln(n_{ABC})$. From the definitions above, the Kd that governs the dissociation of ligand *a* from the 3-ligand complex is:

$[ABbCc] * [a] / [AaBbCc] = \big([ABC] * [a] / [AaBC]\big) * \big\{[ABC] * [ABbCc] / \big([ABbC] * [ABCc]\big)\big\} / \big\{[ABC]^2 * [AaBbCc] / \big([AaBC] * [ABbC] * [ABCc]\big)\big\}$

$= K_A * n_{BC} / n_{ABC}$,

and similarly for the other dissociation constants from the 3-ligand complex.

This tells us that the Kd for dissociation of ligand *a* from the 3-ligand complex equals the Kd for dissociation of ligand *a* from the DNA (in the absence of the other ligands) multiplied by the cooperativity factor governing the dissociation of the remaining 2-ligand complex, then divided by the 3-ligand cooperativity factor. As defined in Fig. 2A for the cooperativity factor associated with a 2-ligand complex, $1 / n_{BC}$ represents the free energy associated with cooperativity in the 2-ligand complex containing ligands *b* and *c*, and we now know that $1 / n_{ABC}$ represents the free energy associated with cooperativity in the 3-ligand complex. So, it makes sense that the equilibrium constant for dissociation of ligand *a* from the 3-ligand complex would equal $K_A$ multiplied by $n_{BC}$ (representing the free energy due to cooperativity remaining in the 2-ligand complex containing *b* and *c*), divided by $n_{ABC}$ (representing all of the free energy due to cooperativity within the 3-ligand complex).

For $n_{ABC}$ in the case where the free energies of interaction within the complex consist solely of those found within the respective 2-ligand complexes, the energy change that occurs when ligand *a* dissociates from the 3-ligand complex is the sum of those that occur when ligand *a* dissociates from each of the 2-ligand complexes containing it, minus the energy change that occurs when ligand *a* dissociates from the single-ligand complex (because the sum includes two such dissociations from the DNA itself, one from each 2-ligand complex). This means that the Kd that governs the dissociation of ligand *a* from the 3-ligand complex will be the product of those that govern the dissociation of ligand *a* from each of the 2-ligand complexes that contain it, divided by the Kd that governs the dissociation of ligand *a* from the single-ligand complex. In symbols, this means that:

$K_A * n_{BC} / n_{ABC} = (K_A / n_{AB}) * (K_A / n_{AC}) / K_A$.

Cancelling $K_A$'s and rearranging gives:

$n_{AB} * n_{AC} * n_{BC} = n_{ABC}$.

So, <u>if</u> the free energies of association within the 3-ligand complex are simply the sum of those that occur within the 2-ligand complexes, $n_{ABC}$ is the product of the cooperativity factors of the three 2-ligand complexes, and we already have enough information from analysis of the 2-ligand complexes to predict the behavior of the entire 3-ligand system (see below).

If, instead, we find by studying the formation of the 3-ligand complex that the above relationship does not hold, we can then measure $n_{ABC}$, and develop and test hypotheses that can explain its deviation from the product of the 2-ligand cooperativity factors.

How can we determine $n_{ABC}$ to see if the above relationship holds? Perhaps the most straightforward requires us to distinguish and quantify only the 3-ligand complex and the free DNA, and use the individual Kd's and protein concentrations (determined by the methods described in this paper) along with the following, from above:

$n_{ABC} = [AaBbCc] * K_A * K_B * K_C / ([ABC] * [a] * [b] * [c])$.

If this is done under conditions where the total [DNA] is much less than each of the total protein concentrations, then the free protein concentrations are essentially equal to the total protein concentrations, and we have all the information necessary to obtain $n_{ABC}$. Once $n_{ABC}$ is determined, we can use computational methods to solve the system and obtain the concentrations of each species over the full range of total protein and DNA concentrations.

<u>This system consists of 22 variables</u>, namely, the 3 single-ligand Kd's, the 4 cooperativity factors, the concentrations of 9 forms of DNA (including total, bound, and unbound), and the total and free concentrations of each of the 3 proteins. <u>It is constrained by the following 11 equations:</u> the definitions of the 3 individual Kd's and the 4 cooperativity factors, and the 4 "continuity" equations where the total concentrations of each protein and of the DNA are set equal to the sum of the free and bound forms of each.

<u>The system is determined if we specify 11 of the variables.</u> For example, if we know the 3 single-ligand Kd's and the 4 cooperativity factors, along with the total concentrations of DNA and of each of the 3 proteins, then solving the system gives us the values of the other 11

variables:  the concentrations of the 7 complexes and the free concentrations of the DNA and of each of the 3 proteins.

In principle, we can follow an analogous procedure to characterize a 4-ligand system, by first determining the Kd's of each of the 4 ligands individually, along with the cooperativity factors for each of the 4 possible 3-ligand complexes (as described above), and then determining the "new" cooperativity factor for the 4-ligand complex.  This will be:

$$n_{ABCD} = [AaBbCcDd] * K_A * K_B * K_C * K_D / ([ABCD] * [a] * [b] * [c] * [d]),$$

and it can be determined by quantifying the [4-ligand complex] and the free [DNA] under conditions where the latter is much lower than each of the total [protein], as described above for the 3-ligand complex.

If all of the interactions leading to cooperativity are contained within pairwise interaction domains that are not significantly affected by higher-order complex formation, then the sum of the free energies from the pairwise interactions equals the total cooperative free energy of the entire complex, and:

$$n_{ABCD} = n_{AB} * n_{AC} * n_{BC} * n_{AD} * n_{BD} * n_{CD}$$

For $j$ ligands binding to distinct sites on a substrate and cooperating solely through pairwise interactions, the cooperativity factor is the product of the

$$j! / [2 * (j - 2)!]$$

possible pairwise cooperativity factors, which can each be measured by studying the ternary complex containing those two ligands, using the methods given here.

## Acknowledgements

This research was supported by the National Institutes of Health, Awards R01GM050231 and R01GM117458 to JBJ.

## References

[1] Peacock and Jaynes, Using competition assays to quantitatively model cooperative binding by transcription factors and other ligands, BBA-General Subjects, in press.

[2] Ernesto Freire, Arne Schön, and Adrian Velazquez-Campoy, "Isothermal Titration Calorimetry: General Formalism Using Binding Polynomials", Methods in Enzymology, Volume 455, pp. 127-155 (Chapter 5) (2009) Elsevier Inc., ISSN 0076-6879, DOI: 10.1016/S0076-6879(08)04205-5.

[3] W.G. Bardsley, Factorability of the Allosteric Binding Polynomial and Graphical Manifestations of Cooperativity in Third Degree Saturation Functions, J. Theor. Biol. 67, 407-431 (1977).

[4] J. Haiech, Y. Gendrault, M.-C. Kilhoffer, R. Ranjeva, M. Madec, C. Lallement, A general framework improving teaching ligand binding to a macromolecule, Biochimica et Biophysica Acta (BBA) - Mol. Cell Res. 1843 (2014) 2348–2355.

[5] T.R. Riley*, M. Slattery*, N. Abe, C. Rastogi, D. Liu, R.S. Mann, H.J. Bussemaker, SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes, pp. 255-278 ( Chapter 16) of Y. Graba and R. Rezsohazy (eds.), Hox Genes: Methods and Protocols, Methods in Molecular Biology, vol. 1196, DOI 10.1007/978-1-4939-1242-1_16, © Springer Science+Business Media New York 2014 [*these authors made equal contributions]

[6] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen, J. Taipale, Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity (2006), Cell 124:47–59.

[7] Y. Jin, H. Zhong, A.K. Vershon, The Yeast a1 and α2 Homeodomain Proteins Do Not Contribute Equally to Heterodimeric DNA Binding, Mol. Cell. Biol. 19 (1999) 585-593.

[8] M. Fujioka, B. Gebelein, Z.C. Cofer, R.S. Mann, J.B. Jaynes, Engrailed cooperates directly with Extradenticle and Homothorax on a distinct class of homeodomain binding sites to repress sloppy paired, Dev. Biol. 366 (2012) 382–392. doi:10.1016/j.ydbio.2012.04.004