Supplementary information

R. G. Smock, O. Rivoire, W. P. Russ, J. F. Swain, S. Leibler, R. Ranganathan, and L. M. Gierasch

CONTENTS

I. SUPPLEMENTARY TEXT

The definition of an inter-domain sector in the Hsp70/110 family is based on the principles introduced in Refs. (1) and (2). The alignment of this family does not, however, satisfy the condition of "good sampling" of the sequences required by the approach taken in Ref. (2). This supplementary note presents our approach to the problem raised by this alignment. A more detailed and general description of the SCA approach to inferring protein sectors from multiple sequence alignments will be reported elsewhere (O. Rivoire, S. Leibler and R. Ranganathan, in preparation).

A. Framework for analyzing positional and sequence correlations

As in previous work (2), we start from a binary approximation of the alignment, which is represented by a binary array X with elements satisfying $X_{si} = 1$ if the most frequent amino at position i is present in sequence s, and $X_{si} = 0$ otherwise. X has size $M \times L$, where M denotes the number of sequences in the alignment, and L the number of positions. Correlations between conserved positions are measured by a SCA matrix $\tilde{C}$ with elements defined by[1]

$$\tilde{C}_{ij} = \phi_i \phi_j \left(\langle X_{si} X_{sj} \rangle_s - \langle X_{si} \rangle_s \langle X_{sj} \rangle_s \right), \tag{1}$$

where $\langle \cdot \rangle_s$ denotes an average over the M sequences of the alignment, and the weights ϕ_i are defined as in previous work (1) by

$$\phi_i = \ln \frac{f_i (1 - q^{(a_i)})}{(1 - f_i) q^{(a_i)}}. \tag{2}$$

ϕ_i quantifies the deviation of the frequency $f_i = \langle X_{si} \rangle_s = \sum_s X_{si}/M$ of the prevalent amino acid a_i at position i from the expected frequency $q^{(a_i)}$ of this amino acid; it is a positive and increasing function of f_i when $f_i \geq q^{(a_i)}$ (a condition that holds for all positions here). The SCA matrix of positional correlations can also be written

$$\tilde{C} = \tilde{X}^\top \tilde{X}/M \tag{3}$$

where the $M \times L$ matrix $\tilde{X}$ is defined by

$$\tilde{X}_{si} = \phi_i (X_{si} - \langle X_{si} \rangle_s), \tag{4}$$

and where $\tilde{X}^\top$ denotes the transpose of $\tilde{X}$. Correspondingly,

$$\tilde{S} = \tilde{X} \tilde{X}^\top / L. \tag{5}$$

defines a matrix of correlations between sequences. $\tilde{S}_{st}$ measures the similarity between two sequences s and t with the contribution of each position weighted based on its conservation: differences between sequences occurring at more

[1] In (2), we used $|\tilde{C}|$ instead of $\tilde{C}$. The difference is, however, minor and for simplicity we do not take the absolute value.

conserved positions are thus emphasized.

A relation between the eigenvectors of $\tilde{C}$ and those of $\tilde{S}$ is given by the singular value decomposition of $\tilde{X}$. This decomposition always exists and has the form

$$\tilde{X} = U\Sigma V^\top, \tag{6}$$

where U and V are respectively $L \times L$ and $M \times M$ orthogonal matrices, and Σ is a $L \times M$ diagonal matrix. From this decomposition and Eqs. (3)-(5) it follows that $\tilde{C} = VDV^\top$ and $\tilde{S} = U\Delta U^\top/L$, where $D = \Sigma^\top\Sigma/M$ and $\Delta = \Sigma\Sigma^\top/L$ are both diagonal matrices: the columns of U thus correspond to the eigenvectors $|U_1\rangle, |U_2\rangle, \dots$ of $\tilde{S}$ and those of V to the eigenvectors $|V_1\rangle, |V_2\rangle, \dots$ of $\tilde{C}$.

Following previous work (2), sectors are defined as linear combinations of the top eigenvectors $|V_1\rangle, |V_2\rangle, \dots, |V_k\rangle$ of $\tilde{C}$. Given that statistical independence in the alignment is a defining property of protein sectors, relevant linear combinations can be sought using independent component analysis (ICA), a method for recovering statistically independent signals (see e.g. (3)). Starting from the top k eigenvectors $|V_1\rangle, |V_2\rangle, \dots, |V_k\rangle$ of $\tilde{C}$, ICA calculates a $k \times k$ unmixing matrix W^p, defined so that the vectors $|V_1^p\rangle, |V_2^p\rangle, \dots, |V_k^p\rangle$ obtained by $|V_n^p\rangle = \sum_{m=1}^k W_{nm}^p |V_m\rangle$ are maximally independent[2]; details on the algorithm used in this study are presented in Sec. I.C.

If a sector is associated with $|V_1^p\rangle = \sum_{m=1}^k W_{1m}^p |V_m\rangle$, the same linear transformation W^p applied to the eigenvectors of $\tilde{S}$, namely $|U_1^p\rangle = \sum_{m=1}^k W_{1m}^p |U_m\rangle$, indicates a direction in the sequence space along which the sequences are classified based on differences in these sector positions. Alternatively, we may also apply ICA to obtain an unmixing matrix W^s based on the top k eigenvectors of the matrix of sequence correlations $\tilde{S}$, instead of the matrix of positional correlations $\tilde{C}$. In this case, we expect the rotated vectors $|U_n^s\rangle$ to indicate subfamilies of sequences and the corresponding vectors $|V_n^s\rangle$ to point to positions that have distinct patterns of amino acids in these subfamilies. We refer here to these two complementary ways of analyzing $\tilde{X}$ as "sequence ICA" (seqICA) and "positional ICA" (posICA).

B. Correlations in the Hsp70/110 family

Inspection of the spectrum of $\tilde{C}$ for the Hsp70 alignment indicates that 4 eigenvalues clearly stand out (Fig. S1). A projection of the sequences along the top 2 eigenvectors $|U_1\rangle$ and $|U_2\rangle$ of $\tilde{S}$ reveals a structured distribution of the sequences with mainly 4 subfamilies of sequences[3] (Fig. S2). $|U_3\rangle$ reinforces this conclusion by making more apparent the distinction between the cyan and white subfamilies, while $|U_4\rangle$ displays subgroups of sequences within the white subfamily (Fig. S2). ICA applied to the vectors $|U_1\rangle, |U_2\rangle, |U_3\rangle$ leads to an unmixing matrix W^s, which maps $|U_1\rangle, |U_2\rangle, |U_3\rangle$ to new vectors $|U_1^s\rangle, |U_2^s\rangle, |U_3^s\rangle$ (seqICA). These maximally independent directions can be used to define subfamily of sequences. Thus, we colored here the sequences in purple, white, cyan or orange based on Fig. S3.

The 4 subfamilies can be interpreted from the available annotations of the sequences: the white and cyan subfamilies correspond to the two subclasses of Hsp70 proteins known as DnaK and HscA. The orange subfamily is found to comprise chaperones from several other classes of Hsp70 proteins and the purple subfamily proteins that are non-allosteric, including proteins from the Hsp110 family. More subdivisions are found along the following eigenvectors of $\tilde{S}$ which can also be interpreted in terms of functional subfamilies (sequences from particular orthologous families of proteins) and phylogenic subfamilies (sequences from particular clades). The subdivisions displayed by $|U_4\rangle$ thus correspond to subgroups of DnaK sequences from different clades of bacteria.

The non-allosteric nature of the purple sequences found along $|U_1^s\rangle$ suggests that when the same unmixing matrix W^s is applied to the eigenvectors $|V_1\rangle, \dots |V_3\rangle$ of $\tilde{C}$, the resulting vector $|V_1^s\rangle$ should point to correlated positions involved in the allosteric interaction between the two domains, i.e., to positions consistently conserved in all Hsp70

---

[2] Note that ICA does not prescribe an order between the vectors $|V_n^p\rangle$, in contrast with principal component analysis (PCA) which orders the eigenvectors $|V_n\rangle$ in descending value of their associated eigenvalues.

[3] Note that in presence of a structured distribution of sequences, the first eigenvector of $\tilde{S}$, $|U_1\rangle$, may be associated with a particular subfamily rather than with a "global phylogenetic trend"; in such a case, it should not be subtracted as done in (2).

proteins except for the purple sequences. The vectors $|V_1^s\rangle, |V_2^s\rangle, |V_3^s\rangle$ are represented in Fig. S4 and are the basis for the definition of the allosteric sector (see also Fig. S5). Consistently, the same sector can be identified by posICA, i.e., ICA based on the eigenvectors $|V_1\rangle, \ldots, |V_3\rangle$ of $\tilde{C}$. This leads to an unmixing matrix W^p and transformed vectors $|V_1^p\rangle, \ldots, |V_3^p\rangle$ displayed in Fig. S6 (both the order and the sign of $|U_1^s\rangle, \ldots, |U_3^s\rangle$ and $|V_1^p\rangle, \ldots, |V_3^p\rangle$ are arbitrary and were chosen here to facilitate the comparison between the various figures). Note that the equivalence between the results of seqICA and posICA regarding $|V_1^s\rangle$ and $|V_1^p\rangle$ may not hold for other protein families.

The use of the binary approximation of the MSA is a necessary simplification of the MSA for usage of the singular value decomposition method described in this work. Generalization to consider the full, unreduced alignment will be subject of future work. However, we note that for instances such as the Hsp70/110 family in which the function of interest (e.g. allostery) is a property of a major subfamily, sector identification is robust to the binary approximation (2).

C. ICA calculations

Different implementations of ICA use different measures of independence and different algorithms for optimizing them. In this work, we used one of the simplest implementations of ICA, proposed in Ref. (4) with modifications introduced in Ref. (5) (we also verified that the results were robustly recovered when using other algorithms for ICA). For seqICA, the input of the algorithm is the $k \times M$ matrix Z whose rows correspond to $|U_1\rangle, \ldots, |U_k\rangle$, while for posICA it is the $k \times L$ matrix Z whose rows correspond to $|V_1\rangle, \ldots, |V_k\rangle$. The algorithm iteratively updates the unmixing matrix W, starting from the $k \times k$ identity matrix $W = I_k$, with increments ΔW given by

$$\Delta W = \epsilon \left(I_k + \left(1 - \frac{2}{1 + \exp(-WZ)} \right) (WZ)^\top \right) W. \tag{7}$$

The parameter ϵ is a learning rate that has to be sufficiently small for the iterations to converge; in the present study, we found that with $\epsilon = 10^{-4}$ both seqICA and posICA converged after 1000 iterations (for $k = 3$ components). The iterations lead to W^s for seqICA and W^p for posICA. The vectors $|U_n^s\rangle$ and $|V_n^s\rangle$ are obtained by applying W^s to $|U_n\rangle$ and $|V_n\rangle$, and the vectors $|U_n^p\rangle$ and $|V_n^p\rangle$ by applying W^p to them. The resulting vectors were finally normalized to unit length[4].

SUPPLEMENTARY REFERENCES

[1] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286:295–299, 1999.

[2] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–86, 2009.

[3] J. V. Stone. *Independent component analysis: A tutorial introduction*. The MIT Press, 2004.

[4] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind source separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[5] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in neural information processing systems*, volume 8, pages 757–763, Cambridge MA, 1996. MIT Press.

[4] Note that in contrast with seqICA, the input matrix Z for posICA is not centered: $\sum_s \langle s|U_n\rangle = 0$ but in general $\sum_i \langle i|V_n\rangle \neq 0$. In this general case, it may be appropriate to subtract the mean and "sphere" the data so that $ZZ^\top = I_k$ before computing an unmixing matrix W^p. In the present case however, these operations only result in an irrelevant translation of the vectors $|V_n^p\rangle$ and $|U_n^p\rangle$.

II. SUPPLEMENTARY TABLE

position i	$\langle i\|V_1^s\rangle$	position i	$\langle i\|V_1^s\rangle$	position i	$\langle i\|V_1^s\rangle$
7	0.070	340	0.052	468	0.088
9	0.062	344	0.052	470	0.060
11	0.100	353	0.051	471	0.060
12	0.101	367	0.111	472	0.060
14	0.078	371	0.069	474	0.093
16	0.050	375	0.068	475	0.064
32	0.088	378	0.089	476	0.081
38	0.058	384	0.092	479	0.102
62	0.064	390	0.064	480	0.067
112	0.069	391	0.097	482	0.070
116	0.084	392	0.079	486	0.080
127	0.059	394	0.093	488	0.104
132	0.072	396	0.069	494	0.073
138	0.074	397	0.102	499	0.080
139	0.051	398	0.050	512	0.069
141	0.103	399	0.086	515	0.111
144	0.088	400	0.099	519	0.056
145	0.109	401	0.051		
146	0.066	402	0.100		
148	0.059	403	0.099		
152	0.092	405	0.096		
154	0.113	406	0.091		
155	0.096	412	0.093		
165	0.075	414	0.062		
171	0.080	415	0.087		
172	0.111	416	0.076		
175	0.061	417	0.064		
178	0.081	418	0.078		
181	0.061	424	0.075		
182	0.080	426	0.052		
192	0.051	428	0.114		
195	0.092	431	0.085		
197	0.106	433	0.122		
198	0.102	436	0.050		
199	0.105	438	0.071		
200	0.110	440	0.092		
201	0.115	442	0.062		
205	0.074	443	0.099		
216	0.094	444	0.083		
217	0.089	445	0.114		
218	0.064	450	0.075		
221	0.063	454	0.081		
223	0.106	457	0.079		
225	0.051	459	0.094		
227	0.064	462	0.097		
231	0.066	463	0.079		
269	0.072	464	0.061		
316	0.056	466	0.050		
326	0.089	467	0.055		

TABLE S1 - List of allosteric sector positions i and magnitude of contribution to the first independent component of the $\tilde{C}$ positional correlation matrix $|V_1^s\rangle$.

III. SUPPLEMENTARY FIGURES

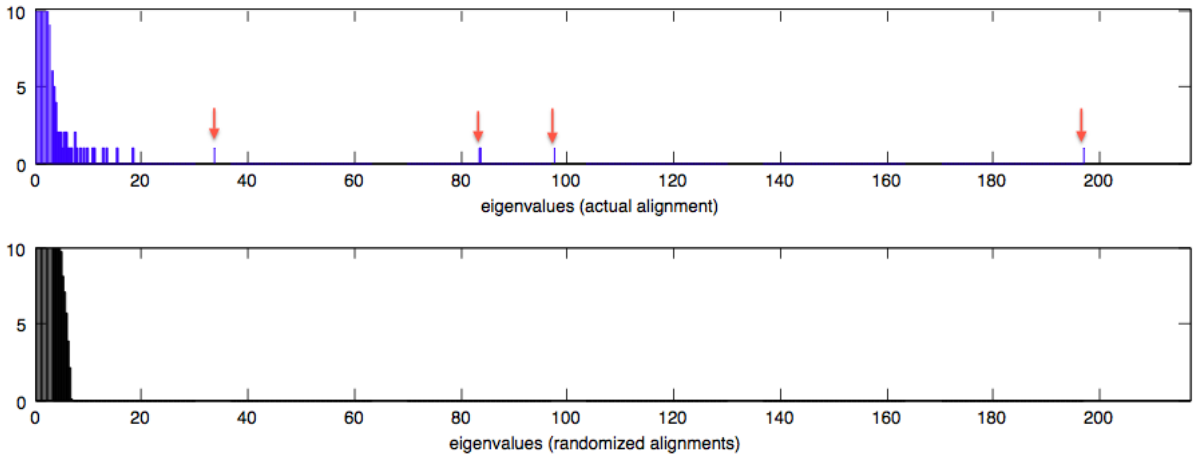

FIG. S1 - Histogram of eigenvalues of $\tilde{C}$ for the actual alignment (top) and randomized alignments (bottom). From the comparison, 10 to 12 eigenvalues may be considered statistically significant. In the top figure, 4 eigenvalues stand apart very clearly.

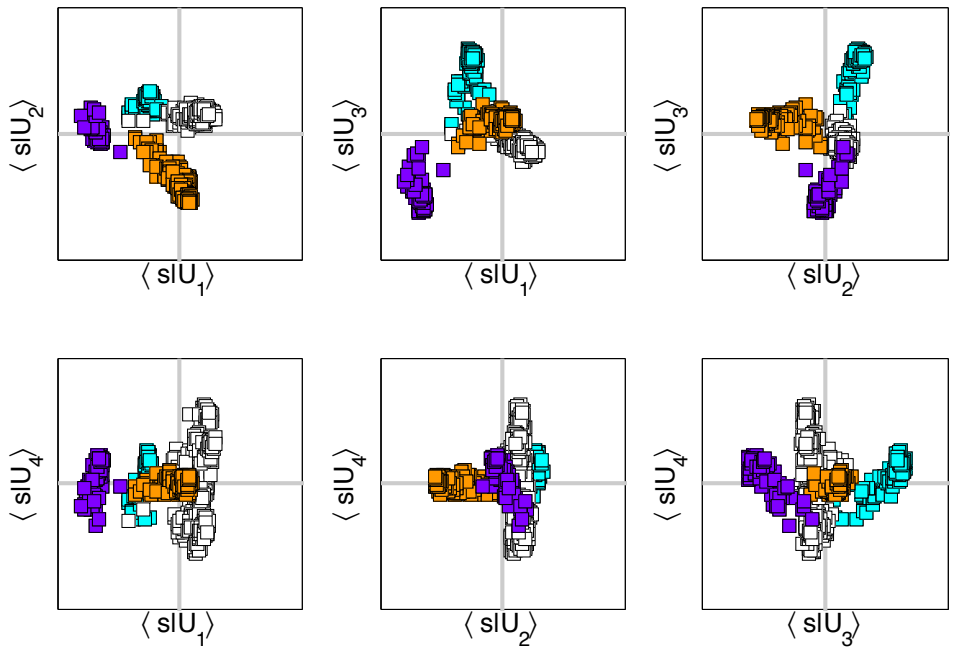

FIG. S2 - Distribution of the sequences represented by projection on the top 4 eigenvectors of $\tilde{S}$. The different panels present all pairwise combinations of the 4 vectors $|U_1\rangle, |U_2\rangle, |U_3\rangle, |U_4\rangle$. The first panel represents for instance each sequence s as a square with two-dimensional coordinates $(\langle s|U_1\rangle, \langle s|U_2\rangle)$. Four subfamilies of sequences are apparent along $|U_1\rangle, |U_2\rangle, |U_3\rangle$, while $|U_4\rangle$ exposes variation within the white sequences (colors assigned based on Fig. S3). See also Box Fig. 3A for a 3-D representation of the first three panels

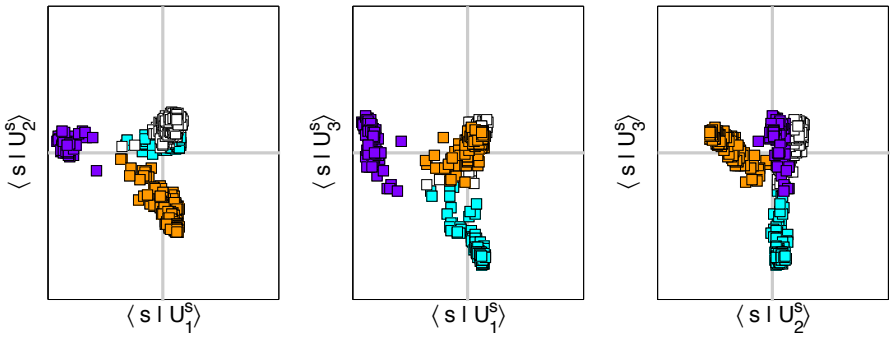

FIG. S3 - Sequences projected along $|U_1^s\rangle, |U_2^s\rangle, |U_3^s\rangle$, the maximally independent axis of sequence variations obtained by seqICA, i.e., by applying ICA to the top 3 eigenvectors of $\tilde{S}$ (unmixing matrix W^s). Colors were assigned to the sequences based on these projections. Note that this assignment of colors is here to facilitate the visualization of the subfamilies and to help the comparison with Fig. S1, but plays no role in the analysis. See also Box Fig. 3B for a 3-D representation of these panels

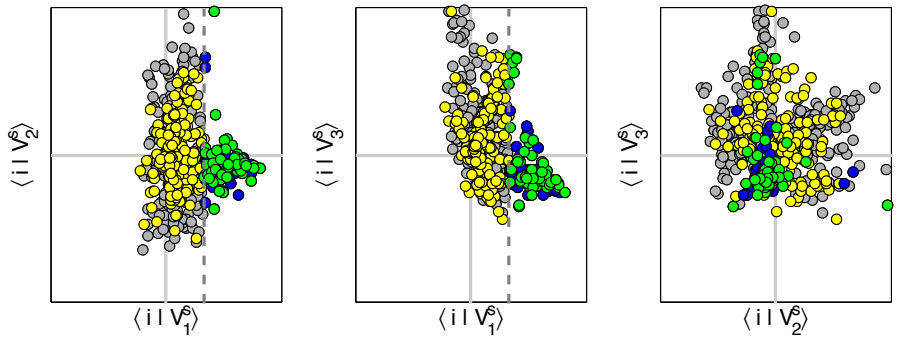

FIG. S4 - Positions projected along $|V_1^s\rangle, |V_2^s\rangle, |V_3^s\rangle$, the axis of positional variations obtained by seqICA, i.e., by applying W^s to the 3 top eigenvectors of $\tilde{C}$. Non-sector positions are represented in gray if belonging to the nucleotide-binding domain and in yellow if belonging to the substrate-binding domain (colors as in Fig. 1). A sector is defined as the group of positions i satisfying $\langle i|V_1^s\rangle > \epsilon$, where $\epsilon = 0.05$ indicates a threshold of statistical significance (dashed vertical line); sector positions are represented in blue if belonging to the nucleotide-binding domain and in green if belonging to the substrate-binding domain (colors as in Fig. 2). Note that not all the positions are clearly visible here as some are plotted on top of others.

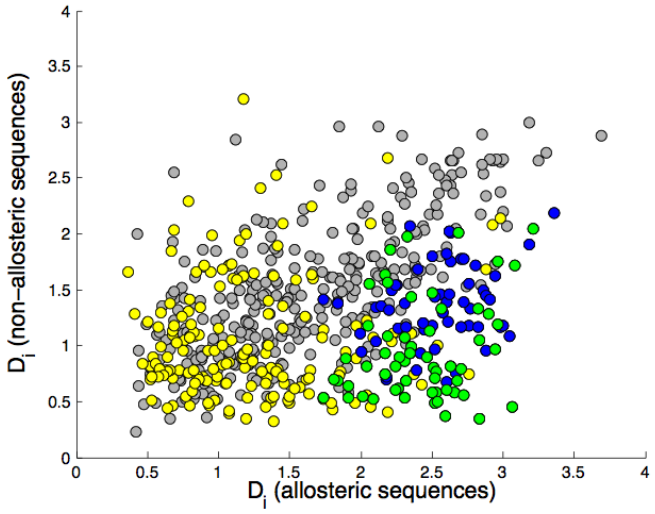

FIG. S5 - Comparison of conservation of each position in the allosteric sequences (white, cyan, and orange in Fig. S3) and non-allosteric sequences (purple in Fig. S3). Conservation of position i is measured using the relative entropy D_i (see Ref. (2)). The sector positions, in blue or green, are conserved in the allosteric sequences but not in the non-allosteric sequences, consistently with the interpretation that the sector represents a conserved functional property specific to the allosteric sequences (colors as in Fig. S4).

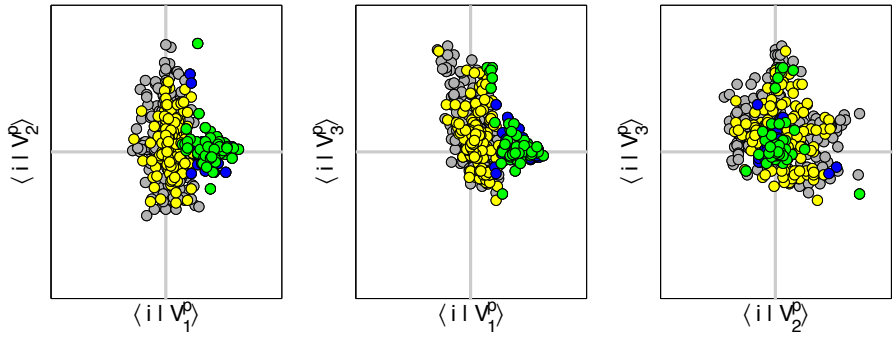

FIG. S6 - Positions projected along $|V_1^p\rangle, |V_2^p\rangle, |V_3^p\rangle$, the maximally independent axis of positional variations obtained by posICA, i.e., by applying ICA to the top 3 eigenvectors of $\tilde{C}$ (unmixing matrix W^p). The colors are as in Fig. S4, showing that the sector can be obtained either by seqICA or posICA with only minor differences.

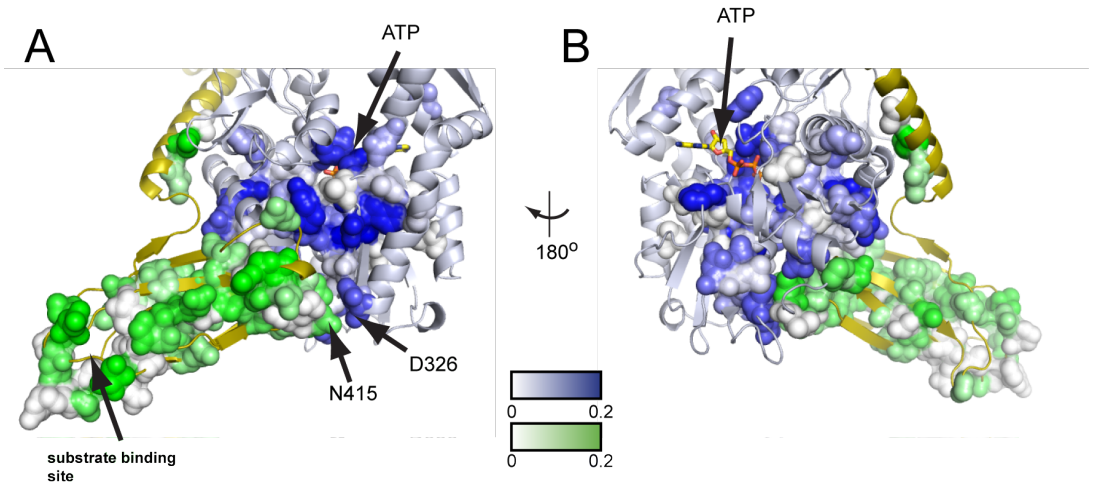

FIG. S7 - Two views of the allosteric sector mapped onto a model of the ATP-bound state of E. coli DnaK. Sectors positions are colored as in Fig. 2C (blue for the nucleotide binding domain and green for the substrate binding domain, but with a color gradient that is proportional to the magnitude of contribution of each position to the allosteric sector (i.e. its weight along the first independent component of the $\tilde{C}$ positional correlation matrix ($|V_1^S\rangle$)). See also Fig. 2B and Supplementary Table 1.

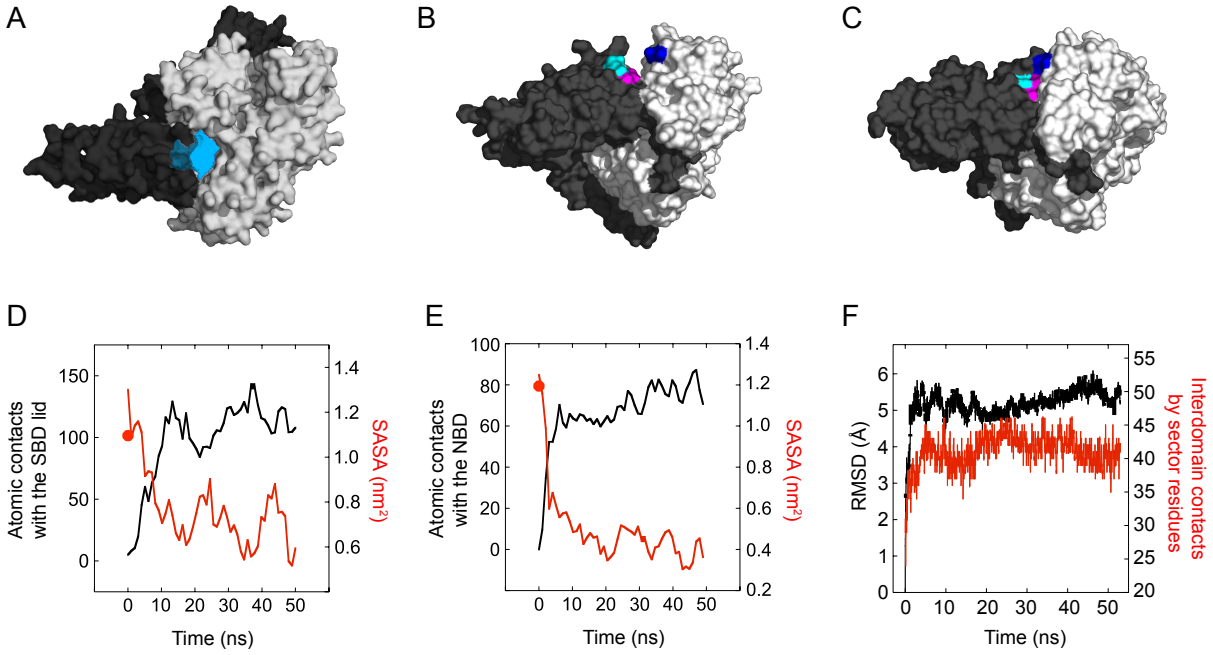

FIG. S8 - Modeling Hsp70 structure in an ATP-bound, domain-docked conformation. **(A)** An initial Hsp70 homology model. Threading the amino acid sequence of the *E. coli* Hsp70 DnaK directly onto an Hsp110(ATP) crystal structure (PDB code 2QXL) leads to poorly optimized interdomain contacts with a cavity of 286 Å^3 volume within the interface (shown in blue). **(B)** Prior to simulation, the domain surfaces near sector residues D326 (blue), K414 (cyan) and N415 (magenta) are not contiguous in the DnaK(ATP) homology model. **(C)** During simulation, the interdomain cavity collapsed and the interdomain sector positions joined into a single contiguous group across the interface (shown on the median structure of the trajectory). **(D-E)** Molecular dynamics simulation is accompanied by formation of specific interdomain contacts consistent with experimental data. Nucleotide-binding domain residue W102 (D) and substrate-binding domain residue K414 (E) initially have solvent-accessible surface areas (SASA) nearly the same as in single-domain DnaK crystal structures (red circles, PDB codes 1DKG and 1DKZ). As the simulation progresses, both sites become buried through the formation of interdomain contacts. Data are plotted as moving averages with a window size of 1 ns. Solvent-exposed residues are defined as sites with fractional side chain accessible surface area greater than 0.25 relative to extended Gly-Xaa-Gly peptides, as calculated by the VADAR web server using the default parameters. **(F)** Overall, the number of interdomain contacts between sector residues increases as structural changes (RMSD) reach a plateau. Interdomain sector contacts are calculated as sector positions of separate domains that contain side chain atoms within 5.0 Å of each other.

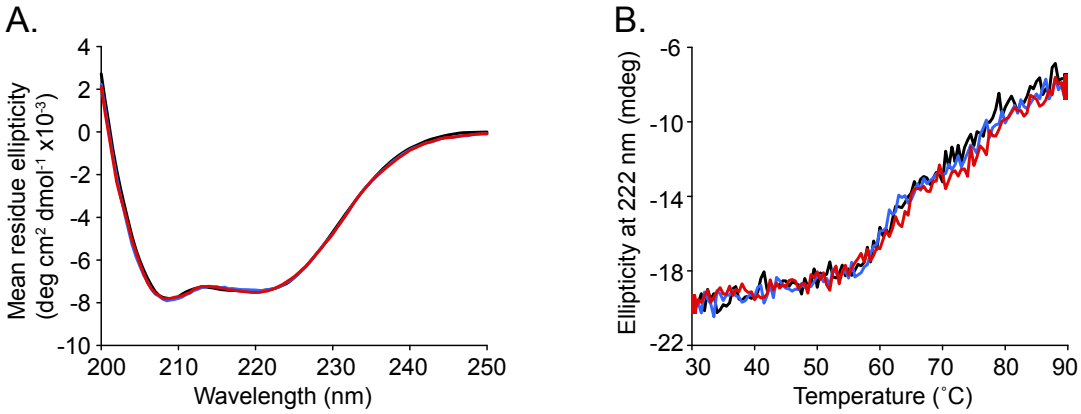

FIG. S9 - Sector mutants DnaK D326V and N415G retain native secondary structural content and are thermally stable. **(A)** Circular dichroism (CD) spectra and **(B)** thermal denaturation curves for D326V and N415G mutant DnaK proteins are indistinguishable from those of wild type, indicating that these sector mutants are well-folded (wild type, black; D326V, blue; N415G, red). Sector mutant CD spectra in (A) showed less than 5% intensity difference from that of wild type and were normalized to allow comparisons of curve shapes, which are diagnostic for secondary structural content.

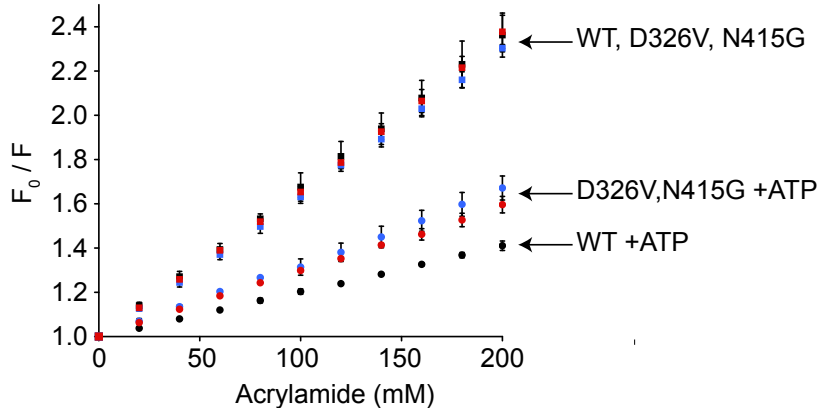

FIG. S10 - Sector mutants DnaK D326V and N415G show incomplete conversion to the ATP-bound state. Acrylamide quenching of W102 fluorescence intensity is similar for wild type DnaK and sector mutants in the absence of ATP. However, upon addition of ATP, sector mutants are partially defective in the extent to which they are quenched by acrylamide relative to wild type (wild type, black; D326V, blue; N415G, red).