

## Supplemental Information

### Dissecting Heterogeneous Molecular Chaperone

### Complexes Using a Mass Spectrum

### Deconvolution Approach

Florian Stengel, Andrew J. Baldwin, Matthew F. Bush, Gillian R. Hilton, Hadi Lioe, Eman Basha, Nomalie Jaya, Elizabeth Vierling, and Justin L.P. Benesch

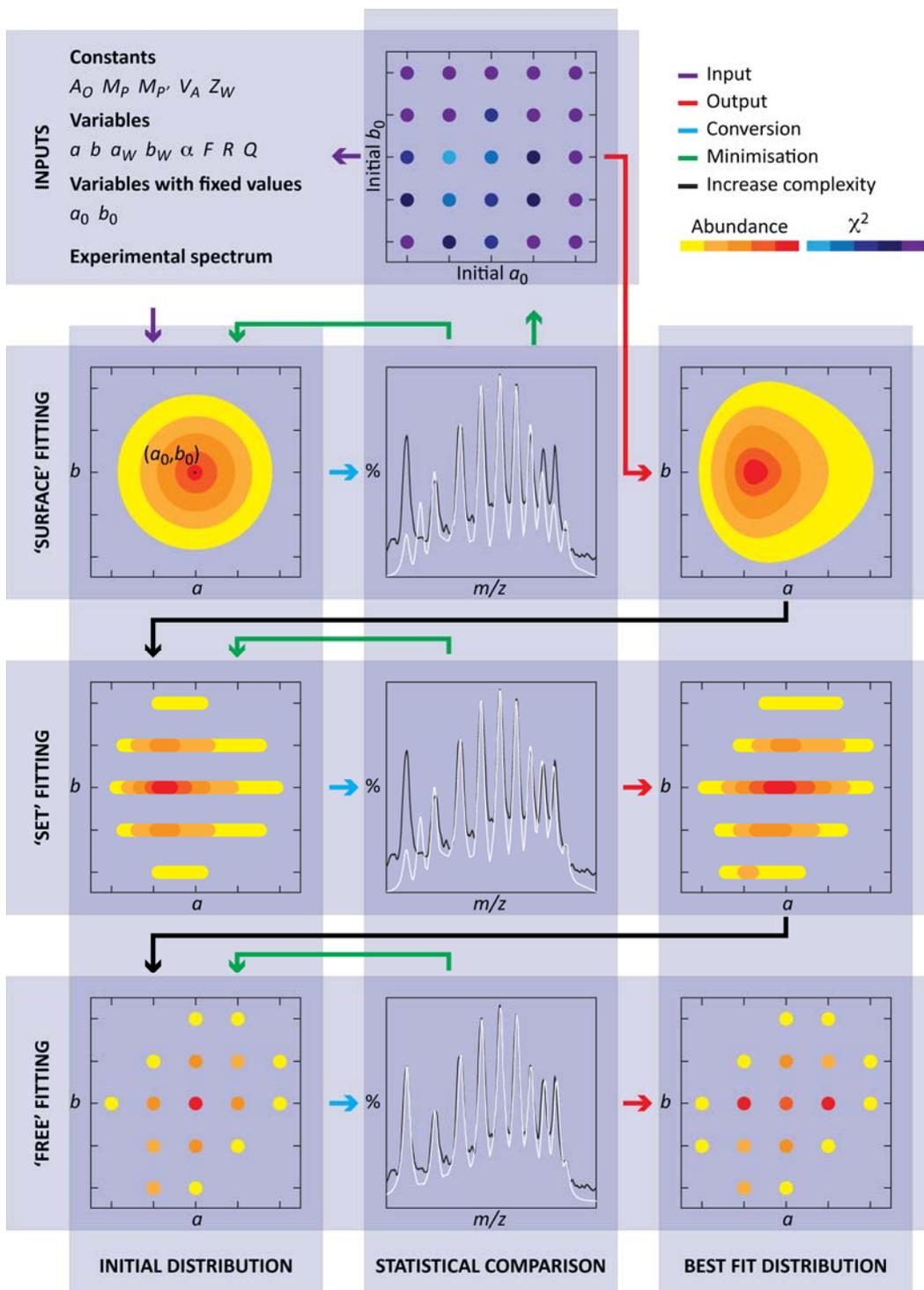
#### Inventory of Supplemental Information

Fig S1 CHAMP overview (relates to Fig 2)

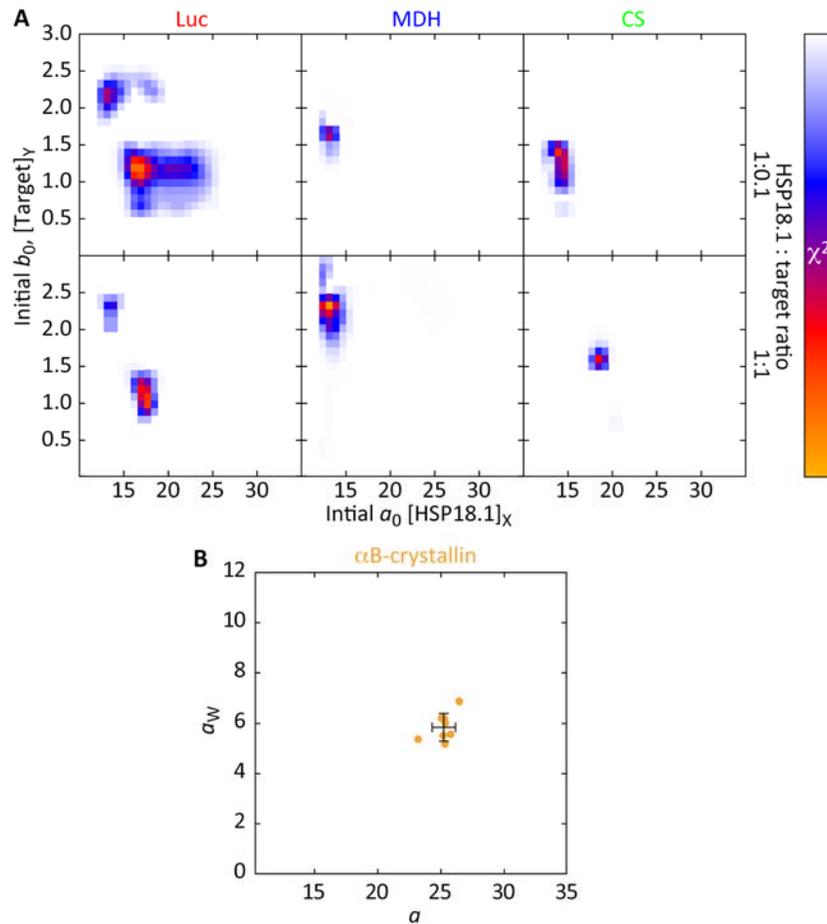
Fig S2  $\chi^2$  minimisation and reproducibility (relates to Fig 4)

Supplemental Experimental Procedures S1 CHAMP Model

Supplemental Experimental Procedures S2 CHAMP Fitting



**Fig. S1, related to Figure 2:** Flow diagram of CHAMP. Three incremental complexities of fit are possible, with the best fits found at each stage using  $\chi^2$  minimisation.



**Fig. S2, related to Figure 4:** A Minimisation of  $\chi^2$  for surface fit, as described in Figure S1, for the distributions shown in Figure 4. B Scatter plot of  $a$  and  $a_w$  for seven independent replicate nanoES spectra (red) of the polydisperse protein  $\alpha$ B-crystallin. The data are tightly clustered, with the standard deviations (error bars) less than 10% of the mean in both dimensions.

## S1 CHAMP Model

For specific application to sHSPs see Experimental Procedures. Calculated mass spectra of a heterogeneous ensemble of protein complexes are generated *in silico* using the model described below. The resulting spectra are shown to be a function of the overall mass of the complex under consideration, and three oligomer-independent parameters that are optimized for each individual spectrum as discussed in the text. In the following, the physical origin of these individual parameters is described, and we demonstrate how they propagate through our model.

If we consider a binary complex of the form  $P_aP'_b$ , where  $P$  and  $P'$  represent two separate proteins (in the present work sHSP and target), its intrinsic ‘sequence’ mass,  $M_S$ , is simply given by

$$M_S(a,b) = aM_P + bM_{P'}$$

[Eqn. S1]

where  $M_P$  and  $M_{P'}$  are the masses derived from the amino-acid composition of the two protein components. The effective mass,  $M_E$ , of protein assemblies observed experimentally in nanoES mass spectra typically have a small positive deviation from that calculated from the sequence due to the binding of buffer ions and solvent molecules (McKay, et al., 2006), such that

$$M_E(a,b) = M_S + M_A$$

[Eqn. S2]

where  $M_A$  is the additional mass due to these adducts. It stands to reason that the mass of adducts, both specific and non-specific, will be proportional to the exposed surface area of the protein oligomers, which has been empirically shown to be proportional to  $M_S^{0.76}$  (Miller, et al., 1987). Therefore we implement

$$M_E(a, b, Q) = M_S + QM_S^{0.76}$$

[Eqn. S3]

where  $Q$  is a parameter that is optimized for each individual spectrum to best account for the binding of adducts to all complexes.

The average charge state,  $Z_A$ , populated by a protein is approximately related, by considerations of the Rayleigh limit, by a square root relationship to its mass (de la Mora, 2000; Kebarle and Verkerk, 2010). The same form has been shown to apply to protein complexes (Heck and van den Heuvel, 2004; Kaltashov and Mohimen, 2005), with a relationship of  $Z_A = 0.0467M^{0.533}$  providing the best fit to the combined data sets (Stengel, et al., 2010). The variations in internal diameter between different emitters (Hernández and Robinson, 2007) leads to small fluctuations in the average protein charge states observed in nanoESI mass spectra (Li and Cole, 2003). Correspondingly, the average charge of a single complex is here modeled as

$$Z_A(a, b, F, Q) = 0.0467M_E^{0.533} + F$$

[Eqn. S4]

where  $F$  is a parameter optimized for each individual mass spectrum to best account for charge-state fluctuations across all complexes under consideration. When the complex

under consideration populates a single conformation, the distribution of charge states around  $Z_A$  is well approximated by a Gaussian distribution (Dobo and Kaltashov, 2001).

The intensity of individual charge states of a given complex will therefore be given by

$$W_C(a, b, z, F, Q) = \frac{\exp(-(z - Z_A)^2 / 2Z_W^2)}{\sum_{z=0}^{\infty} \exp(-(z - Z_A)^2 / 2Z_W^2)} \quad [\text{Eqn. S5}]$$

where  $Z_W$  reflects the width of the charge state distribution. This normalisation ensures that the overall sum over all charge states formed by our oligomers will be equal to unity. The overall intensity of peaks reported by the mass spectrometer will be a function of the detector efficiency  $D_E$ , which varies with mass-to-charge ratio:

$$D_E(a, b, z, A_O, Q, V_A) = A_O \times \left( 1 - \exp \left( -1620 \times \left( \frac{V_A z}{M_E} \right)^{1.75} \right) \right) \quad [\text{Eqn. S6}]$$

where  $A_O$  is the open area of the MCP detector, and  $V_A$  is the accelerating voltage (Fraser, 2002). The overall detected signal intensity of a complex in a charge state  $z$  will therefore be

$$W_D(a, b, z, A_O, F, Q, V_A, Z_W) = W_C D_E \quad [\text{Eqn. S7}]$$

Each charge state gives rise to an approximately Gaussian peak in the overall mass spectrum (McKay, et al., 2006), centred on  $M_E/z$ . The peak width is given by

$$P_w(a, b, z, Q, R) = \frac{M_E}{z} \left( \frac{1}{R} \right)$$

[Eqn. S8]

where the parameter  $R$  accommodates the mass-to-charge-dependence of peak width (McKay, et al., 2006), a parameter optimized for each individual mass spectrum. At any given mass-to-charge ratio  $x$  the contribution to the signal intensity  $S$  from a complex in charge state  $z$  will be

$$S(a, b, A_O, F, Q, R, V_A, Z_W) = W_D \exp\left(-\frac{(x - (M_E / z))^2}{2P_w^2}\right)$$

[Eqn. S9]

The overall signal  $S_C$  coming from all charge states of a given complex at a measured mass-to-charge ratio  $x$  is obtained by summing over all charge states

$$S_C(a, b, A_O, F, Q, R, V_A, Z_W) = \sum_{z=0}^{\infty} S$$

[Eqn. S10]

Considering the total solution concentration of a particular oligomer as  $C_{ab}$ , the total signal at a given  $x$  is given by

$$S_T(a, b, A_O, F, Q, R, V_A, Z_W) = \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} C_{ab} S_C$$

[Eqn. S11]

These above equations therefore provide a means for generating mass spectra for candidate distributions of protein complexes. In CHAMP this is implemented as a

function of only three variable oligomer-independent parameters,  $F$ ,  $Q$  and  $R$ , with  $A_O$ ,  $V_A$  and  $Z_W$  kept constant, see main text. As such the challenge therefore becomes to employ Eqn. S11 to determine the relative abundances of the individual complexes which contribute to a given experimental mass spectrum.

## S2 CHAMP fitting

To obtain the best fit between calculated and experimental mass spectra the set of oligomer concentrations  $C_{ab}$ , as well as the spectrum parameters  $F$ ,  $Q$  and  $R$ , are obtained by CHAMP that minimize the pseudo- $\chi^2$  function

$$\chi^2 = \sum_x^N (S_{\text{exp}} - S_T)^2$$

[Eqn. S12]

where there are  $N$  experimentally recorded  $m/z$  data-points. In all cases, Levenberg-Marquardt minimization methods are used, and the maximum intensity of the experimental spectrum  $S_{\text{exp}}$  and calculated spectrum  $S_T$  were both set to 100.

CHAMP has three levels of complexity (termed ‘surface’, ‘set’, and ‘free’ fits) of sequentially increasing effective resolution, but at the expense of correspondingly more fitting parameters (Fig. S1). As with any case where experimental data is compared to a model, increasing the number of fitting parameters will result in lower  $\chi^2$  and corresponding better fit. F-statistics can then be calculated to estimate whether the overall reduction in  $\chi^2$  incurred by introducing an additional free parameter can be statistically

justified. As discussed in the text, our finding here is that the appropriate complexity of model, unsurprisingly, depends on the effective resolution of the experimental data.

**Surface Fit:** A simple candidate for the distribution of oligomers is obtained by considering their concentrations as a two-dimensional Gaussian function:

$$C_{ab} = \exp\left(-\frac{(a-a_0)^2}{2a_w^2}\right) \exp\left(-\frac{(b-b_0)^2}{2b_w^2}\right)$$

[Eqn. S13]

which is parameterized by four fitting parameters,  $a_0$ ,  $a_w$ ,  $b_0$  and  $b_w$ . In the case of SHSP<sub>a</sub>:target<sub>b</sub> complexes here, Eqn. S13 was found to be an insufficient model to fit the experimental data. Motivated by our previous study (Stengel, et al., 2010), we therefore established the surface fit instead as the slightly more complex form in which the concentration of complexes at a given  $a$  to scale as a Gaussian, and the distribution of complexes at a given  $b$  to vary by a skewed Gaussian (Fig. S1), such that

$$C_{ab} = \exp\left(-\frac{(b-b_0)^2}{2b_w^2}\right) \exp\left(-\frac{(a-a_0)^2}{2a_w^2}\right) \int_{-\infty}^{\frac{a-a_0}{a_w}} \exp\left(\frac{-t^2}{2}\right) dt$$

[Eqn. S14]

Such a distribution is parameterized by five parameters  $a_0$ ,  $a_w$ ,  $b_0$ ,  $b_w$  and a ‘skew’ parameter,  $\alpha$ . The inclusion of the parameters  $F$ ,  $Q$  and  $R$  stemming from the model (see S1) means therefore that the surface fit employs a total of 8 different fitting parameters.

Blind minimization of Eqn. S12 with Eqn. S14 from random starting parameters with steepest descent algorithms was found to converge on a vast range of solutions of widely varying quality of fit, demonstrating a rugged  $\chi^2$  surface. To increase the likelihood that it finds a global rather than local  $\chi^2$  minimum, CHAMP performs a grid search at fixed values of  $a_0$  and  $b_0$  and steepest descent of the remaining parameters. This approach was found to be a robust method for approaching the global minimum  $\chi^2$  of the surface fit.

**Set Fit:** The next level of complexity is reached by separating the two-dimensional skewed Gaussian distribution obtained from the surface fit into a set of independent one-dimensional skewed Gaussian ‘slices’, one for each (integer) value of  $b$  (Fig. S1), each of the form

$$C_{ab} = N(b) \exp\left(-\frac{(a - a_0(b))^2}{2a_w^2(b)}\right) \int_{-\infty}^{\frac{a - a_0(b)}{a_w(b)}} \exp\left(-\frac{t^2}{2}\right) dt$$

[Eqn. S15]

Each ‘slice’ is defined by parameters, and therefore the total number of free parameters describing the distribution is given by  $(4 \times B) - 1$ , where  $B$  is the total number of different stoichiometries of  $P$  considered. The subtraction of 1 degree of freedom reflects the way the spectra are normalized.

**Free fit:** The highest level of complexity is represented by a distribution in which the concentration of each species  $P_aP'_b$  is considered independently (Fig. S1). This is implemented within CHAMP by taking the result of the set fit and varying each concentration  $C_{ab}$  independently. It is important to note that, because steepest descent minimization is used, both the set and free fits are guided by the initial grid-searching performed to generate the surface fit.

**Uncertainties:** To test the reproducibility of CHAMP for interpreting the spectra of polydisperse ensembles we interrogated a series of seven nanoES replicates obtained for the sHSP  $\alpha$ B-crystallin. This protein populates a Gaussian-like distribution of in excess of 30 oligomeric states detectable at equilibrium (Baldwin, et al., 2011). Employing CHAMP to analyse these spectra ( $b = 0$ ), allows us to extract the reproducibility of the centroid  $a$ , and width  $a_w$  of the distribution for a series of independent replicates (Fig. S2B). We find the procedure to be very reproducible, with standard deviations in  $a$  and  $a_w$  of 3.7% and 9.5% of the mean, respectively.

Furthermore, using the minimum values  $\chi^2_{\min}$  of the pseudo- $\chi^2$  function defined above (Eqn. S12), we can estimate the corresponding values of uncertainty  $\sigma$  on each data-point from

$$\frac{\chi^2_{\min}}{d} = \sigma^2$$

[Eqn. S16]

where the degrees of freedom,  $d$ , is the difference between the number of independent experimental measurements and the number of free parameters of the model, and by noting that if we have an optimally fitting model, the overall reduced- $\chi^2$  will be equal to unity. The value of  $\sigma$  using the set fit was generally found to be on the order of 2%, supporting the notion that the models used are giving reasonable reproductions of our experimental data.

### S3 Supplemental References

- Baldwin, A.J., Lioe, H., Robinson, C.V., Kay, L.E., and Benesch, J.L.P. (2011).  $\alpha$ B-crystallin polydispersity is a consequence of unbiased quaternary dynamics. *J. Mol. Biol* 413, 297-309.
- de la Mora, J.F. (2000). Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Anal Chim Acta* 406, 93-104.
- Dobo, A., and Kaltashov, I.A. (2001). Detection of multiple protein conformational ensembles in solution via deconvolution of charge-state distributions in ESI MS. *Anal Chem* 73, 4763-4773.
- Fraser, G.W. (2002). The ion detection efficiency of microchannel plates (MCPs). *Int J Mass Spectrom* 215, 13-30.
- Heck, A.J.R., and van den Heuvel, R.H.H. (2004). Investigation of intact protein complexes by mass spectrometry. *Mass Spectrom Rev* 23, 368-389.
- Hernández, H., and Robinson, C.V. (2007). Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat Protoc* 2, 715-726.
- Kaltashov, I.A., and Mohimen, A. (2005). Estimates of protein surface areas in solution by electrospray ionization mass spectrometry. *Anal Chem* 77, 5370-5379.
- Kebarle, P., and Verkerk, U.H. (2010). On the Mechanism of Electrospray Ionization Mass Spectrometry (ESIMS). In *Electrospray and MALDI Mass Spectrometry: Fundamentals, Instrumentation, Practicalities, and Biological Applications*, R.B. Cole, ed. (Wiley).

- Li, Y., and Cole, R.B. (2003). Shifts in peptide and protein charge state distributions with varying spray tip orifice diameter in nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* 75, 5739-5746.
- McKay, A.R., Ruotolo, B.T., Ilag, L.L., and Robinson, C.V. (2006). Mass measurements of increased accuracy resolve heterogeneous populations of intact ribosomes. *J Am Chem Soc* 128, 11433-11442.
- Miller, S., Lesk, A.M., Janin, J., and Chothia, C. (1987). The accessible surface area and stability of oligomeric proteins. *Nature* 328, 834-836.
- Stengel, F., Baldwin, A.J., Painter, A.J., Jaya, N., Basha, E., Kay, L.E., Vierling, E., Robinson, C.V., and Benesch, J.L.P. (2010). Quaternary dynamics and plasticity underlie small heat shock protein chaperone function. *Proc Natl Acad Sci U S A* 107, 2007-2012.