ELECTRONIC SUPPLEMENTARY MATERIAL

C-terminal Interactions Mediate the Quaternary Dynamics of αB-Crystallin

Gillian R. Hilton, Georg K. A. Hochberg, Arthur Laganowsky,

Scott I. McGinnigle, Andrew J. Baldwin* & Justin L.P. Benesch*

Department of Chemistry, Physical & Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, Oxfordshire, OX1 3QZ, U.K.

Contains:

SUPPLEMENTARY METHODS

Ligand binding model

In the absence of peptide ligands, L, the truncated αB-crystallin constructs, P, exist in equilibrium between monomeric, P_1 and dimeric, P_2, forms. Given that each molecule of P has the capacity to bind one molecule of L, and that oligomerisation rarely proceeds past the level of the dimerisation in these constructs, we have six species to consider, P_1, P_1L, P_2, P_2L, P_2L_2, and L. Interpreting binding data therefore requires consideration of six coupled equilibria and their corresponding association constants.

$$P_1 + L \rightleftarrows P_1L, K_{A,0} = \frac{[P_1L]}{[P_1][L]}$$

$$P_2 + L \rightleftarrows P_2L, K_{A,1} = \frac{[P_2L]}{[P_2][L]}$$

$$P_2L + L \rightleftarrows P_2L_2, K_{A,2} = \frac{[P_2L_2]}{[P_2L][L]}$$

$$P_1 + P_1 \rightleftarrows P_2, K_{A,3} = \frac{[P_2]}{[P_1]^2}$$

$$P_1L + P_1 \rightleftarrows P_2L, K_{A,4} = \frac{[P_2L]}{[P_1L][P_1]}$$

$$P_1L + P_1L \rightleftarrows P_2L_2, K_{A,5} = \frac{[P_2L_2]}{[P_1L]^2}$$

There is some degeneracy within the association constants in this scheme: $K_{A,2}K_{A,4} = K_{A,5}K_{A,0}$ and $K_{A,1}K_{A,3} = K_{A,4}K_{A,0}$. These equalities have an important physical interpretation, in that weakening the monomer/dimer interface results in a decrease in species like P_2L and P_2L_2. The dimer interface free energy and the effective association constants for the dimer species are therefore closely linked. Due to the degeneracy, rather than being described by six equilibrium constants, the system is instead completely defined by four constants. In what follows, we choose to define properties in terms of $K_{A,0}$, $K_{A,3}$ $K_{A,4}$ and $K_{A,5}$, with $K_{A,2}$ and $K_{A,1}$ determined using the relations stated above. The total concentration of ligand is given by:

$$
\begin{aligned}
[L_{Tot}] &= \sum_{i=1}^{2}\sum_{j=0}^{i} j[P_iL_j] + [L] \\
&= [L] + [P_1L] + [P_2L] + 2[P_2L_2] \\
&= [L] + K_{A,0}[L][P_1] + K_{A,4}K_{A,0}[P_1]^2[L] + 2K_{A,5}K_{A,0}^2[P_1]^2[L^2]
\end{aligned}
$$

Which is quadratic in [L]. Solving for [L] and taking the positive solution yields:

$$
[L] = \frac{\left(1 + K_{A,0}[P_1] + K_{A,4}K_{A,0}[P_1]^2\right) + \sqrt{\left(1 + K_{A,0}[P_1] + K_{A,4}K_{A,0}[P_1]^2\right)^2 + 8L_{Tot}K_{A,5}K_{A,0}^2[P_1]^2}}{2L_{Tot}}
$$

The quantity of free peptide can be directly determined if the four equilibrium constants, total peptide concentration and quantity of free monomeric protein is known. Similarly, the total concentration of protein is given by:

$$
\begin{aligned}
[P_{Tot}] &= \sum_{i=1}^{2}\sum_{j=0}^{i} i[P_iL_j] \\
&= [P_1] + [P_1L] + 2\left([P_2] + [P_2L] + [P_2L_2]\right) \\
&= [P_1] + K_{A,0}[P_1][L] + 2K_{A,3}[P_1]^2 + 2K_{A,0}K_{A,4}[P_1]^2[L] + 2K_{A,5}K_{A,0}^2[P_1]^2[L]^2
\end{aligned}
$$

Inserting the expression for [L] into this expression gives an expression for [P_{tot}] in terms of [P_1] and the four equilibrium constants. In the binding experiments described in the manuscript, the total quantity of protein has been experimentally determined, and so the equation can be solved numerically for [P_1] for any set of four trial equilibrium constants. With these values, all remaining concentrations [P_iL_j] are trivially back-calculated.

Fitting the model to the data

From the IM-MS experiments, we have experimental measurements of the signal intensity from each species, $I\left(P_iL_j\right)$, such that the total signal intensity is $I(Tot) = \sum_{i=1}^{2}\sum_{j=0}^{i} I\left(P_iL_j\right)$. For any given set of equilibrium constants, we can evaluate the following χ^2 function:

$$
\chi^2 = \sum_{i=1}^{2}\sum_{j=0}^{i}\left(\left(\frac{I\left(P_iL_j\right)}{I\left(Tot\right)}\right)_{\exp} - \left(\frac{[P_iL_j]}{[P_{Tot}]}\right)_{calc}\right)^2
$$

Where $I\left(P_iL_j\right)$, $I(Tot)$ and [P_{Tot}] are determined from experiment, and the set of [P_iL_j] are obtained as a function of the four equilibrium constants, as described above. Assuming that signal intensities are proportional to the concentrations of the corresponding species in solution, the set of equilibrium constants that return the lowest χ^2 value are those that best describe the species present in solution.

Simplifying the model

The equilibrium scheme above implicitly allows for ligand binding to influence the strength of the dimer interface, and vice versa. For example, if $K_{A,5}>K_{A,3}$ the dimer interface is effectively strengthened by the binding of two ligands. This model is a function of four equilibrium constants and is the most complex considered in this study. We can make simplifying assumptions that result in models of reduced complexity.

If we make the assumption that $K_{A,3}=K_{A,4}$, which results in the further equality $K_{A,0}=K_{A,1}$, the simplified scheme states that a single ligand binds both monomers and dimers with equal affinity. This model still allows the second ligand to bind with a different affinity, and is specified by three free parameters rather than four.

Finally, we can assume that $K_{A,3}=K_{A,4}=K_{A,5}$ then $K_{A,0}=K_{A,1}=K_{A,2}$. This model assumes that binding the ligand does not alter the dimer interface in any way, and that ligand binding is entirely independent of oligomerisation. This relatively simple model is specified by only two free parameters, and is the least complex of those tested in this study.

Which model is 'correct'?

Fitting data to a model with an increased number of free parameters will result in a lower χ^2. It is important to distinguish whether this lowering in χ^2 comes from an improved physical model, or if it has instead arisen due to unfortunate patterns in the noise. This distinction can be evaluated statistically by performing an F-test.

In this study, three models have been considered as described above, each with increasing statistical complexity being defined by 2, 3 or 4 free parameters. To determine if the increased complexity is statistically justified, we determined the lowest χ^2 of each model, and noted the number of associated 'degrees of freedom', ν, equal to the number of free parameters in the model minus the number of data points. We could then assess the

probability of whether the reduction in χ^2 in moving from a simple to a complex model is justified by the reduction in degrees of freedom by evaluating the P-level:

$$P = \beta_R\left(\frac{\nu_{complex}}{2}, \frac{\nu_{simple} - \nu_{complex}}{2}, \frac{\chi^2_{complex}}{\chi^2_{simple}}\right)$$

Where β_R is the regularised β function, defined as $\beta_R(x,y,z) = \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)}\int_0^z t^{x-1}(1-t)^{y-1}dt$, and Γ is the Gamma function, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}t$, in the relevant $z>0$ limit. Should the probability be less than 0.06 (6%), then we can conclude that the improved fit using the more complex model is statistically justified. If the probability is greater than 6%, then noise in the data is most likely responsible for the improve χ^2. In such a situation, we cannot reliably distinguish between the two models and so we stay at the complexity of the simpler model.

In the data described in the tests, the 3 parameter model was found to be statistically justified for both the αB_{Pal} and αB_{Ext} binding datasets according to these definitions. In the case of the αB_{Cnt} peptide, there was no significant reduction in moving from the 2 to the 3 parameter model. There was no instance here in which the 4 parameter model was statistically justified.

Determination of errors in fitted binding affinity parameters

Multiple repeats of the samples were run on independent protein separations on different days in order to get reliable error bars. Typically, 3-6 independent measurements were performed on each peptide / concentration combination. The error bars shown in the isotherms in figure S2 represent the standard error of the mean $\sigma_M = \sigma/\sqrt{N-1}$ where σ is the standard deviation of the repeated measurements. To obtain numerical error estimates of the affinity constants, a boot-strapping procedure was employed. Here, a new dataset was calculated from randomly sampling the data points in a manner that preserves the total number of data points at the end, with some data points randomly double-sampled and some not at all. The fitting is repeated to obtain a new set of binding affinities. By repeating this procedure 1000 times, a histogram of the affinity coefficients can be created whose mean is identical to the best fit from the raw data, and whose width reflects the numerical uncertainty in the parameter in question (Figure 4) .

Adjustments due to non-specific association in the gas phase

In a typical nanoelectrospray MS experiment, the protein will be at concentrations on the order of 10 μM. In the case where K_D is on the order of 100 μM or weaker, it is necessary to mix with ligand at concentrations that approach 100 μM. Under these conditions of relatively high concentration, the probability of having a peptide and an αB-crystallin domain in the same droplet becomes appreciable, appearing in the spectrum as a low concentration bound species (1). Such "non-specific" binding effects occlude the effects of "specific" binding present in the solution phase.

Considering the case of a binding equilibrium where both specific (present in solution) and non-specific (formed during the electrospray process) are detected, the equilibria and corresponding association constants are given by:

$$P + L \rightleftarrows PL^{spec}, K_A^{spec} = \frac{[PL^{spec}]}{[P][L]}$$
$$P + L \rightleftarrows PL^{non}, K_A^{non} = \frac{[PL^{non}]}{[P][L]}$$

The total protein concentration will be given by

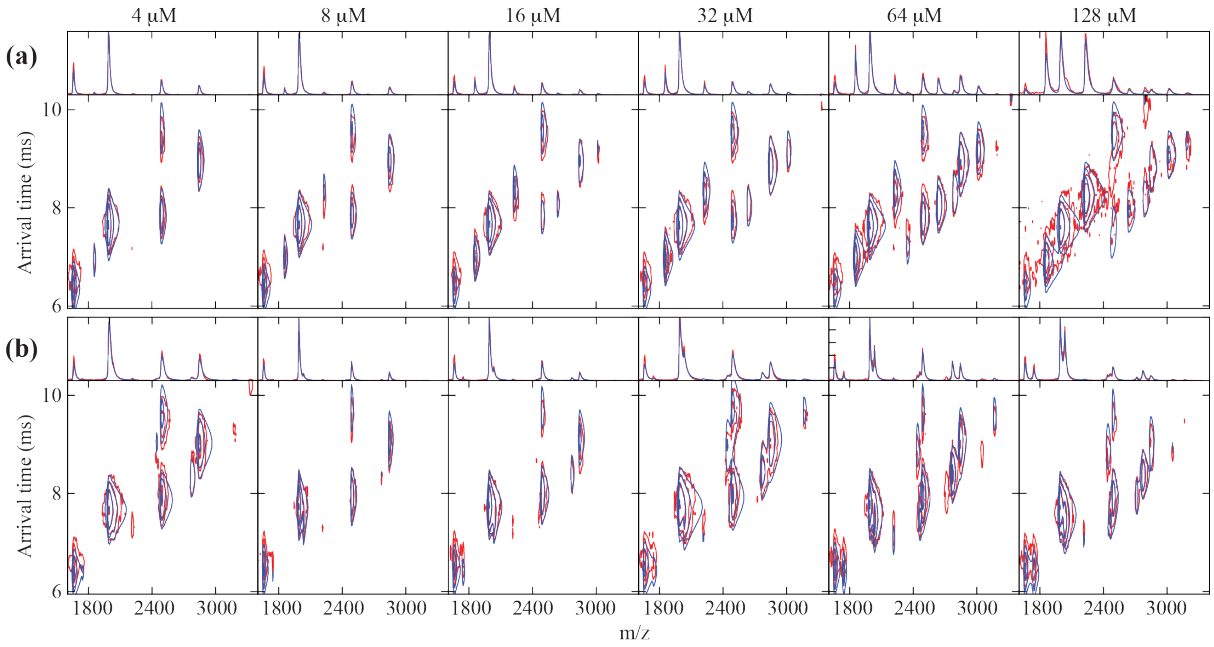
$$\begin{aligned} P_{Tot} &= [P] + [PL^{spec}] + [PL^{non}] \\ &= [P] + [P][L]\left(K_A^{spec} + K_A^{non}\right) = [P] + K_A^{eff}[P][L] \end{aligned}$$

This leads to a useful general result, $K_A^{eff} = K_A^{non} + K_A^{spec}$. The effective association constant is that which is determined through neglecting the effects of non-specific association. We arrive at an estimate of the required association constant through $K_A^{spec} = K_A^{eff} - K_A^{non}$. Providing we can estimate the magnitude of non-specific binding effects, they can be trivially subtracted. When the association constant is more than an order of magnitude larger than the effects of non-specific binding, the correction is negligible. However, when the effective constant and the non-specific constant are on the same order of magnitude, the correction will be large.

In this work, in order to estimate the magnitude of effects due to non-specific association, 128uM of the αB_{Pal} peptide was added to the unrelated protein cytochrome C. In the resulting spectra, a bound state populated to approximately 10% state was detected. This corresponds to an effective K_D of 1 mM. In this study, this value is taken as an approximation for K_D^{non}. The association constants that describing peptide binding, $K_{A,3}$, $K_{A,4}$ and $K_{A,5}$ are susceptible

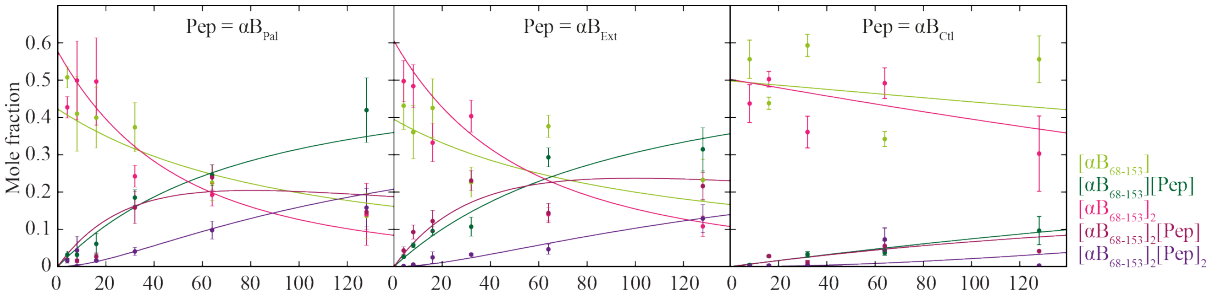
to effects of this kind. Values reported in the manuscript for these constants have had non-specific effects subtracted. It is important to note that this correction is small for all K_D values presented in this work.

FIGURE S1



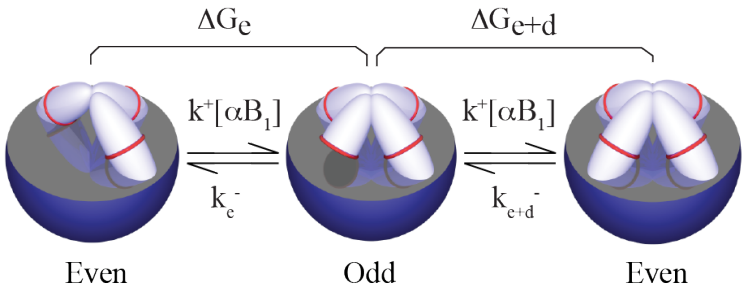
Comparison of both IM-MS experimental data (blue) and fits (red) obtained using the CHAMP algorithm, as described in the text, for titrations of **(a)** αB_{Pal} and **(b)** αB_{Ext} with αB_{Dom}. The fitted spectra are in excellent agreement with the experimental data, as can be seen from the one-dimensional projections.

FIGURE S2



Two dimensional IM-MS spectra were acquired for each combination of peptide concentration and type binding to with αB_{Pal} between 3-6 times, allowing a mean and standard error of each observed species to be defined. The data were then fitted to the binding affinity model described above. The two, three and four parameter models were fitted to the data, and an F-test used to decide if the fitting quality is sufficiently improved as to justify the inclusion of additional parameters. In the case of αB_{Pal} (left) and αB_{Ext} (middle), using the three parameter model rather than the two parameter was justified (P-level < 6%). Physically, the two parameter fit was poor as far less $[P_2L_2]$ was observed than the model would predict. The reduction in χ^2 stems from the three parameter model allowing for this, which enables us to conclude that binding the second peptide is significantly more difficult than binding the first. By contrast, no significant reduction in χ^2 was found moving from the two to the three parameter model in the case of the negative-control peptide αB_{Ctl} (right). In no instance was fitting using the four parameter model statistically justified.

TABLE S1



$$\Delta G_{e+d} = -RTlnM_A$$
$$\Delta G_d = -RT\ln N$$
$$\Delta G_e = -RT\ln\left(M_A / N\right)$$
$$\Delta G_{e+d}^{\ddagger} = -RT\ln\left(k_{e+d} / \nu\kappa\right)$$
$$\Delta G_e^{\ddagger} = -RT\ln\left(k_e / \nu\kappa\right)$$
$$\Delta G_+^{\ddagger} = -RT\ln\left(k_+[P_1] / \nu\kappa\right)$$

	ΔG_{e+d}	ΔG_d	ΔG_e	$\Delta\Delta G_{e+d}$	$\Delta\Delta G_d$	$\Delta\Delta G_e$	$\Delta\Delta G_e^R$	$\Delta\Delta G_{e+d}^{\ddagger}$	$\Delta\Delta G_e^{\ddagger}$	$\Delta\Delta G_+^{\ddagger}$
αB_{WT}	-8.8	-7.1	-1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
αB_E	-9.0	-7.3	-1.8	-0.2	-0.1	-0.1	2.6	-1.1	-1.2	-1.4
αB_R	-9.3	-8.2	-1.1	-0.5	-1.0	0.6	1.8	-1.4	-2.5	-1.9
αB_{ER}	-9.1	-8.2	-0.9	-0.4	-1.1	0.8	4.3	-1.2	-2.3	-1.6
αB_{ERT}	-9.2	-8.7	-0.5	-0.4	-1.6	1.2	7.4	-1.7	-3.3	-2.1
αB_{II}	-9.1	-8.0	-1.1	-0.3	-0.8	0.5	38.6	-1.6	-2.4	-1.9
αB_{IPI}	-9.8	-10.1	0.3	-1.0	-3.0	2.0	40.0	-0.6	-3.5	-1.6
αB_{TRE}	-8.8	-6.9	-2.0	-0.1	0.2	-0.3	10.0	-1.2	-1.0	-1.3

Table of thermodynamic and kinetic quantities obtained from our analysis of with αB-crystallin mutants. ΔG and ΔΔG values are in kJ/mol. $\Delta\Delta G=\Delta G_{Mutant}-\Delta G_{WT}$ for both thermodynamic and kinetic parameters. ΔG values are defined according the following, where M_A, N, k_{e+d}, k_e and $k^+[\alpha B_1]$ are parameters extracted from fitting the data to the model as described previously (3), R is the gas constant, T is the thermodynamic temperature, ν is the characteristic vibration frequency across the reaction co-ordinate, and κ is the transmission coefficient. The terms in νκ are unimportant in this instance as they cancel when calculating the $\Delta\Delta G^{\ddagger}$ values. $\Delta\Delta G^R$ values were calculated using a Rosetta point mutation protocol as described in the text.

SUPPLEMENTARY REFERENCES

1. Lane LA, Ruotolo BT, Robinson CV, Favrin G, Benesch JLP. A Monte Carlo approach for assessing the specificity of protein oligomers observed in nano-electrospray mass spectra. Int J Mass Spectrom. 2009;283(1-3):169-77.

2. Baldwin AJ, Lioe H, Hilton GR, Baker LA, Rubinstein JL, Kay LE, et al. The Polydispersity of alphaB-Crystallin Is Rationalized by an Interconverting Polyhedral Architecture. Structure. 2011 Dec 7;19(12):1855-63.

3. Baldwin AJ, Lioe H, Robinson CV, Kay LE, Benesch JLP. alphaB-crystallin polydispersity is a consequence of unbiased quaternary dynamics. J Mol Biol. 2011 Oct 21;413(2):297-309.