Methods 53 (2011) 318–324

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

# Approaches for probing the sequence space of substrates recognized by molecular chaperones

Pradeep Kota, Nikolay V. Dokholyan *

*Department of Biochemistry and Biophysics, Program in Cellular and Molecular Biophysics, Center for Computational and Systems Biology, University of North Carolina at Chapel Hill, NC 27599-7260, USA*

ARTICLE INFO

*Article history:*
Available online 30 December 2010

*Keywords:*
Chaperones
Epitope identification
Neurodegenerative diseases
Eris

ABSTRACT

Neurodegeneration, the progressive loss of function in neurons that eventually leads to their death, is the cause of many neurodegenerative disorders including Alzheimer's, Parkinson's, and Huntington's diseases. Protein aggregation is a hallmark of most neurodegenerative diseases, where unfolded proteins form intranuclear, cytosolic, and extracellular insoluble aggregates in neurons. Mounting evidence from studies in neurodegenerative disease models shows that molecular chaperones, key regulators of protein aggregation and degradation, play critical roles in the progression of neurodegeneration. Although chaperones exhibit promiscuity in their substrate specificity, specific molecular features are required for substrate recognition. Understanding the basis for substrate recognition by chaperones will aid in the development of therapeutic strategies that regulate chaperone expression levels in order to combat neurodegeneration. Many experimental techniques, including alanine scanning mutagenesis and phage display library screening, have been developed and applied to understand the basis of substrate recognition by chaperones. Here, we present computational algorithms that can be applied to rapidly screen the sequence space of potential substrates to determine the sequence and structural requirements for substrate recognition by chaperones.

## 1. Introduction

Acquired or inherited mutations that result in modifications to protein structure can cause various genetic disorders, including cancers, where the mutant gene product interferes with cellular processes like synthesis, transport, stability, or enzymatic activity. In contrast, conformational diseases, which include neurodegenerative diseases like Alzheimer's (AD), Parkinson's (PD), Huntington's disease (HD), and prion encephalopathies, develop due to conformational rearrangements in a specific protein. These rearrangements lead to the formation of protein aggregates in ordered fibrillar structures known as amyloids, which can be found inside and outside of brain cells [1]. A histological feature shared by these diseases is the occurrence of lesions in brains, consisting of the intra- or extracellular accumulation of misfolded, aggregated, and ubiquitinated proteins that are intimately associated with neurodegeneration. In AD, lesions are observed as intracellular neurofibrillary tangles that contain hyperphosphorylated Tau protein and extracellular plaques that contain β-amyloid (Aβ) peptides [2]. In PD, the cytoplasmic lesions, called Lewy bodies, are composed primarily of the protein α-synuclein [3]. In HD, intranuclear and cytoplasmic inclusion bodies comprised of the protein huntingtin are formed [4]. Another critical feature shared by many neurodegenerative disorders is the presence in the brain lesions of molecular chaperones and components of the ubiquitin–proteasome degradation system [5,6]. The ubiquitous role of molecular chaperones in neurodegenerative disease presents an interesting conundrum: how do chaperones recognize their substrates among other proteins *in vivo*? Many hypotheses have been posited that led to the development of experimental techniques to determine the sequence/structural basis for substrate recognition by chaperones. Here, we review the most common computational techniques employed to rapidly identify critical determinants of protein–protein interactions that can directly be applied to chaperone–substrate systems.

### 1.1. Molecular chaperones in neurodegenerative diseases

Cells are equipped with potential mechanisms of defense against physical and chemical stress, the most common of which is elevated expression of a set of highly conserved genes that encode heat shock proteins (Hsps) [7,8]. This set of proteins, including members of the Hsp70, Hsp90, Hsp110, HSP104, Hsp40, and many small Hsps, function as molecular chaperones defending the cell against the accumulation of damaged or mutant proteins

* Corresponding author.
*E-mail address:* dokh@med.unc.edu (N.V. Dokholyan).

1046-2023/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2010.12.034

[9–11]. In addition to their role in protein folding, chaperones are involved in the translocation of many proteins across cell membranes and macromolecular assembly and disassembly, as well as facilitating the transfer of misfolded proteins to the proteasome for degradation [11–13]. Chaperones function by transiently binding to exposed hydrophobic surfaces in target proteins in an ATP-dependent manner, shielding them from aberrant interactions with other proteins and folding intermediates in the cellular environment [14]. Over-expression of chaperones of the Hsp70 and Hsp40 families has been shown to suppress the aggregation and toxicity of polyQ-containing proteins [6]. The most compelling evidence for the critical role of chaperones in conformational diseases comes from *in vivo* studies of neurodegenerative disorders in Drosophila [15–18]. In one study, the expression of expanded polyQ controlled by an eye-specific promoter caused severe degeneration of external eye structures, resulting in loss of the retina. This effect could be rescued by over-expression of human Hsp70 [18]. Additionally, mounting biochemical evidence establishes that chaperones play a critical role in neurodegeneration and aging. Hence, this class of proteins holds much promise as a therapeutic target for the treatment of neurodegenerative disorders [19].

### 1.2. Mechanism of chaperone-assisted protein (re)folding

The diverse functions of molecular chaperones typically involve iterative ATP-dependent substrate binding and release cycles, until the substrate has gained its active conformation or has been marked for degradation [20,21]. Chaperones from different Hsp families work in concert to carry out the multitude of cellular processes involving chaperone activity. Here, we focus on the Hsp70 chaperone machinery, which has been implicated in posttranslational protein assembly and translocation [22]. The Hsp70 machinery typically includes a member of the 70-kDa Hsp70 family, assisted by a DnaJ-like protein and a nucleotide exchange factor (NEF). DnaJ-like proteins (referred as "J-proteins" from here on) are 40 kDa molecular chaperones (Hsp40s) that specifically regulate Hsp70s by direct interaction via their J-domains [23,24]. The cofactors of Hsp70s (J-protein and NEF) are critical for their function, as they regulate binding of Hsp70 to its substrates. ATPase activity is stimulated by the J-protein, promoting substrate binding; dissociation of ADP is stimulated by the NEF, initiating the recycling of Hsp70 due to substrate release. Therefore, the canonical model of Hsp70 activity is marked by two iterative steps: first, the initial binding of an unfolded substrate by the J-protein prevents its aggregation and delivers it to Hsp70; second, the dissociation of the substrate caused by the nucleotide exchange provides an occasion for the substrate to fold into its native conformation [25,26]. These two steps are tightly regulated, since overstimulation by the J-protein may prevent substrate binding and excess of NEF activity may cause premature substrate release. An elaborate scheme of chaperone-assisted protein folding and degradation is shown in Fig. 1.

### 1.3. Structural classification of J-proteins

The only common feature among all J-proteins is the presence of the J-domain. The J-domain is typically (but not always) present at the N-terminus of the protein [23]. This small alpha-helical domain presents a highly conserved tripeptide of histidine, proline, and aspartic acid (HPD motif) that has been shown to interact with Hsp70 [27,28]. J-proteins can be divided into three types based on their structure. Type-1 J-proteins are descendents of *Escherichia coli* DnaJ, and feature an N-terminal J-domain followed by a Gly and Phe-rich unstructured region, four repeats of the CxxCxGxG-type zinc finger motif, and a C-terminal peptide-binding domain (PBD) [29,30]. The C-terminal domain (CTD) comprises two beta barrel topology domains, CTD I and CTD II. CTD I has a hydrophobic pocket implicated in substrate binding, along with the zinc finger motifs extruding from the CTD I [30,31]. The extreme C-terminus mediates dimerization of the J-proteins in solution [32]. J-proteins lacking the zinc-finger domains are classified as type-2, while the remaining J-proteins are designated type-3. However, one must exercise caution while referring to the three types, since the structural features do not reflect their functional characteristics.

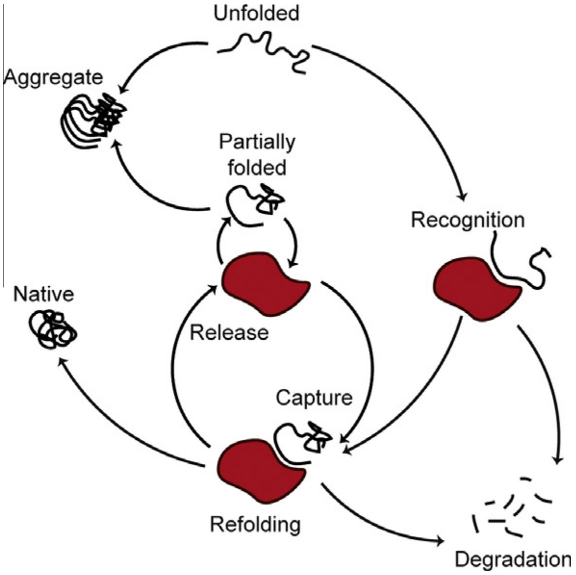

**Fig. 1.** Chaperone-assisted protein folding and degradation. Chaperones (red shape) recognize unfolded/misfolded non-native polypeptides or native proteins in the cell and subject them to refolding or degradation. Once the substrate is captured, it is transiently released to allow refolding. The capture-release cycles take place iteratively until the protein is folded to its native state. If the protein could not be refolded, it is targeted for degradation. The chaperone does not interact with the protein once it is folded to its native form.

### 1.4. Substrate recognition by chaperones

J-proteins with a DnaJ-like fold coordinate with other chaperones in order to facilitate protein folding in the cell [29]. However, the J-domain alone has been shown to substantially substitute for the entire J-protein and recruit the corresponding Hsp70 for efficient refolding, while in Hsp40s the C-terminal PBD mediates substrate recognition [33]. The sequence and local structure of the substrate plays a critical role in mediating specific interactions between the chaperone and the peptide. Structural determinants of enzyme–substrate interactions have been previously systematically analyzed [34]. However, not all residues or groups of residues contribute equally to the binding free energy ($\Delta G_{\text{binding}}$) of substrate binding. Accurate determination of $\Delta G_{\text{binding}}$ is necessary to assess the role of specific residues in the chaperone–substrate interface. Experimental techniques, including phage display screens and alanine scanning, have been developed and applied to many protein–peptide systems. Because the experimental determination of critical residues at the interface is time consuming and expensive, an effort has been made to achieve accurate, predictive computational methodologies for alanine scanning mutagenesis, capable of reproducing the experimental mutagenesis values. In order to apply these generic computational methods to understanding the peptide-binding characteristics of a chaperone of interest, it is essential to accurately calculate the binding free energies of the system and the effects of mutations on the stability of the complex using known three-dimensional structures. Numerous algorithms with varying complexity have been developed to accurately estimate the binding energy between macromolecules. These algorithms can be divided into two broad classes: (1) use of explicit atomistic simulations to estimate free energy changes, either directly or upon the mutation of certain residues in the interacting molecules, and (2) empirical functions that use knowledge-based or physical force fields to evaluate binding. The most rapid methods of estimation of binding energies are the empirical or knowledge-based scoring approaches, in conjunction with simple physical force fields [35,36]. Explicit simulations are computationally intensive and include both rigorous free energy perturbation [37], thermodynamic integration [38], and more approximate methods such as molecular mechanics/Poisson-Boltzmann surface area (MM-PBSA) [38,39]. Here, we review one protocol from each class in detail and present a case study of their application to chaperone–substrate systems in order to identify critical sites that determine substrate specificity.

## 2. Description of methods

### 2.1. Computational alanine-scanning

Alanine-scanning mutagenesis is a simple and widely used technique in the determination of the functional role of specific amino acid side chains in proteins [40]. Alanine is the residue of choice because it is most abundant and is present both in buried and exposed positions as well as in all secondary structural elements. Because it eliminates the side chain beyond the β-carbon atom, alanine substitution is helpful to study the role of specific side chains in imparting substrate specificity. In addition, alanine is chiral and, unlike glycine, sufficiently structurally rigid not to fundamentally alter the backbone configuration of the protein. However, alanine-scanning mutagenesis rapidly becomes cumbersome with the increasing length of the peptide under investigation.

Computational alanine scanning, introduced by Massova and Kollman, is an elegant alternative for rapidly screening parts of a protein to deduce the role of specific side-chains in protein–peptide interactions [41]. In this approach, independent molecular dynamics (MD) simulations from separate starting structures, one each of the native and the alanine mutant structure, are performed. The structure of the alanine mutant is modeled by truncating the coordinates of the desired residue at the $C_\beta$ atom in the crystal structure. Hydrogen atoms of the methyl group are added depending on the force field used for simulations. MD simulations of the starting complexes (wildtype and mutant), along with the individual components (protein and peptide), are then performed, and snapshots of the structure from simulation trajectories are saved at equally-spaced intervals. The binding free energies ($\Delta G_{\text{binding}}$) for each complex can then be estimated using the following expression:

$$\Delta G_{\text{binding}} = \Delta G(\text{complex}) - [\Delta G(\text{protein}) + \Delta G(\text{peptide})]$$

where the individual components are obtained from their corresponding simulations. The difference in free energies ($\Delta\Delta G$) represents the energetic contribution of the residue to the protein–peptide interface. The flow of events related to this algorithm is summarized in Fig. 2A. This MD protocol, where simulations are performed for both native and mutant proteins, is theoretically a representative of any structural rearrangements at the interface that may occur upon mutation, and hence represents the biological reality of an alanine mutation. This algorithm can be directly applied to a chaperone–substrate system in order to identify residues on the substrate peptide that determine the binding specificity of the substrate to the chaperone. Since the wild type and mutant complexes and their individual components are simulated independently of one another, this method results in a high standard deviation in the estimated energy values, which will likely increase with increase in the size of the protein–peptide complex. Massova and Kollman proposed a variant of the algorithm outlined above, in which independent simulations are not performed for the wild type and mutant complexes [41]. Instead, the side chain of the desired residue is truncated up to the $C_\beta$ atom in every frame of the simulation trajectory (Fig. 2B). This ‘post-process’ approach is guaranteed to decrease the standard deviation in estimation of binding free energies at the cost of inaccurate estimation in cases where local backbone rearrangements upon substrate recognition are expected. A systematic analysis of both approaches in comparison with experimental results shows that the post-process approach, although not suitable for all cases, offers superior accuracy and faster prediction than the traditional MD approach [42].

Incorporation of implicit solvation in place of explicit water in MD simulations will further reduce the time taken for such computational alanine scans. Recently, Moreira and colleagues introduced an improved methodology with reduced computational cost, involving molecular dynamics simulations performed in a continuum medium using the Generalized Born (GB) model. This approach has been used on multiple systems to predict $\Delta\Delta G_{\text{binding}}$ within 1 kcal/mol of experimentally determined values [43].

### 2.2. Monte Carlo driven approach for consensus motif determination

Phage display library screening has been shown to be a powerful tool for characterizing protein–peptide interactions. Using this methodology, large peptide libraries can be rapidly screened to isolate candidate peptides that are recognized by the protein of interest. One of the most successful applications of phage display has been the isolation of monoclonal antibodies using large phage antibody libraries [44]. This approach has been applied to identify candidate peptides that are recognized by the yeast chaperone Ydj1, a representative member of the type-1 DnaJ like proteins [45]. The results from the screen not only produced a candidate peptide library, but also provided valuable insight into the potential consensus sequence required for substrate-recognition by Ydj1. One drawback of this approach is the time taken for the characterization of peptides that are potential binding partners to a given chaperone. The same library may not be useful for screening with multiple chaperones of interest. Recently, we have employed computational methods to identify a consensus peptide-binding motif for Ydj1p, and experimentally verified this motif [31]. The trend observed in our computational studies agrees with the phage display screening conducted by Li and Sha, establishing the potential of predictive computational algorithms in determining the binding properties of proteins in general, and their specific application to understanding substrate recognition by chaperones.

The algorithm for determining whether a given peptide is a potential substrate for a chaperone is based on a Monte Carlo driven iterative search for the best combination of side chain orientations that energetically favor the configuration of the substrate in the chaperone–substrate complex. Such an algorithm can also be used to design novel peptides targeted for recognition by a specific chaperone. In this iterative approach, each position on the substrate is computationally mutated to all possible amino acids and the energetically most favorable residue is chosen for each position. This approach allows for screening the entire sequence space in order to identify potential substrate peptides. At each step, the stability of the chaperone–substrate complex is estimated after a chosen mutation. Estimation of the effect of mutation on protein stability is important, considering that mutagenesis is a central tool in molecular biology. The change in stability of the complex upon mutation ($\Delta\Delta G$) can then be computed using

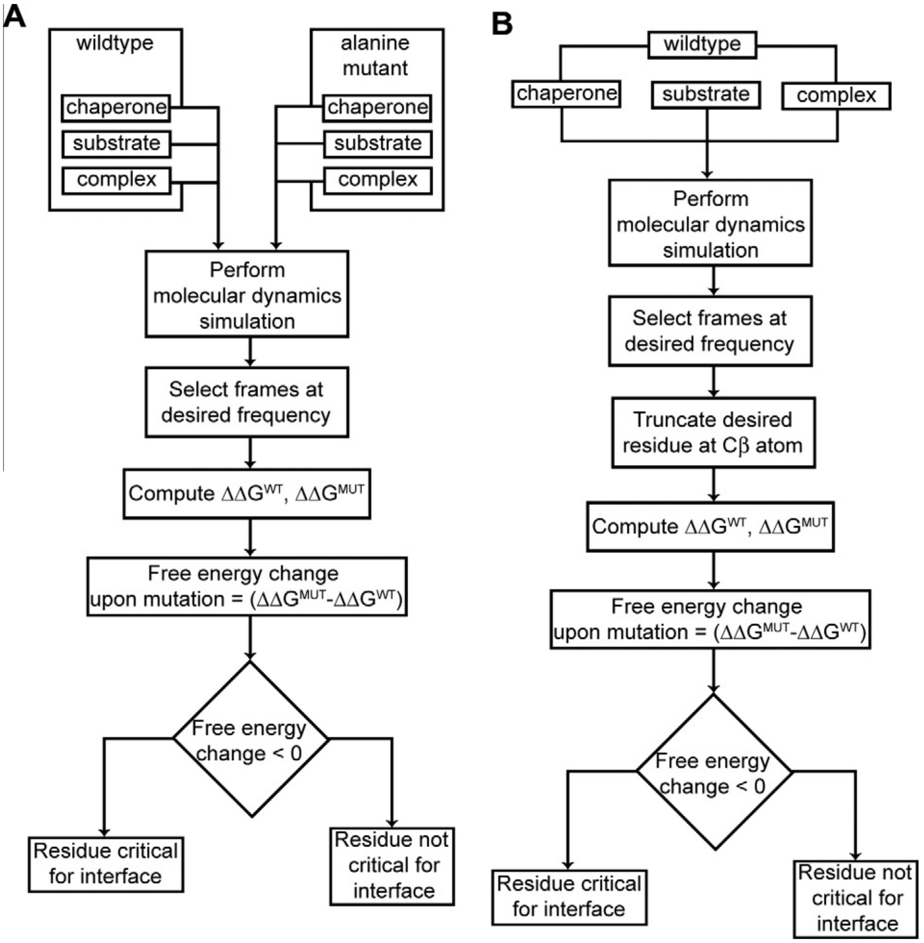
A
wildtype
chaperone
substrate
complex
alanine mutant
chaperone
substrate
complex
Perform molecular dynamics simulation
Select frames at desired frequency
Compute ΔΔG^WT, ΔΔG^MUT
Free energy change upon mutation = (ΔΔG^MUT-ΔΔG^WT)
Free energy change < 0
Residue critical for interface
Residue not critical for interface
B
wildtype
chaperone
substrate
complex
Perform molecular dynamics simulation
Select frames at desired frequency
Truncate desired residue at Cβ atom
Compute ΔΔG^WT, ΔΔG^MUT
Free energy change upon mutation = (ΔΔG^MUT-ΔΔG^WT)
Free energy change < 0
Residue critical for interface
Residue not critical for interface

**Fig. 2.** Flow chart for computational alanine-scanning mutagenesis. (A) Explicit MD simulations of individual components of both the wild type and alanine mutant are performed. Data is collected at discrete time points for computation of change in free energy upon mutation. The mutation is considered for further experimental analysis if it leads to a favorable change in free energy and is discarded otherwise. (B) Alternate flow of the MD protocol with improved accuracy in computation of free energy differences. Explicit MD simulations are performed only of the components of the wild type complex and the alanine mutants are constructed by truncating side chain atoms of the desired residue.

$$\begin{aligned}\Delta\Delta G^{\text{mutation}} &= \Delta\Delta G^{\text{MUT}}_{\text{Complex}} - \Delta\Delta G^{\text{WT}}_{\text{Complex}} \\ &= \left(\Delta G^{\text{WT}}_{\text{Chaperone}} - \Delta G^{\text{MUT}}_{\text{Substrate}}\right) - \left(\Delta G^{\text{WT}}_{\text{Chaperone}} - \Delta G^{\text{WT}}_{\text{Substrate}}\right)\end{aligned}$$

Here, a mutant is defined as a peptide in which one of the positions is mutated to a different amino acid, while the chaperone remains intact. The change in free energy ($\Delta\Delta G$) can be compared across different substrates to arrive at the best-fit peptide for the chaperone. This process is considerably faster than using the computational alanine scanning approach described above, because free energy estimation protocols using molecular dynamics simulation require several hours of computing time, and thus are inefficient when estimating $\Delta\Delta G$ from multiple mutations [37]. In addition, many heuristic force fields currently available are limited by the dataset on which the parameters are initially trained [46–49]. To avoid these limitations, we have developed a methodology that uses a physical force field (hence applicability is not limited by a training data set), Medusa, with atomic modeling as well as fast side chain packing and backbone relaxation algorithms [35,50].

In Medusa, the protein is modeled using the united atom model, which includes all heavy atoms as well as polar hydrogen atoms. The free energy of the chaperone–substrate complex is then expressed as a weighted sum of van der Waals forces, solvation, hydrogen bonding, and backbone-dependent statistical energies [35], as shown below.

$$\begin{aligned}E = {} & W_{vdw\_attr}E_{vdw\_attr} + W_{vdw\_rep}E_{vdw\_rep} + W_{solv}E_{solv} + W_{bb\_hb}E_{bb\_hb} \\ & + W_{sc\_hb}E_{sc\_hb} + W_{bb\_sc\_hb}E_{bb\_sc\_hb} + W_{\phi,\psi|aa}E_{\phi,\psi|aa} \\ & + W_{\phi,\psi,aa|rot}E_{\phi,\psi,aa|rot} - E_{ref}\end{aligned}$$

where $E$ is the total energy of the system; $E_{vdw\_attr}$, $E_{vdw\_rep}$ are the attractive and repulsive portions of the van der Waals interaction, respectively; $E_{solv}$ is the solvation energy; and $E_{bb\_hb}$, $E_{sc\_hb}$, and $E_{bb\_sc\_hb}$ are the hydrogen bonding energies among backbone atoms, side chain atoms, and between backbone and side chain atoms, respectively. $E_{\phi,\psi|aa}$ and $E_{\phi,\psi,aa|rot}$ correspond to the internal energy of an amino acid and its rotamer state given the backbone dihedrals, $\phi$ and $\psi$. $E_{ref}$ is the reference energy for the unfolded state. We use the weights ($W$) to estimate the contribution of each term to the total energy. The weighting parameters were independently trained to recapitulate the native amino acid sequences for 34 proteins using high-resolution X-ray structures [35].

### 2.3. Protein design with fixed backbone

In order to design the optimal side chain orientations for the residues in the substrate, we fix the protein's backbone and use a Monte Carlo-based simulated annealing approach to search for low-energy structural configurations of the substrate. Using the Metropolis criterion, we accept or reject a trial mutation – either an amino acid substitution or a side chain rotation – by computing the energy difference between the original and altered states. Using Monte Carlo, we accept a change in the state of the system if the change results in decrease of the total energy of the system. Additionally, we incorporate the Metropolis criterion in making decisions during the Monte Carlo simulation. The advantage of the Metropolis criterion is that it allows us to accept changes to the system despite an increase in energy with a given probability. In doing so, the system can be rescued from being trapped in a local energy minimum. This methodology improves the sampling efficiency in the overall energy landscape. During the last step of simulated annealing, we perform a quenching simulation, in which conjugate-gradient minimization is used to find the lowest energy state in the sub-rotameric conformation of each trial rotamer. Due to the stochastic nature of the design algorithm, one needs to perform multiple simulations in order to attain statistical significance of the outcomes. However, since the simulations are independent of one another, they can be performed in parallel, thereby decreasing the time taken for free energy estimation. Using this approach, we can rapidly identify the best-fit amino acid side chains for each position on a given substrate for the chaperone. Given the promiscuity of substrate recognition by chaperones, it is critical to evaluate all possible substitutions at all positions on the substrate before choosing the best-fit substrate for the chaperone. The steps involved in determining the sequence space that best represents the substrate sequences are outlined as a flow diagram in Fig. 3.

### 2.4. Backbone flexibility improves protein stability estimation

The drawback of fixing the protein backbone during redesign is that the methodology does not accurately estimate the change in free energy upon small side chain to large side chain substitutions and vice versa, where movement of the backbone is necessary to reach the energetic minimum. In order to be able to allow small to large as well as large to small side chain substitutions, we modeled backbone flexibility in our protein redesign algorithm. Proteins often adapt their structure to changes in the sequence by slightly reorienting the backbone. We apply the same in our design approach to achieve a realistic estimation of $\Delta\Delta G$ upon mutation. Backbone relaxation also helps to relieve any nonphysical atomic interactions in the protein that might bias the van der Waals energy, and hence the total energy, of the chaperone–substrate complex.

### 2.5. Eris – an automated tool for estimating stability changes upon mutation

We have developed a web-based tool, Eris (http://eris.dokhlab.org), for automated estimation of the change in protein stability upon mutation. We benchmarked Eris on 595 mutants with experimentally measured $\Delta\Delta G$ values. We observe that there is significant correlation (0.75; $P = 2 \times 10^{-108}$) between the Eris-estimated and experimentally determined $\Delta\Delta G$ values, an overall performance that is comparable with that of other methods [46–48]. However, unlike other methods, Eris accurately predicts the effect of small-to-large side chain mutations by effectively relaxing the backbone structures and resolving the clashes introduced by larger side chains. In direct comparison with other methods available on web servers, we computed stability changes upon small-to-large residue mutations, and we found that Eris outperforms other methods. Additionally, Eris provides a protein structure pre-relaxation option, which remarkably improves the prediction accuracy when a high-resolution protein structure is not available, such as when only a homology model of the protein exists. Because of its unbiased force field, side-chain packing, and backbone relaxation algorithms, Eris is applicable to a broad spectrum of mutations evaluated during protein engineering and design. Additionally, an integral step of Eris is backbone relaxation, when severe atom clashes or backbone strains are detected during calculation [50,51].

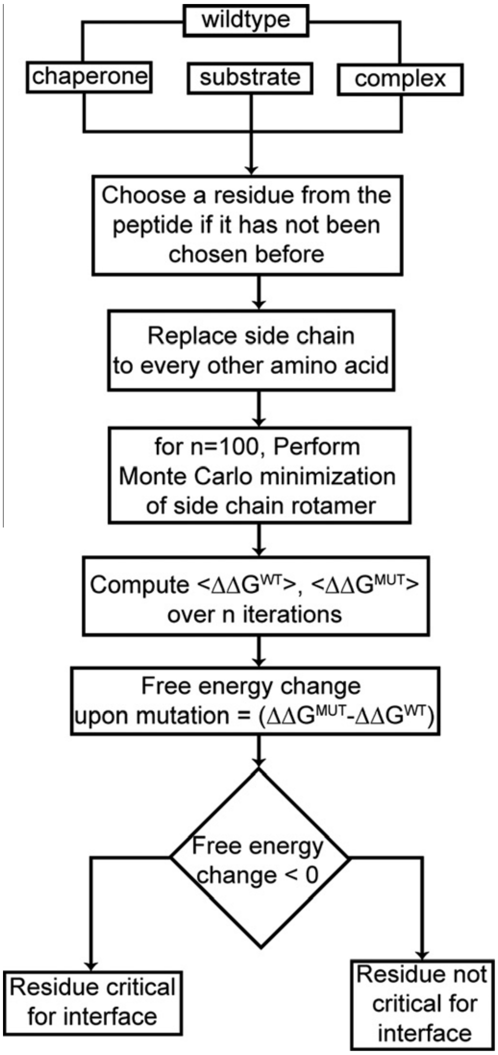
wildtype
chaperone
substrate
complex
Choose a residue from the peptide if it has not been chosen before
Replace side chain to every other amino acid
for n=100, Perform Monte Carlo minimization of side chain rotamer
Compute <ΔΔG^WT>, <ΔΔG^MUT> over n iterations
Free energy change upon mutation = (ΔΔG^MUT-ΔΔG^WT)
Free energy change < 0
Residue critical for interface
Residue not critical for interface

**Fig. 3.** Flow chart for side chain redesign using Medusa. Every amino acid in the peptide substrate is mutated to all other possible amino acids. This process is repeated several times for statistical significance. $\Delta\Delta G_{\text{binding}}$ is computed as the difference in free energies of the mutant and wild type complexes. Mutations that lead to a drop in $\Delta\Delta G$ to a value less than zero are considered stabilizing.

### 2.6. Ensemble-based profile prediction

In the Monte Carlo based approach outlined above, we fix the backbone of the chaperone and the substrate throughout the modeling exercise. However, we have also shown that incorporating backbone flexibility into our design principles improves estimation of the binding energy of a protein–peptide complex. A more recently reported methodology takes into account backbone flexibility of both the protein and the peptide, under the assumption that local conformational changes in the protein and/or peptide might improve the prediction of binding affinity between the two entities [52]. In this methodology, an ensemble of structurally distinct states is generated, reasoning that proteins are flexible to varying extents. The ensemble is used to generate a tolerance profile that represents a set of amino acids that are energetically tolerated at the protein–protein interface. The steps followed in this algorithm are described in detail below.

*(i)* Ensemble generation. Small changes in side chain rotamer orientation and/or hydrogen bonding partners lead to significant side chain motion perpendicular to the chain direction. As a result, the corresponding residue and its adjacent peptides twist slightly around the backbone. This phenomenon is termed "backrub" motion in proteins [53]. Backrub is observed in sub-Angstrom resolution crystal structures, where the alternate conformations of the protein backbone exhibit a highly localized plasticity of small amplitude coupled to a larger, two-state conformational change of the side chain [53]. Smith and Kortemme take advantage of this minor perturbation in the protein backbone to computationally generate an ensemble of structurally distinct states using Monte Carlo simulations. In these simulations, side chain rotamer moves are performed, retaining the lowest-scoring structures obtained during the simulation [52]. Smith and Kortemme used the Rosetta protein-modeling suite with backrub and side chain rotamer moves to generate the structural ensemble of states [54].

*(ii)* Profile prediction. A tolerance profile for a given set of positions in the protein–protein interface can be obtained using the genetic algorithm based approach proposed by Humphris and Kortemme [55]. After every generation of sequence propagation using the genetic algorithm, Monte Carlo-simulated annealing will be used to minimize the energy function over the entire protein complex. During this process, optimal side chain rotameric orientations are chosen for the selected sequence from the Dunbrack rotamer library [56]. During the repacking, only the amino acids that are within 4 Å of any other amino acid are allowed to change their rotamer orientations, while the others are not considered for redesign. The efficiency of binding (the binding score) between the protein and peptide is then estimated based on the inter-chain score across the interface. The difference between the overall complex score and the binding score will be used as an estimate for folding, assuming that the score for which the sequence remains unchanged is constant. With the binding score of the native complex as a benchmark, sequences that offer a binding score within 1% of that of the native complex are considered a part of the tolerance profile. This procedure is repeated for all the near-native structural states generated using the backrub algorithm, and the profiles thus obtained are included in the tolerance profile.

The outcomes of this approach are similar to the Monte Carlo based consensus motif identification algorithm outlined above. The set of sequences generated following this protocol can be used to rationally design the interface between the chaperone and substrate, such that a target peptide can be engineered to be a substrate for a specific chaperone.

### 2.7. Case study: substrate recognition by the yeast chaperone Ydj1p

Ydj1p (Yeast DnaJ 1) is the yeast homolog of *E. coli* DnaJ and a representative member of the 40 kDa heat shock proteins (Hsp40s). These proteins are essential for normal cell growth and the survival of yeast from heat stress, and are involved in protein translocation across the membrane as well as protein folding and degradation [57]. Ydj1p has been shown to influence the assembly-state of endogenous yeast prions, and it influences the aggregation of fragments of huntingtin in yeast models [58,59]. The structure of this protein with a co-crystallized 7-residue peptide substrate (GWLYEIS) in its C-terminal peptide-binding domain (PDB: 1NLT) clearly indicates that the substrate binds to the chaperone via beta-strand extension on the surface of the protein [30]. Our goal in this study was to identify peptides that potentially bind Ydj1 with affinities comparable to the co-crystallized peptide. We applied the fixed backbone redesign protocol described above in order to identify a consensus sequence from the computationally designed peptide library, which had energetically favorable substrates for Ydj1. We systematically mutated each position on the peptide to all possible amino acids, and estimated the change in stability upon mutation. This process is analogous to the generation of a random peptide library, often performed in phage display screening. We then compared the binding affinity of each designed peptide to that of the peptide co-crystallized with Ydj1 (referred to as wild type peptide). We isolated a set of ~2500 peptides featuring a binding affinity of >75% of that of the wild type peptide. We computed the propensity of each amino acid to appear in a given position on the peptide obtained from the computationally generated peptide library, and arrived at a consensus sequence (GX[LMQ]{P}X{P}{CIMPVW}, where [XY] denotes either X or Y and {XY} denotes neither X nor Y) that acts as a sufficient condition for recognition by Ydj1 [31]. We experimentally verified that the computationally generated consensus sequence is in good agreement with the trend observed by phage display screening [31,45]. We therefore demonstrated that our peptide redesign methodology could be generally applied to computationally mimic phage display screening for the identification of potential binding motifs recognized by chaperones.

We conducted a yeast proteome-wide screen for peptides satisfying the identified consensus motif. Interestingly, the hits obtained from our proteome screen include proteins that are either known substrates of Ydj1 (e.g. prions) or other chaperones with which Ydj1 has been shown to interact (e.g. Hsp70, Hsp90), along with other uncharacterized proteins. These results indicate that our methodology can be applied to identify potential new candidate substrates for a given chaperone. We acknowledge that these observations require further experimental verification to validate our claim.

## 3. Conclusions

Chaperones are versatile molecules that recognize unfolded and/or misfolded protein fragments in the cell and target them for refolding or degradation. The role of molecular chaperones in neurodegenerative disorders has been recognized in various animal models. The indispensability of chaperones in conformational disorders has made them potential targets for drug discovery. Recent pharmacological advances in the development of small molecules, like geldanamycin and its derivatives, that induce the synthesis of multiple endogenous molecular chaperones have shown promise with their effectiveness in neurodegenerative disease models. Understanding the basic features of substrate recognition by molecular chaperones will aid in the design of therapeutics that target aggregates enriched by specific peptide fragments for chaperone-assisted degradation. Computational estimation of the binding free energy of peptide–chaperone complexes will be beneficial for the rapid identification of potential substrates for a given chaperone. Many computational algorithms have been developed to accurately estimate the binding affinity of protein–protein interfaces. This report reviews three algorithms that can be directly applied in order to identify potential substrates to a chaperone. In order to accurately represent the chaperone–substrate complex, modeling backbone flexibility is a critical consideration, especially when the chaperone or substrate is likely to undergo conformational change upon recognition. In conclusion, a computational algorithm must be chosen based on the system under consideration in order to efficiently design a set of potential substrates.

## Acknowledgments

We thank Elizabeth Proctor for critical reading of the manuscript. Grant sponsor: National Institutes of Health; Grant Nos.:

R01GM080742, ARRA supplements GM080742-03S1, GM066940-06S1.

## References

[1] R.W. Carrell, D.A. Lomas, Lancet 350 (1997) 134–138.
[2] D.J. Selkoe, Physiol. Rev. 81 (2001) 741–766.
[3] M. Goedert, Nat. Rev. Neurosci. 2 (2001) 492–501.
[4] H.Y. Zoghbi, H.T. Orr, Annu. Rev. Neurosci. 23 (2000) 217–247.
[5] J.I. Clark, P.J. Muchowski, Curr. Opin. Struct. Biol. 10 (2000) 52–59.
[6] M.Y. Sherman, A.L. Goldberg, Neuron 29 (2001) 15–32.
[7] R.I. Morimoto, Science 259 (1993) 1409–1410.
[8] S. Lindquist, E.A. Craig, Annu. Rev. Genet. 22 (1988) 631–677.
[9] U. Jakob, M. Gaestel, K. Engel, J. Buchner, J. Biol. Chem. 268 (1993) 1517–1520.
[10] J.L. Johnson, E.A. Craig, Cell 90 (1997) 201–204.
[11] C. Georgopoulos, W.J. Welch, Annu. Rev. Cell Biol. 9 (1993) 601–634.
[12] D.M. Cyr, W. Neupert, EXS 77 (1996) 25–40.
[13] D.A. Parsell, S. Lindquist, Annu. Rev. Genet. 27 (1993) 437–496.
[14] F.U. Hartl, M. Hayer-Hartl, Science 295 (2002) 1852–1858.
[15] P.K. Auluck, H.Y. Chan, J.Q. Trojanowski, V.M. Lee, N.M. Bonini, Science 295 (2002) 865–868.
[16] P. Fernandez-Funez et al., Nature 408 (2000) 101–106.
[17] P. Kazemi-Esfarjani, S. Benzer, Science 287 (2000) 1837–1840.
[18] J.M. Warrick et al., Nat. Genet. 23 (1999) 425–428.
[19] U.K. Jinwal et al., Mol. Cell Pharmacol. 2 (2010) 43–46.
[20] F.U. Hartl, M. Hayer-Hartl, Nat. Struct. Mol. Biol. 16 (2009) 574–581.
[21] B. Bukau, J. Weissman, A. Horwich, Cell 125 (2006) 443–451.
[22] R.P. Beckmann, L.E. Mizzen, W.J. Welch, Science 248 (1990) 850–854.
[23] D.M. Cyr, T. Langer, M.G. Douglas, Trends Biochem. Sci. 19 (1994) 176–181.
[24] M.A. Scidmore, H.H. Okamura, M.D. Rose, Mol. Biol. Cell 4 (1993) 1145–1159.
[25] A. Szabo et al., Proc. Natl. Acad. Sci. USA 91 (1994) 10345–10349.
[26] T. Laufen et al., Proc. Natl. Acad. Sci. USA 96 (1999) 5452–5457.
[27] F. Hennessy, W.S. Nicoll, R. Zimmermann, M.E. Cheetham, G.L. Blatch, Protein Sci. 14 (2005) 1697–1709.
[28] M.K. Greene, K. Maskos, S.J. Landry, Proc. Natl. Acad. Sci. USA 95 (1998) 6108–6113.
[29] Z. Lu, D.M. Cyr, J. Biol. Chem. 273 (1998) 5970–5978.
[30] J. Li, X. Qian, B. Sha, Structure 11 (2003) 1475–1483.
[31] P. Kota, D.W. Summers, H.Y. Ren, D.M. Cyr, N.V. Dokholyan, Proc. Natl. Acad. Sci. USA 106 (2009) 11073–11078.
[32] Y. Wu, J. Li, Z. Jin, Z. Fu, B. Sha, J. Mol. Biol. 346 (2005) 1005–1011.
[33] C. Sahi, E.A. Craig, Proc. Natl. Acad. Sci. USA 104 (2007) 7163–7168.
[34] S.J. Hubbard, S.F. Campbell, J.M. Thornton, J. Mol. Biol. 220 (1991) 507–530.
[35] F. Ding, N.V. Dokholyan, PLoS Comput. Biol. 2 (2006) e85.
[36] B. Kuhlman et al., Science 302 (2003) 1364–1368.
[37] P. Kollman, Chem. Rev. 93 (1993) 2395–2417.
[38] H. Gouda, I.D. Kuntz, D.A. Case, P.A. Kollman, Biopolymers 68 (2003) 16–34.
[39] P.A. Kollman et al., Acc. Chem. Res. 33 (2000) 889–897.
[40] B.C. Cunningham, J.A. Wells, Science 244 (1989) 1081–1085.
[41] I. Massova, P. Kollman, J. Am. Chem. Soc. 121 (1999) 8133–8143.
[42] R.T. Bradshaw, B.H. Patel, E.W. Tate, R.J. Leatherbarrow, I.R. Gould, Protein Eng. Des. Sel. 24 (2011) 197–207.
[43] I.S. Moreira, P.A. Fernandes, M.J. Ramos, J. Comput. Chem. 28 (2007) 644–654.
[44] G. Winter, A.D. Griffiths, R.E. Hawkins, H.R. Hoogenboom, Annu. Rev. Immunol. 12 (1994) 433–455.
[45] J. Li, B. Sha, Biol. Proc. Online 6 (2004) 204–208.
[46] K. Saraboji, M.M. Gromiha, M.N. Ponnuswamy, Biopolymers 82 (2006) 80–92.
[47] R. Guerois, J.E. Nielsen, L. Serrano, J. Mol. Biol. 320 (2002) 369–387.
[48] D. Gilis, M. Rooman, J. Mol. Biol. 272 (1997) 276–290.
[49] E. Capriotti, P. Fariselli, R. Calabrese, R. Casadio, Bioinformatics 21 (Suppl. 2) (2005) ii54–ii58.
[50] S. Yin, F. Ding, N.V. Dokholyan, Nat. Methods 4 (2007) 466–467.
[51] S. Yin, F. Ding, N.V. Dokholyan, Structure 15 (2007) 1567–1576.
[52] C.A. Smith, T. Kortemme, J. Mol. Biol. 402 (2010) 460–474.
[53] I.W. Davis, W.B. Arendall 3rd, D.C. Richardson, J.S. Richardson, Structure 14 (2006) 265–274.
[54] C.A. Smith, T. Kortemme, J. Mol. Biol. 380 (2008) 742–756.
[55] E.L. Humphris, T. Kortemme, PLoS Comput. Biol. 3 (2007) e164.
[56] R.L. Dunbrack Jr., Curr. Opin. Struct. Biol. 12 (2002) 431–440.
[57] A.J. Caplan, M.G. Douglas, J. Cell Biol. 114 (1991) 609–621.
[58] K.C. Gokhale, G.P. Newnam, M.Y. Sherman, Y.O. Chernoff, J. Biol. Chem. 280 (2005) 22809–22818.
[59] H.Y. Lian et al., J. Biol. Chem. 282 (2007) 11931–11940.