PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

BiPPred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP

Markus Schneider,[1] Mathias Rosam,[2] Manuel Glaser,[1] Atanas Patronov,[1,4] Harpreet Shah,[1] Katrin Christiane Back,[2] Marina Angelika Daake,[2] Johannes Buchner,[2,3] and Iris Antes[1,4]*

[1] Department Biowissenschaftliche Grundlagen, Technische Universität München, Freising, Germany

[2] Department Chemie, Technische Universität München, Garching, Germany

[3] Center for Integrated Protein Science, Department of Chemistry, Technische Universität München, Munich, Germany

[4] Center for Integrated Protein Science, Departments of Bioscience, Technische Universität München, Munich, Germany

ABSTRACT

Substrate binding to Hsp70 chaperones is involved in many biological processes, and the identification of potential substrates is important for a comprehensive understanding of these events. We present a multi-scale pipeline for an accurate, yet efficient prediction of peptides binding to the Hsp70 chaperone BiP by combining sequence-based prediction with molecular docking and MMPBSA calculations. First, we measured the binding of 15mer peptides from known substrate proteins of BiP by peptide array (PA) experiments and performed an accuracy assessment of the PA data by fluorescence anisotropy studies. Several sequence-based prediction models were fitted using this and other peptide binding data. A structure-based position-specific scoring matrix (SB-PSSM) derived solely from structural modeling data forms the core of all models. The matrix elements are based on a combination of binding energy estimations, molecular dynamics simulations, and analysis of the BiP binding site, which led to new insights into the peptide binding specificities of the chaperone. Using this SB-PSSM, peptide binders could be predicted with high selectivity even without training of the model on experimental data. Additional training further increased the prediction accuracies. Subsequent molecular docking (DynaDock) and MMGBSA/MMPBSA-based binding affinity estimations for predicted binders allowed the identification of the correct binding mode of the peptides as well as the calculation of nearly quantitative binding affinities. The general concept behind the developed multi-scale pipeline can readily be applied to other protein-peptide complexes with linearly bound peptides, for which sufficient experimental binding data for the training of classical sequence-based prediction models is not available.

Proteins 2016; 84:1390–1407.

Key words: peptide binding prediction; protein–peptide docking; MMPBSA calculations; MMGBSA calculations; Hsp70-substrate specificity.

INTRODUCTION

The heat shock protein 70 kDa (Hsp70) family is a major chaperone class, which can be found throughout all kingdoms of life.[1] Hsp70s bind to their protein substrates through extended peptide stretches, thus suppressing protein aggregation and assisting folding.[2] Hsp70s consist of an N-terminal nucleotide binding domain (NBD) and a C-terminal substrate binding domain (SBD) (Fig. 1). Structurally, the SBD is composed of a β-sheet "sandwich" harboring a cleft for substrate binding and an α-helical domain, the so-called lid.[2,4] Substrate affinity in the SBD is regulated by ATP hydrolysis in the NBD: The ATP bound state shows low substrate affinity, while the hydrolysis of ATP to ADP leads to efficient substrate binding.[5–7] Upon substrate binding and ATP hydrolysis, the chaperone undergoes large conformational changes.[8] In the ADP state, both domains are weakly coupled and the substrate-binding site is closed

Additional Supporting information may be found in the online version of this article.

Grant sponsor: Deutsche Forschungsgemeinschaft; Grant number: SFB1035; Grant sponsors: CIPSM Cluster of Excellence, TUM-Laura Bassi Award.

Markus Schneider and Mathias Rosam contributed equally to this work.

*Correspondence to: Iris Antes, Department Biowissenschaftliche Grundlagen, Technische Universität München, Freising, Germany. E-mail: antes@tum.de

Received 10 February 2016; Revised 8 May 2016; Accepted 19 May 2016
Published online 10 June 2016 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.25084

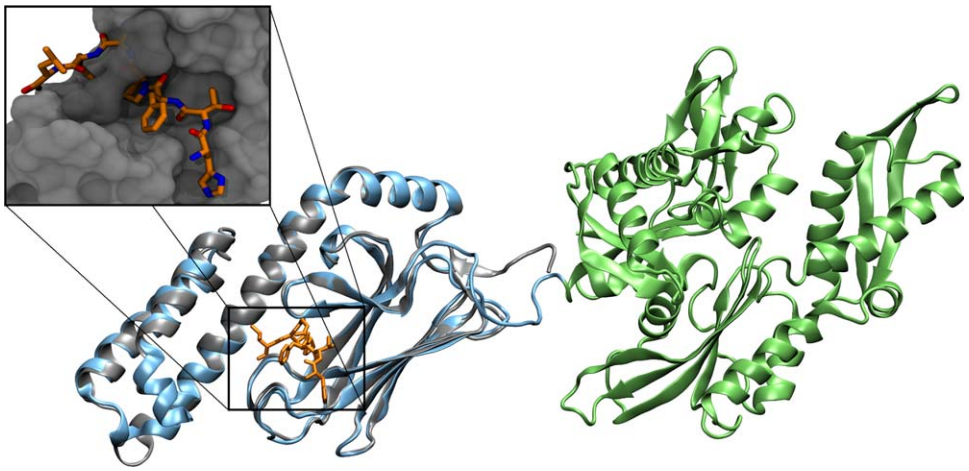


Figure 1

Structure of DnaK in its ADP-bound state (NMR, PDB-ID: 2KHO[3]) (green, nucleotide binding domain (NBD); iceblue, substrate binding domain (SBD)) with the superimposed homology model of the BiP-SBD (gray cartoon) with bound substrate peptide (HTFPAVL, orange licorice). The inset shows BiP's substrate binding cavity in surface representation with bound HTFPAVL.

(Fig. 1). Upon ATP binding, the domain interaction increases, leading to a respositioning of the lid, which in return allows the opening of the substrate binding site and substrate release.

The substrate specificity of Hsp70 chaperones has been studied extensively and it was observed that Hsp70s originating from different organisms and compartments differ in their substrate recognition.[9–11] It is commonly accepted that extended stretches of five to seven, preferably hydrophobic, residues are recognized.[1,2,12,13] The conserved binding site consists of an elongated groove with a deep, mainly hydrophobic binding pocket at the center (Fig. 1, inset). Structural and electrostatic complementarity of the central peptide residue with this central pocket is crucial for stable peptide binding.[14] Once stable side-chain interactions with this pocket are established, a hydrophobic bridge consisting of two binding site loops closes over the peptide backbone thus further stabilizing the peptide in its bound position.[14–16] Differences in the binding site in the various Hsp70s lead to variations in their substrate binding affinity and peptide kinetics and specificity.[14,17] For example, BiP, the endoplasmic Hsp70, accepts tryptophan residues in the binding site, whereas DnaK, the Hsp70 homolog in *Escherichia coli*, does not.[13,17,18] Both share the preference for leucine-enriched peptides.

In the present study, we focus on the peptide binding properties of BiP.[1,19] BiP is involved in many biological processes such as protein folding, quality control, and translocation into the ER.[20–23]

To identify binding sites of Hsp70 chaperones in proteins, computational prediction algorithms are needed as they allow a high-throughput, proteome-wide approach. However, despite long lasting efforts, the prediction of potential Hsp70 binding sites in proteins still remains challenging.[17,18,24,25] Several prediction models based on the sequences of known peptide substrates and nonbinders were developed for the identification of peptide stretches binding to Hsp70 binding sites. On the basis of the data from a bacteriophage-based study of the binding of random octa- to decapeptides to BiP, Blond-Elguindi *et al.* constructed a sequence-based, position-specific scoring matrix (PSSM) using the experimentally observed amino acid probabilities at each specific sequence position of the bound peptide as matrix terms. Evaluation of the prediction method on independent data sets[17,26] showed that the prediction accuracy varies between 60 and 80% depending on the tested protein substrate. The score was further improved by the inclusion of sequence information from known substrate proteins.[17] Rüdiger *et al.*[13,18] performed a systematic study of the substrate specificity of DnaK, the bacterial Hsp70, using peptide array data of 4260 tridecapeptides. This prediction approach is also based on the amino acid probability at each specific sequence position of the bound peptide. The method clusters them into a core and two flanking regions, and assigns "region-specific" scores for each amino acid. The algorithm can correctly predict about 80% of the strong binders and nonbinders in the benchmark set.

More recently, van Durme *et al.*[25] performed another study for DnaK, testing 172 decapeptides from seven known DnaK substrate proteins using the same experimental setup. The experimental data was used to first create a classical sequence-based PSSM based on the amino acid probability at each specific sequence position. Afterward, a combined sequence- and structure-based PSSM was established by performing an *in silico* structural position scan. For this the FoldX force field was

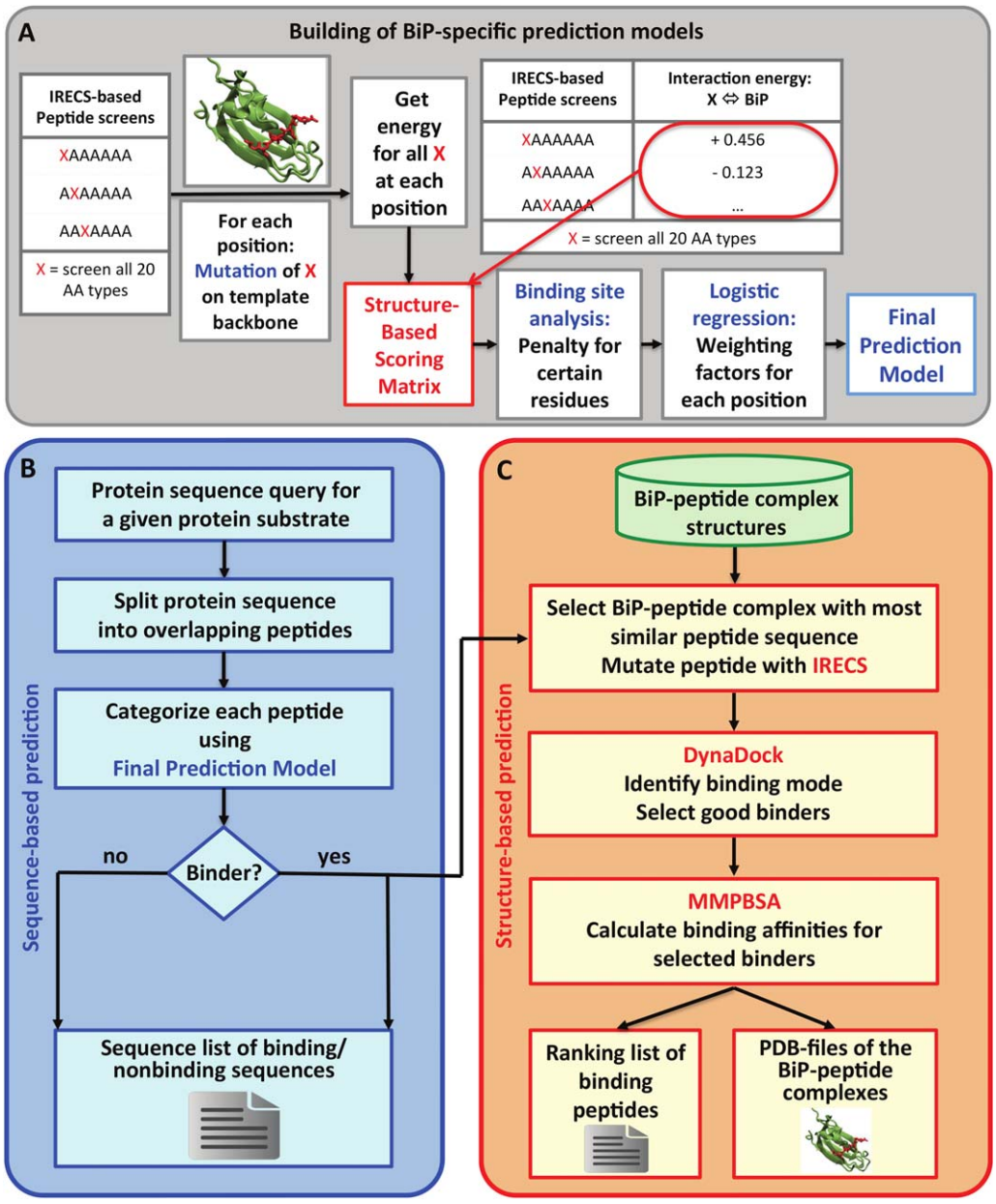


Figure 2

Pipeline for the development of the prediction models (**A**). Flowchart of the sequence-based (**B**) and of the structure-based (**C**) prediction steps of the final prediction pipeline.

used to thread each amino acid onto each position of the peptide using a polyalanine heptapeptide backbone, and to calculate its interaction energy with the DnaK binding site.[25] These interaction energies were added to the PSSM. The additional use of structural data improved the prediction accuracy considerably for the validation set, indicating that the use of structure-based information leads to more robust predictions than models based on peptide sequence data only. However, the differences in the performances of the individual models were strikingly larger for the validation set than for the benchmark set, which might be due to the small size of the validation set.

Regarding the overall accuracies of the prediction methods for Hsp70-peptide interactions, these are not as high as for other protein–peptide systems like for example major histocompatibility complex (MHC) peptide complexes. Extensive studies in the context of MHC-peptide binding led to three crucial prerequisites for the development of accurate sequence-based prediction models of protein-peptide binding.[27] First, the amount and comparability of the available experimental data is

Table I
Pentadecapeptides Tested for BiP Binding in Solution

ID	Sequence[a] (N- to C-terminus)	Parent protein	Solubility[b]	DE[c] (%)	Pred. binders[d]	$Score_{max}$[e]
0	HTFPAVL	C_H1	+	101.9	1	1.00
75	SSLGTQTYICNVNHK	C_H1	– – –	–	-	-
81	TYICNVNHKPSNTKV	C_H1	– –	–	-	-
83	ICNVNHKPSNTKVDK	C_H1	+	35.1	7	0.84
213	QHNKCECRPKKDRAR	VEGF	–	–	-	-
225	RARQENPCGPCSERR	VEGF	+	11.2	3	0.69
238	RRKHLFVQDPQTCKC	VEGF	+	44.0	9	0.97
251	KCSCKNTDSRCKARQ	VEGF	+	n.d.	1	0.80
260	RCKARQLELNERTCR	VEGF	+	38.4	8	0.95

The displacement efficiency for Lucifer Yellow-labeled HTFPAVL (HTFPAVLGSC) was determined in fluorescence anisotropy measurements and represents the ratio of the change in anisotropy (Δr) for dissociation and Δr for association. Dashes indicate that the parameter could not be determined due to artifacts during the measurement caused by peptides with poor solubility.
[a]To increase peptide stability, the N- and C-termini were synthesized in acetylated and amidated form, respectively.
[b]Subjective assessment of solubility upon addition of HKM buffer, that is, observation of visible aggregates in solution.
[c]DE: Displacement efficiency; n.d. not detectable.
[d]Number of heptamers within the 15mer peptide, predicted as binders with the CD-fitted IE/BA model corresponding to "Model 6" in Table II.
[e]$Score_{max}$: highest score obtained within the group of predicted binders from column 5.

crucial. For a decent prediction accuracy, sequence-based prediction models should be based on at least 200 binding and 200 nonbinding sequences. Second, the data sets need to be well-balanced (equal amount of binding and nonbinding sequences). Third, a drop of 10–20% in accuracy can be observed if the exact binding register is not known and has to be predicted computationally.

In the case of Hsp70-peptide prediction models, the size of the experimental data sets used so far is either below or at the lower limit of 200 peptides in most studies. Furthermore, as there exist only a few binding sites in each protein substrate, a large imbalance between the number of binding and nonbinding sequences can lead to a bias in the prediction model. Third, in all studies, the length of the measured peptides ranges from 8 to 13 residues, thus the exact binding heptamer stretches (i.e., the exact binding registers) are experimentally not known. Therefore, the actual binding heptamer stretches were selected using various computational procedures, which might be error-prone. Next to these general accuracy-limiting factors, it was observed recently that peptides can bind to DnaK with their backbone placed alternatively in a so called "forward" or "reverse" direction[15] in the symmetrical binding site (Supporting Information Fig. S1); that is, the peptide is flipped by 180° with respect to its backbone direction. As the binding direction of the peptides in the binding assays is not known, the same (forward) direction (same as observed in the first experimental structures) is assumed for all peptides during the development of sequence-based prediction models. Therefore, the potential existence of a "reverse" binding mode introduces an additional error into the sequence-based models. Nevertheless, regarding the experimental structures available in the PDB at this time, 15 different peptides are bound in forward direction, while only 7 peptides were observed to bind in reverse orientation. As, in addition, the latter were all obtained by the same group and are thus structurally similar,[15] the introduced error should be tolerable.

Because of these challenges for the prediction of Hsp70-peptide binding, the major goal of this work was to design a prediction approach, which is based predominantly on structural "*ab initio*" modeling, using this data for the design of a structure-based position-specific scoring matrix (SB-PSSM). This allows predicting peptide binding to BiP reliably with high accuracy without the need of extensive high quality experimental binding data and knowledge of the exact binding mode of each peptide binder. We present a hierarchical approach which combines such a SB-PSSM-based prediction model with subsequent molecular docking and MMPBSA calculations allowing for identification, structural characterization, and binding affinity estimation of peptide sequences binding to BiP (Fig. 2). In addition, peptide array (PA) data and florescence anisotropy measurements were performed for the optimization and verification of the prediction results.

MATERIALS AND METHODS

Peptide array design and experiments

CelluSpot™ peptide array chips were purchased from Intavis (Cologne, Germany) and comprised 384 peptides spotted in duplicates onto the chip. The array was designed to contain peptides with a length of 15 amino acids overlapping by 14 residues, that is, 1 amino acid offset from spot to spot. The specific sequences of the parent proteins are summarized in Supporting Information Table SV. The peptide array was prepared according to the manufacturer's instructions. 14 μ*M* BiP was incubated with the peptides on the chip surface in HKM buffer (50 mM HEPES/KOH pH 7.5, 150 m*M* KCl, 10 m*M* $MgCl_2$) containing 1 m*M* ADP at 37°C for 2 h.

Table II
Performance of the Different Prediction Models

Model	SB-PSSM	Data sets[a]		Performance	
		Training set	Evaluation set	AUC_{train}[b]	AUC_{eval}[b]
1	IE[c]	—	CD	—	0.48
2	IE/4	—	CD	—	0.61
3	IE/BA	—	CD	—	0.83
4	IE	CD	—	0.71	—
5	IE/4	CD	—	0.72	—
6	IE/BA	CD	—	0.85	—
7	IE	PA_{train}	PA_{eval}	0.68	0.74
7	IE	PA_{train}	CD	0.68	0.59
8	IE/4	PA_{train}	PA_{eval}	0.70	0.65
8	IE/4	PA_{train}	CD	0.70	0.51
9	IE/BA	PA_{train}	PA_{eval}	0.65	0.57
9	IE/BA	PA_{train}	CD	0.65	0.58

[a]For the definition of the data sets see Material and Methods section.
[b]Area Under the Curve (AUC) values of the corresponding training sets (AUC_{train}) and the independent evaluation sets (AUC_{eval}).
[c]For the definition of the SB-PSSM models see Results section.

Bound BiP was detected by a primary anti-BiP antibody (1:5000) kindly provided by Linda Henderhot (St. Jude Children's Research Hospital, Memphis, TN) and a secondary anti-rabbit IgG antibody coupled to horseradish peroxidase (1:10,000; Sigma–Aldrich, St. Louis, MO).

Peptide preparation and fluorescence anisotropy spectroscopy measurements

Synthesized peptides

All peptides from Tables I–III were ordered from Biomatik (Cambridge, Canada) at a purity grade of 95% or higher. As published, HTFPAVL and SVFPLAP were synthesized without modification at their termini.[8,14] Because the termini of the peptides on the array were not charged, the individually ordered peptides were acetylated at the N terminus and amidated at the C terminus to increase stability.

Peptide labeling and preparation of peptide stocks

HTFPAVLGSC was labeled with Lucifer Yellow as described before[14] and the synthesized peptides were dissolved in HKM buffer to a final concentration of 10 mM. Because of the lack of aromatic side chains, no extinction coefficients could be determined for the peptides. Therefore, the exact synthesized quantity (masses around 5 mg per peptide) and the molecular mass described in the quality control report of the manufacturer were used to calculate the amount of buffer needed to achieve a final concentration of 10 m*M* in the stock solution.

Table III
Predicted Heptapeptides Tested for BiP Binding in Solution

ID	Sequence[a] (N- to C-terminus)	Parent protein	Solubility[b]	M3[c]	DE[d] (%)
0	HTFPAVL	C_H1	+	B	113.1
HP1	PGHPPRF	VpreB	+	NB	n.d.
HP2	GPCSERR	VEGF	+	NB	–
HP3	PQVPPRF	VpreB	+	*B*	n.d.
HP**4**	KDVARNR	VpreB	+	**B**	**39.4**
HP5	QPEDEAM	VpreB	−	NB	n.d.
HP6	MGARSSE	VpreB	+	NB	n.d.
HP7	HPIETLV	VEGF	− −	B	48.6
HP**8**	PMAEGGG	VEGF	+	**B**	**75.9**
HP9	FMDVYQR	VEGF	− − −	B	49.3
HP**10**	PPRFLLR	VpreB	+	**B**	**107.0**

The displacement efficiency for Lucifer Yellow-labeled HTFPAVL (HTFPAVLGSC) was determined in fluorescence anisotropy measurements and represents the ratio of the change in anisotropy (Δr) for dissociation and Δr for association. Dashes indicate that the parameter could not be determined due to artifacts during the measurement caused by peptides with poor solubility.
[a]To increase peptide stability, the N- and C-termini were synthesized in acetylated and amidated form, respectively.
[b]Subjective assessment of solubility upon addition of HKM buffer, that is, observation of visible aggregates in solution.
[c]M3: corresponds to "Model 3" in Table II; wrong predictions are highlighted as *italic* typeface. Peptides highlighted in **bold** typeface were found to bind to BiP and were further used for titration experiments. B, binder; NB, nonbinder.
[d]DE: Displacement efficiency; n.d., not detectable.

Fluorescence anisotropy spectroscopy

For the detection of peptide association and dissociation kinetics to BiP, a Jasco FP-8500 spectrofluorimeter equipped with polarizers was thermostated at 37°C. Samples containing 1 μ*M* Lucifer Yellow (LY)-labeled HTFPAVLGSC and 1 m*M* ADP were equilibrated and measured in a 1-cm quartz cuvette at 37°C for ~15 min before 15 μM BiP was added. LY was excited at 428 nm and BiP-peptide association was followed at 525 nm with bandwidths of 5 and 10 nm for excitation and emission, respectively. Sensitivity was set to high and the time interval was 1 s. After reaching steady-state, a 150-fold molar excess of unlabeled peptide was added to the cuvette and dissociation was recorded. The kinetic parameters were derived from a single-exponential model similar to Ref. 14. Titration experiments were performed with increasing concentrations of the competing unlabeled peptide and fitted using Eq. (1)

$$y = F_p - \left((P_t + x + K_d) - \sqrt{(P_t + x + K_d)^2 - (4P_t x)}\right) \times \left(\frac{F_p - F_{pl}}{2P_t}\right) \quad (1)$$

with the fluorescence signal of the peptide F_p, the total concentration of the labeled peptide P_t, the total concentration of the ligand (BiP) x, the affinity between peptide and ligand K_d, and the fluorescence signal of the peptide-ligand complex F_{pl}.

BiP homology model

For the construction of the SB-PSSMs a previously described homology model of the BiP substrate binding domain (SBD) with a bound peptide (HTFPAVL)[14] was used (Fig. 1). It was created using the DnaK structure 1DKX.[4] The alignment of the protein sequences was performed with the align2d module of MODELLER.[28] Afterwards 400 structural models were created using MODELLER and the model with the best DOPE score was used for the further studies after manual inspection for soundness. The final model was energy minimized and slowly heated up to 300 K and equilibrated. A more detailed description of the modeling procedure is provided in Ref. 14

Sequence-based prediction model

For the sequence-based prediction, a model was developed which uses an interaction energy-based structure-based position-specific scoring matrix (SB-PSSM), and is independent from the available experimental binding data [Fig. 2(A)]. Experimental data-based statistical learning was only used to adjust additional position-specific weights, which scale the contribution of each position of the bound peptide in the final score. All structural calculations use a previously built homology model of BiP.[14]

Structure-based position-specific scoring matrix

For the elements of this matrix, first, a peptide library was created by introducing point mutations at each position of the bound peptide in the BiP homology model, while restraining its backbone conformation [Fig. 2(A)]. For this SB-PSSM the base peptide AAAPAAA was used instead of HTFPAVL to avoid any bias from the neighboring residues. A proline residue was placed in the central binding pocket instead of an alanine, as it was observed that this increases the overall stability of the peptide's position in the binding site. The mutations were performed with our in-house tool for protein side-chain prediction, IRECS,[29,30] by mutating all 20 proteinogenic amino acids onto all seven peptide sequence positions. This way a structural library of 7 × 20 BiP-peptide complexes was assembled and used for the derivation of the SB-PSSM. Afterwards, all 140 complexes were energy minimized using DynaCell[31] with the OPLS all-atom parameter set.[32] For the energy optimization, a step size of 0.002 nm and an energy convergence criterion of 1 kJ mol^{-1} was used. Afterwards the pepscore-weighed Coulomb and Lennard–Jones interaction terms between the mutated residues and BiP were calculated according to Ref. 31 (for details about the pepscore scoring function see section "molecular docking simulations") for all energy minimized structures. These energy values were normalized over the whole matrix and the resulting normalized values form the basis for the SB-PSSM.

To further increase the prediction performance of the SB-PSSM, the binding site of the equilibrated BiP-HTFPAVL complex was comprehensively analyzed and for all peptides, which were mutated at position 4, molecular dynamics (MD) simulations were performed. The reason for the latter was that the central binding site was previously found to be very flexible, requiring a dynamic treatment to be able to accurately judge the binding properties of the different amino acids in that pocket.[14] For the molecular dynamics simulation, the same conditions as described in the MMPBSA section were used, but with the OPLS all-atom force field[32] instead of ff99SBildn, to be consistent with the IRECS mutation calculations. The SB-PSSM was further modified according to the results of these two analyses. A more detailed discussion of the modifications can be found in the Results section and a detailed description of the analyses is given in the Supporting Information.

In addition, in the final matrix, the values of the first and seventh position were set to zero as the variation in these amino acids was very small in the available binding data set (CD data set, see below). The reason for this simply is that they are considered unimportant for binding

Table IV
Experimental (K_d) and Computational Binding Data from DynaDock Docking, MMPBSA, and MMGBSA Calculations

ID	Sequence[a] (N- to C-term)	K_d (μM)	FF-score[b] (kJ mol^{-1})	Pepscore[c] (kJ mol^{-1})	ΔG MMPBSA[d] (kcal mol^{-1})	ΔG MMGBSA[d] (kcal mol^{-1})
HP4	KDVARNR	12.0 ± 3.1	−2806.33	−375.30	−9.78 (0.50)	−63.44 (0.38)
	KDVARNR_r	–	−3113.55	−402.35	−15.46 (0.61)	−82.49 (0.43)
HP8	PMAEGGG	17.9 ± 6.0	−2313.11	−127.53	5.15 (0.39)	−39.09 (0.27)
	PMAEGGG_r	–	−2239.15	−139.90	14.29 (0.50)	−34.92 (0.29)
HP10	PPRFLLR	9.7 ± 4.1	−2290.83	−439.79	−23.23 (0.40)	−72.87 (0.31)
	PPRFLLR_r	–	−2051.27	−436.52	−5.71 (0.64)	−58.80 (0.37)

All complexes were docked in forward and reverse binding mode ("_r") and the two binding modes were analyzed separately.
[a]To increase peptide stability, the N- and C-termini were synthesized in acetylated and amidated form, respectively.
[b]Docking pose with the best FF-score among the cluster representatives of the five biggest structural clusters.
[c]Docking pose with the best pepscore among the cluster representatives of the five biggest structural clusters (chosen for MMPBSA/MMGBSA analysis).
[d]The standard error of the mean is provided in parentheses.

and thus are (in contrary to the central three residues) not mutated in the corresponding experimental studies. This leads to an artificially high "sequence conservation," which biases the statistical learning step leading to unnaturally high weights for these positions and thus no gain in the prediction performance of the final model could be observed by the inclusion of these positions.

Data sets for statistical learning

Experimental peptide binding data was collected from different resources: First, we included the data from the peptide array (PA) assay. As we assumed that each binding core should be defined as a heptamer, there are multiple possibilities for binding stretches in the tested 15mers and it is unknown which one of the possible heptapeptides in the corresponding 15mer peptide is responsible for the signal. Additionally, the signal of known heptameric binders can fluctuate depending on their position in the 15mer sequence. Therefore, we identified all 15mers in which the same heptamer is present and used the average over the PA signal intensities of these 15mers as a score to represent the heptamer's binding affinity. To create a robust data set, only the sequences with an average intensity >20% (binders) and those with an intensity <2% (nonbinders) were included in the PA final data set. Second, binding peptide sequences were taken from Knarr *et al.*[17,26] In this case we included all peptide sequences as binders in our training set, for which an ATPase stimulation factor of >1.5 was detected. Third, we included DnaK binding sequences from Zahn *et al.*[15] Data from both chaperones can be used in the statistical learning step as this step is only used to adjust the weights of the individual sequence positions. These weights are predominantly defined by the overall geometry of the binding site, which is the same in both chaperones. Fourth, we included several peptide sequences, which were identified in previous studies in our groups.[14]

On the basis of this data we defined three final data sets for the training of the SB-PSSM. The full PA data set was divided into two data sets, a training set (PA_{train}) and an independent evaluation set (PA_{eval}, 20% of the entire PA data set). The PA_{train} data set consists of 117 peptides (41 binding, 76 nonbinding sequences), while 30 peptides (19 binding, 11 nonbinding sequences) are included in the PA_{eval} set, respectively. The third data set (CD, Collected Data) consists of the binding peptides from resources 2–4[14,17,26] to which we added the definite nonbinding sequences from all sources (Supporting Information Table SIV).

Statistical learning protocol

Training of the models was conducted using logistic regression realized by a C++ implementation of LIBLINEAR version 1.96 or Python 2.7 with the additional libraries skikit-learn 0.16.1, SciPy 0.16 and NumPy 1.9.2. L2-regularized logistic regression was used and the regularization strength parameter was optimized by threefold cross-validation for the best area under the curve (AUC) score in a response operator characteristic (ROC) curve. The samples were weighted inversely to the frequency of their respective class in the data set. The performance of the different prediction models was evaluated using ROC analysis. In the shown ROC plots the True Positive Rate (TPR) is plotted against the false positive rate (FPR) for different minimum scores required for a data point to be predicted as a binder. Good models have a fast-climbing TPR and a slow-climbing corresponding FPR, which leads to a high AUC in the plot. Random prediction algorithms are expected to yield an AUC of 0.5 (diagonal from (0,0) to (1,1)), while a perfect algorithm would have an AUC of 1.0 (step from (0,0) to (1,0)).

Molecular docking simulations

As there exist no experimental structures of BiP-peptide complexes, the performance evaluation of the DynaDock method for Hsp70-peptide complexes was performed using seven DnaK-peptide complexes with known X-ray structures (PDB-ID: 1DKX,[4] 3DPO,[33]

3QNJ, 4E81, 4EZT, 4EZY, and 4EZZ[15]). For the three predicted binding peptides of BiP (Tables III and IV), the equilibrated homology model of the BiP-SBD with the HTFPAVL peptide (see above) was used and the amino acid side chains of the predicted peptides were mutated using the HTFPAVL peptide backbone with the side-chain placement tool IRECS.[29,30] In accordance to the experiment, the termini of the BiP peptide ligands were capped with N-terminal acetyl or C-terminal amide groups, respectively. To model the reverse binding mode of peptide ligands, the template peptide ligand was flipped along its backbone direction by 180°. The molecular docking simulations were performed using the DynaDock module of our in-house modeling program DynaCell.[31] A detailed description of the docking conditions is provided in the Supporting Information Text S1.

MMPBSA and MMGBSA calculations

For the final, energetically best scoring docking poses from the DynaDock calculations 20 ns of molecular dynamics simulations were performed using the AmberTools14 package[34] and standard, established simulation conditions, as detailed in the Supporting Information Text S1. To estimate the free energy of binding of the bound peptides Molecular Mechanics-Poisson Boltzmann Surface Area (MMPBSA)[35–37] and Molecular Mechanics-Generalized Born Surface Area (MMGBSA)[36] calculations were performed, based on snapshot ensembles collected in 20 ps intervals from the last 5 ns of the 20 ns MD trajectories. MMPBSA and MMGBSA calculations were conducted with the MMPBSA.py script[38] from the AmberTools14 software package.[34] Further details (GB method, SA models, etc) are provided in the Supporting Information Text S1.

RESULTS

In the following we present a new hierarchical approach for the prediction of peptides binding to the Hsp70 chaperone BiP. First, we will present the experimental data determined in this study, followed by the optimization and performance evaluation of the sequence-based prediction methods. Finally, we will discuss the results of the structural docking and binding free energy estimations. In Figure 2 a schematic overview over the approach is provided.

Experimental peptide array and anisotropy studies

To obtain sufficient data for the parameterization of a sequence-based prediction model, a peptide array containing 384 overlapping peptides was designed from three secreted human proteins, the IgG1 antibody C_H1 domain, the vascular endothelial growth factor A (VEGF), and the surrogate light chain component (VpreB) (Fig. 3 and Supporting Information Table SV). As a control, the formerly identified binders SVFPLAP and HTFPAVL from C_H1 were included[14] together with the previously predicted nonbinder LRAEDMA (lowest BiP score) and the binder FTFSDYY (highest BiP score) from V_H.[26] All peptides had a length of 15 amino acids with an offset of one residue and were linked covalently with their C-terminus to cellulose spots on a glass chip.

Peptide spots associated with BiP were identified by immunodetection (Fig. 3). Both internal array replicates indicate a homogenous incubation with BiP and the absence of spots on the negative control shows the specificity of the detection antibodies. The array results were reproducible qualitatively but some minor differences existed in the quantitative evaluation, albeit the overall trend was not changed. The strongest BiP-peptide interactions were detected in the lower half of both arrays, indicating a stronger interaction with VEGF and VpreB than with C_H1. Quantification of the spot intensities allowed ranking the peptides and grouping them into a selection of protein- or array-wide top 10 binders (Supporting Information Table SI). In the array-wide ranking, the top 10 positions are predominantly occupied by VpreB peptides (seven peptides) showing the consensus sequence PRFSGSKDVARNR, followed by three peptides of VEGF occupying the positions 6, 8, and 10 in the array-wide ranking, respectively. The earliest entry for C_H1 appears at rank 59 (TYICNVNHKPSNTKV, peptide 81), featuring an array-wide relative signal intensity of only ~17%. This categorizes C_H1 in general as the weakest binding partner for BiP compared to VpreB and VEGF.

Surprisingly, from the known C_H1 binding sites SVFPLAP and HTFPAVL, only SVFPLAP containing 15mers were found in the C_H1-specific top 10 (3 appearances), but no HTFPAVL containing peptides (Supporting Information Table SI). Regarding the C_H1-wide comparison of the known binders, SVFPLAP-containing peptides yielded a maximum signal intensity of ~80% (peptide 4), while HTFPAVL peptides only reached a maximum intensity of ~40% (peptide 52; not present in top 10) (Supporting Information Table SI).

In summary, the peptide array data indicate that BiP can recognize peptides immobilized in a cellulose matrix with differing efficiency. Based on the array-wide ranking VpreB was identified as the most potent BiP binder, followed by VEGF and C_H1. However, the known binders SVFPLAP and HTFPAVL, which also served as positive control and share similar affinities for BiP in solution (12.5 and 11.1 μM, respectively[14]), do not show comparable signal intensities in the peptide array. This observation indicates that the affinities in solution do not necessarily correlate with the signal intensity on the chip.

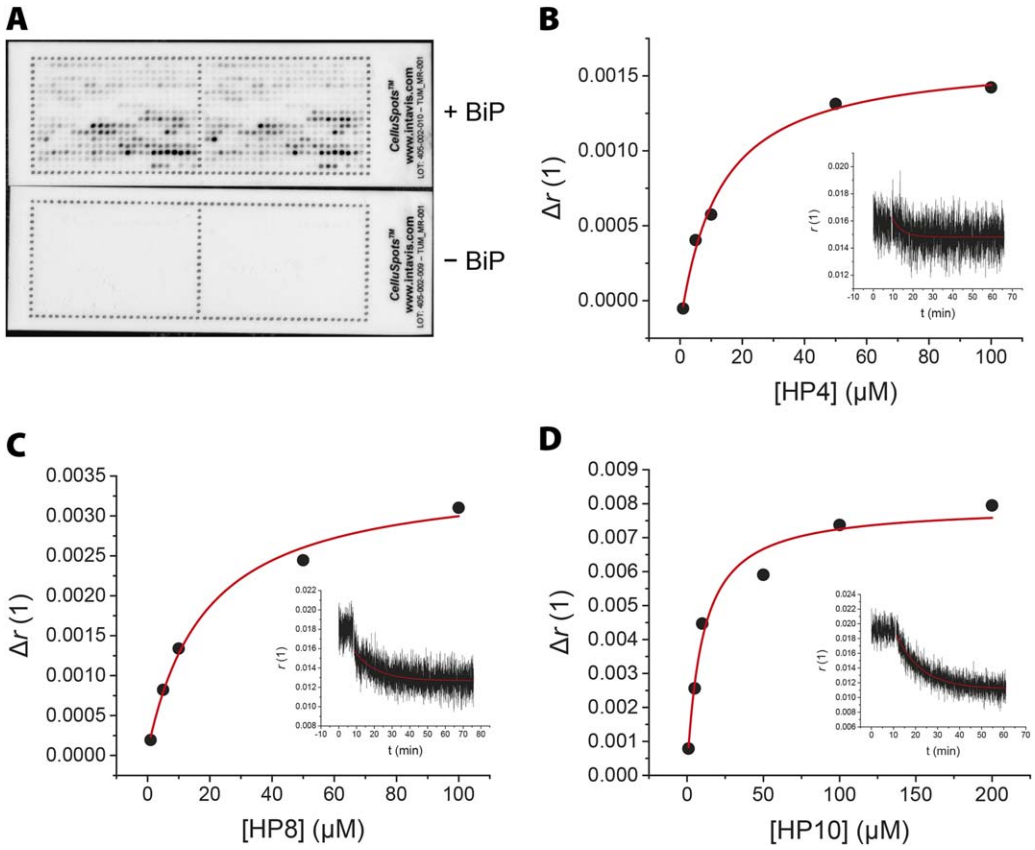


Figure 3

Identification of novel BiP binding sites in C_H1, VEGF, and VpreB and determination of binding affinities for selected true binders. (**A**) BiP was incubated with the peptide chip at 14 μM in HKM buffer containing 1 mM ADP for 2 h at 37°C. Bound BiP was detected with α-BiP (1:5,000) as primary and α-rabbit (1:10,000) as secondary antibody, respectively. Exposure time was 1 min and post-processing (Auto Contrast, overlay of luminescence and visible light images) was performed in Photoshop on the whole images. Each chip contained two identical arrays and the shown images are representative of at least three independent experiments. The addition of BiP is indicated on the right. (**B–D**) The difference in fluorescence anisotropy signal (Δr) between 1 μ*M* of fully bound and fully displaced HTFPAVLGSC by/from BiP was determined at different concentrations of the peptides HP4 (**B**), HP8 (**C**), and HP10 (**D**). Δr was then plotted against the competitor concentration and fitted according to the formula in the material and methods section (red) to derive the K_d values. The inset depicts an exemplary trace of HTFPAVLGSC displacement at 100 μ*M* competitor with a single-exponential fit. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Because of the different BiP association behavior on the chip and in solution, a selection of the array-derived strong and medium binders was synthesized and their binding to BiP was tested *in vitro* by fluorescence anisotropy spectroscopy (Table I and Supporting Information Table SII). For these experiments HTFPAVL coupled to a C-terminal GSC linker (HTFPAVLGSC) was labeled with Lucifer Yellow (LY) and used to monitor the association with BiP. Once in steady-state, the BiP-HTPFPAVLGSC complex was dissociated with an excess of unlabeled HTFPAVL, thus proving that the peptide competes with its labeled counterpart for binding to BiP.

As visible from the changes in the anisotropy signal, all peptides with poor solubility (peptides 75, 81, and 213) produced artifacts during the measurements leading to a sudden signal increase upon their addition to the BiP-HTFPAVLGSC complex (Supporting Information Fig. S2, sharp signal spike between 30 and 40 min). These peptides could not be used for further experiments due to their undefined behavior and the light scattering properties of the formed aggregates.

The peptides 83, 225, 238, and 260 showed weak displacement of the labeled peptide as the anisotropy signal decreased only slightly over time. Even after 80 min, the strongest competitor, peptide 238, could displace only ∼44% of the initially bound HTFPAVLGSC whereas unlabeled HTFPAVL displaced ∼100% (Table I, Supporting Information Fig. S2). With 0.134 min^{-1} the rate of displacement was higher than for HTFPAVL (Supporting Information Table SII). In a similar range, peptides 260 and 83 displaced ∼38 and ∼35% of the BiP-HTFPAVLGSC complex, respectively, but at different rates of 0.107 and 0.280 min^{-1} (Table I, Supporting Information Fig. S2, Table SII). Peptide 225 only

displayed a limited capability of binding to BiP with a displacement efficiency of 11%. These relatively low efficiencies and hence BiP binding capabilities were surprising, as the respective peptides showed strong signals on the peptide chip. This discrepancy indicates that immobilized peptides behave differently once they can move freely in solution and underlines the importance of a verification of binders in solution.

In summary, the fluorescence anisotropy measurements revealed that the binding intensity on the chip does not necessarily correlate with the peptide's ability to displace bound peptides from BiP in solution. For example, the best binder in solution was peptide 238, which did not even appear in the ranking of the top 10 binders on the chip.

Design of the sequence-based prediction model

The study by van Durme *et al.* and several MHC-peptide binding studies showed that using binding energy data from structure-based calculations for the construction of the PSSM generally improves the performance of the prediction models.[25,27,39] However, all these approaches still rely strongly on the available experimental data, as either amino acid propensity-based data is additionally included in the PSSM or complex experimental data-dependent prediction models are trained on the basis of the calculated interaction energies using statistical learning procedures. Thus in both approaches the model's performance is still strongly correlated to the size and quality of the chosen experimental data, that is, the type of data used, the consistency of their measurement, and the overall experimental setup. As already discussed in the Introduction, in the case of Hsp70-peptide binding there exist serious limitations for such approaches due to the limited available data as well as the specific binding properties of the peptides (e.g., forward/reverse binding direction). Thus the aim of this study was to develop a computational procedure, which, in contrast to all previous studies, allows for the derivation of a highly accurate sequence-based prediction model without the use of experimental data. Therefore the derivation strategy presented here is based on the idea of developing an "*ab initio*" (i.e., without the use of experimental data) position-specific scoring matrix, solely obtained by structure-based modeling (SB-PSSM).

Specifically, all previous prediction approaches for Hsp70-peptide binding use either experimental data-based amino acid propensity scores alone or a combination of these with interaction energies from structural calculations as matrix elements in the PSSM. In contrary, we combine force field based interaction energies with results from extensive binding site analyses and molecular dynamics simulations for these terms. This leads to a PSSM, which is *a priori* independent of any experimental binding data (amino acid propensities). Nevertheless, as existing high-quality experimental data can also be used to further improve the performance of such a prediction model, we evaluated this possibility and additionally assigned one overall weighting factor to each sequence position of the peptide, that is, resulting in seven position-dependent weighting parameters (instead of 140 (7×20) amino acid and position-dependent propensity-based scores in the previous models). These weights were fitted on the basis of the available experimental data.

Optimization and evaluation of the SB-PSSM-based prediction model

For the evaluation and optimization of the prediction model we focused on two aspects, namely the influence of the binding site analysis and molecular dynamics simulation-based modifications of the PSSM on the performance of the prediction model, and the influence of additional fitting of the position-based weights using experimental data.

Thus we built three different SB-PSSM matrices: The first matrix only contained the force field-based interaction energies (further referred to as IE, interaction energy), the second also contained MD-based modifications of the scores at peptide position 4 (binding into the central binding pocket) (IE/4, interaction energy/modified position 4), and the third matrix contained also all modifications at the other peptide positions based on the binding site analyses (IE/BA, interaction energy/binding site analyses).

For these three matrices we investigated different settings for the training of the position-dependent weights and the model evaluation using three different experimental data sets: PA_{train}, PA_{eval}, and CD. For the first two data sets, the peptide array data was randomly divided into a training (PA_{train}) and an evaluation (PA_{eval}) set, representing 80 and 20% of the data, respectively. In addition, an independent third data set (CD, collected data) was built by a collection of experimentally verified binding sequences from literature and additional nonbinding sequences from the peptide array, as nearly no verified nonbinding data was available elsewhere. The CD set was collected in response to the discrepancies between experimental anisotropy and peptide array results observed in this study.

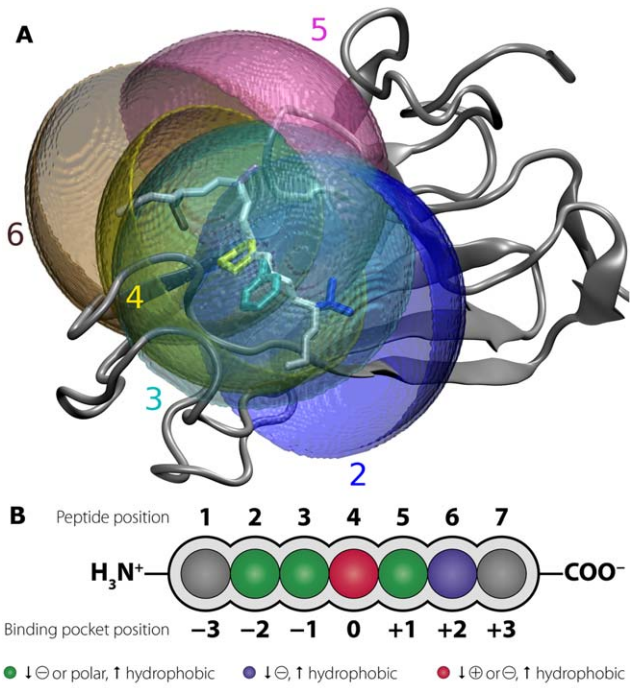
The performance and details of the 9 resulting prediction models are listed in Table II. The evaluation of the original IE matrix on the CD data set showed that the matrix had no distinct predictive value (AUC = 0.48) (see Model 1 in Table II). Additional training on the CD and peptide array (PA_{train}) data sets improved the results by about 20% leading to models with AUC values around 70% (Table II, Models 4 and 7), that is, showing

a decent performance, which is in the same range as the previous Hsp70-peptide prediction models.

However, the performance of the unfitted IE model is about 10% lower than the accuracies of corresponding interaction energy-based matrices for other protein-peptide systems (e.g., for MHC-peptide binding the performance is around 60%). There are two main reasons for this relatively low performance, which can be attributed to the special geometry of the Hsp70 binding site.

First, in our previous work about the BiP-peptide binding features,[14] we observed that the central binding pocket of the binding site, interacting with the 4th residue of the heptamer peptides, is not only the most crucial for peptide specificity, but does also have a distinct conformational flexibility and is predominantly hydrophobic. This allows the pocket to accommodate hydrophobic and aromatic side chains of very different size, whereas charged side chains are normally rejected. In our mutation protocol this incompatibility led the IRECS algorithm to place the charged residues outside the binding pocket. Thus, no meaningful repulsive energies could be obtained, as outside the binding site, the side chains were placed on the protein surface in a more or less "interaction energy neutral position." Therefore, more advanced sampling is necessary to properly describe the interaction between residue four and its binding pocket. As the pocket size and shape can change significantly, depending on the size and type of amino acid bound, we performed MD simulations for all BiP-peptide complexes obtained in our original IRECS-based mutation procedure, in which the peptide was mutated at position 4, that is, each AAAXAAA sequence with X located in or around the central binding pocket. These simulations were expected to provide a realistic picture of the binding site's propensity to adapt to or to reject the corresponding amino acid.

The second accuracy-limiting feature is the open surface-like character of the regions of the binding site surrounding the central binding pocket. As a consequence, different backbone and side-chain positions are possible for the peptide residues in these regions and the exact conformation of the individual residue might depend on its neighboring residues. As the latter interdependence is not included in our alanine-based mutation protocol, this might lead to another drop in accuracy. A straightforward solution to this problem would be to sample all combinations of all 20 amino acids in each of the seven positions. This, however, would lead to 8×10^{15} combinations and is therefore not practically feasible. To overcome this issue, we designed a different, feasible strategy, which is based on a combined structural and interaction analysis of the bound peptide side-chain as obtained by the IRECS-based mutation protocol and the characteristic binding features of the binding site region responsible for binding the respective residue (see Fig. 4). For this purpose, sub-regions of the binding site



C

Peptide residue				
2	3	4	5	6
T429	T429	F452	E428	A455
T435	V430	S453	T429	S456
K436	T454	T454	V430	Q459
Q450	A455	A455	G431	P460
I451	R493	Q459	G432	T461
F452	G494	V462	T463	T463
	V495	T463	K465	K465
		I498	K475	D484
		V500	D484	

Figure 4

Binding site analysis: (**A**) BiP binding site showing the residue centered spheres used for analysis of the individual residue surroundings. (**B**) Schematic drawing of the general properties that each peptide residue should have for good BiP binding properties. (**C**) Residues of BiP, which can interact with the corresponding peptide residue (i.e., which are located within the corresponding sphere).

were defined, including all residues located within a 10 Å sphere around the Cα-atoms of the residue of interest, which will be referred to as "residue sub-pockets." A detailed description of the results of both analyses (MD and binding site analysis) is provided in the Supporting Information (Text S2).

On the basis of the molecular dynamics results (Fig. 5), we first modified the scores at position 4 in the SB-PSSM. A score of 1 was attributed to the side chains that were stable or that moved deeper into the cavity during simulation. To the side chains showing a semi-stable character, low cavity penetration, or missing polar contacts a score of 0.5 was assigned. Finally, amino acids with side chains that could not be placed near or in the binding pocket received a score of 0. The resulting

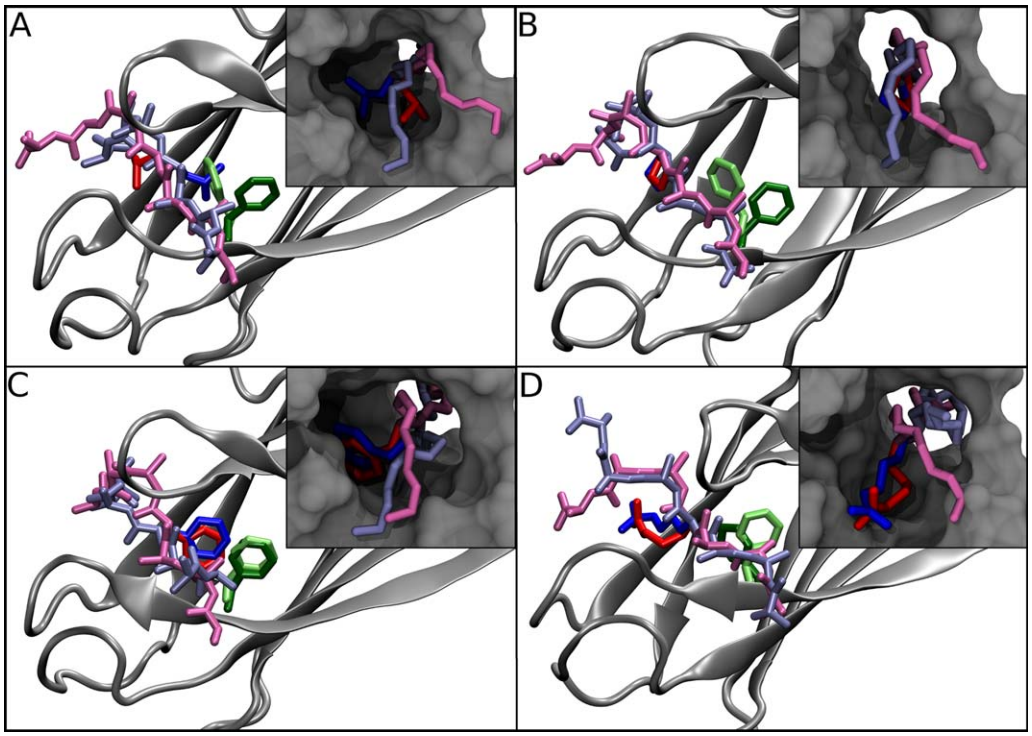


Figure 5

MD results for the AAAXAAA mutants, (**A**) X = L, (**B**) X = P, (**C**) X = F, (**D**) X = R. The starting conformation of the peptides is shown in red (the central amino acid residue in dark red, the rest of the peptide in light red) and the final conformation in blue (the central amino acid residue in dark blue, the rest of the peptide in light blue). The corresponding phenylalanine in the BiP binding site (F452) is shown in light (starting conformation) and dark green (final conformation). The insets show the central binding pocket in surface representation.

matrix is further referred to as IE/4. In a second step, we modified the scores for all other positions based on the binding site analyses. In that case the interaction energy-based scores were retained for "well" placed side chains, which were able to form stable interactions with the binding site. For side chains with medium binding characteristics a score of 0.5 was assigned, and amino acids with side chains unable to form any interaction received a score of 0. This led to the final matrix named IE/BA (Interaction Energy/Binding site Analysis), which contains all modifications (for a more detailed discussion, see Supporting Information).

To train and evaluate the new IE/4 and IE/BA matrices, we applied the same protocol as described above for the IE matrix. All results are listed in Table II, the ROC-curves derived for the Models 1–6 are given in Figure 6 and those for the Models 7–9 are shown in Supporting Information Figure S4. It can be observed that the performance of the unfitted Models (Model 1–3) increases consistently with the introduction of system-specific modifications using the CD data set for evaluation. The MD-based changes for position 4 increase the AUC value from 0.48 (Table II, Model 1) to 0.61 (Table II, Model 2). The final IE/BA matrix (Table II, Model 3), which includes all system-specific changes, features a remarkably high AUC value of 0.83, thus featuring already a very high selectivity even before any experimental data-based fitting of the position-based weights.

Nevertheless, we tried to further improve the performance by fitting weights for each residue position in the peptide, using the CD (Table II, Models 4–6) and PA_{train} (Table II, Models 7–9) data sets. For the IE and IE/4 matrices significant performance improvements could be obtained by fitting these position weights by either of the both training data sets, leading to AUC values for the training sets between 0.68 and 0.72 (Table II, Models 4–5 and 7–8). However, evaluation of the models trained on the PA_{train} data set (Models 7–8) led to inconsistent performance on the PA-based evaluation set (PA_{eval}) (AUCs between 0.57 and 0.74) and only to a very moderate performance of 0.51-0.59 (AUC) on the independent CD data set.

For the final IE/BA model, additional improvement by position-dependent weight fitting was only obtained by using the CD data set (Table II, Model 6), which led to a further increase in performance by 2% (Table II, Model 6 vs. Model 3). However, fitting of the position weights using the PA_{train} data set (Table II, Model 9) led to a

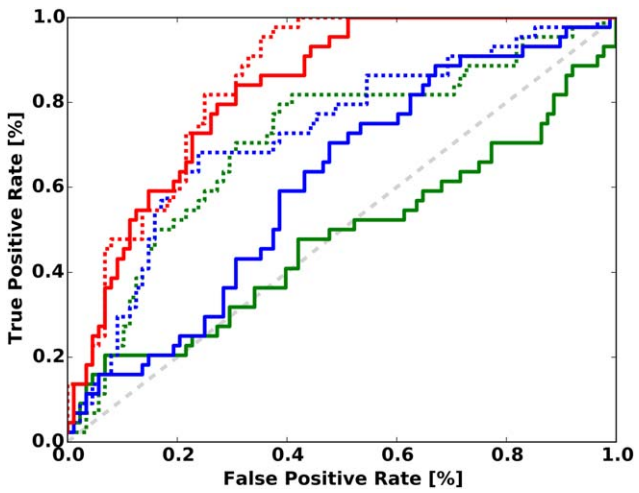


Figure 6

ROC plots for the prediction models from Table II: SB-PSSM: *IE = green, IE/4 = blue, IE/BA = red, original SB-PSSMs = straight line, CD-fitted = dotted line.*

decrease in accuracy for both, on the training set (AUC = 0.65) as well as on the evaluation on the two independent evaluation sets (AUCs of 0.57 and 0.58, respectively).

Overall, all AUC values obtained with models featuring PA_{train} data set-fitted position weights show a lower accuracy than the CD-trained models and vary considerably in their performance. These differences in the performance depend predominantly on the training data set (between 0.51 and 0.74) and only slightly on the SB-PSSM matrix used. This indicates that the predictive capability of the peptide array data is around 65–70%, which is in agreement with the accuracy of the previous prediction models trained on PA data as discussed in the Introduction. It also agrees with the experimental discrepancies between the PA results and the fluorescence anisotropy measurements observed in the present study.

In summary, these results demonstrate the importance of a very careful system-specific design of the SB-PSSM as well as the necessity of a very conservative choice of the experimental data set used for fitting purposes.

Identification of new binding peptide sequences using the final prediction model

Our next goal was to predict new binding heptamers in the PA data set and validate the prediction by measuring their displacement efficiency in fluorescence anisotropy experiments.

In the first study, we predicted the actual binding heptamer stretches within the measured 15mer binding peptides from Table I using the final CD-fitted IE/BA-based prediction model (Model 6 in Table II). As it was experimentally observed that peptides can bind in both sequence directions (Supporting Information Fig. S1),[15] we tested both, the forward and the reverse sequences. The number of resulting binders per 15mer together with the achieved maximum score are provided in Table I. A clear correlation can be found between the maximum score, the number of predicted binders, and the displacement efficiency. The correlation between the number of predicted binders and the displacement efficiency suggests that there might be multiple alternative binding registers, that is, more than one binding heptamer, within one 15mer. This would increase the overall stochastic probability of a binding event, potentially leading to more efficient displacement of HTFPAVLGSC. In general, all 15mers with a high displacement efficiency include at least seven predicted binding heptamers with a score of 0.84 or higher.

In a second study, we performed binding predictions on the whole PA data set to evaluate the prediction capacities of the IE/4 and the IE/BA matrices comprehensively. The predictions on the whole data set were performed using a PA-trained IE/4 model, as this was the best performing model available at that time. From the prediction results we selected the five highest ranking binding heptamer stretches (HP1-HP4, HP10) and the five lowest ranking nonbinding heptamers of the whole PA data set (HP5-HP9) (Supporting Information Table SII). Comparison of the experimental PA data and the prediction results showed a strong correlation, as all predicted strong binding heptamers were part of the top-10 VpreB/VEGF 15mer sequences with the highest signal intensities in the PA assays (Table SI) and all nonbinding heptamers could be located in 15mer peptides with very low intensities, respectively. This can be attributed to the PA-based training of the model. Afterward, the displacement efficiency was measured for all selected peptides by fluorescence anisotropy spectroscopy. The corresponding results are provided in Table III, Supporting Information Table SII, and Figure S3. The peptides HP4, HP7, HP8, HP9, and HP10 were identified as binders, displacing HTFPAVLGSC with different efficiencies. In contrast, the remaining heptapeptides HP1, HP2, HP3, HP5, and HP6 did not show any change in the anisotropy signal indicating that they could not be recognized by BiP. One peptide, HP2, repeatedly gave artifactual signals, perhaps caused by its instability in solution as calculated by ProtParam (http://web.expasy.org/protparam). Comparing these anisotropy data with the PA signal intensities, the correlation is again rather low (40%), in the same range as the correlation for the measured 15mers (Table I). Strikingly, the overall displacement efficiency for the predicted heptapeptide substrates was much higher than for the array-derived pentadecapeptides. Especially, peptide HP10 proved to be a potent BiP binder as it exerted an efficiency of 107%, characteristic for a true substrate of BiP (Table III, Supporting Information Fig. S3). The known binder

HTFPAVL yielded an efficiency of ~113% which is in the same range as HP10. The dissociation kinetics of the former and later peptides were also found to be very similar with 0.096 and 0.100 min^{-1} for HTFPAVL and HP10, respectively (Table III, Supporting Information Fig. S3, Table SII). HP8 was also able to strongly compete with HTFPAVLGSC for BiP binding, as reflected by a displacement efficiency of ~76% (Table III). Here, the competition occurred at a slightly more elevated rate of 0.113 min^{-1}, a value still comparable to the control (Table III, Supporting Information Table SII). The highest displacement rate was achieved by HP7 ($\lambda_{off} = 0.184$ min^{-1}) but this candidate could decrease the anisotropy signal only by about 49% (Table III, Supporting Information Table SII).

Finally, we performed additional predictions for all heptamers using the IE/BA matrix. Because the CD data set contains nonbinding sequences from the PA data set, the prediction on the PA data set might be biased using the CD-fitted IE/BA model. Therefore we used the non-fitted IE/BA matrix (Model 3 in Table II) for these calculations. Comparing these results to the anisotropy and the PA-based experimental binding data revealed a very strong correlation with the anisotropy data, as all peptides except one could be predicted correctly, thus leading to an accuracy of 90% with our non-fitted IE/BA model (Model 3 in Table II). At the same time the correlation to the PA-based data is in the same low range as the correlation between the two experimental data sets. These results demonstrate that the prediction model fitted to the PA data (PA-fitted IE/4 model) predicts this data very well, but not the peptide binding properties in solution, whereas the unfitted "*ab initio*" model shows a very high predictive power for peptide binding in solution. This confirms our original hypothesis that efficient, sequence-based peptide binding prediction should be possible on the basis of structural–biophysical interaction data and properties.

Thus, both independent evaluation studies on the 15mer and 7mer peptide sets demonstrate the high predictive power of the final unfitted as well as the CD-fitted IE/BA-based prediction models.

Molecular docking and MMPBSA/MMGBSA binding affinity estimations

The results described above show that a reliable and selective identification of potential binding sequences of BiP is possible with the IE/BA prediction model. However, due to the special binding site features discussed above, an additional structure-based refinement of the prediction results is still necessary to reliably predict the binding direction (forward or reverse binding mode) and to provide an accurate estimation of the binding affinity for the ranking of the peptides.

Thus, in a second step, the DynaDock approach was evaluated for the prediction of Hsp70-peptide complex structures. The method was already successfully applied in a former study to identify important structural features of BiP-peptide binding.[14] Here we additionally evaluate its capability to predict the correct forward/backward binding mode of the peptide based on a new set of recent experimental structures of DnaK-peptide complexes.[15] As DnaK is the bacterial homolog of BiP, it features the same overall Hsp70-binding site geometry and is thus a suitable evaluation system. A detailed discussion of the evaluation results is provided in the Supporting Information (Text S1). In the performed molecular docking simulations, all peptides were docked in their experimental and in the alternative reverse orientation. These redocking experiments showed that in all cases the best-scoring peptide conformations feature RMSD values for the peptides smaller than 2 Å, demonstrating that with the applied DynaDock protocol it is not only possible to obtain accurately placed peptide conformations but also to identify them by the Hsp70-peptide interaction energy (Supporting Information Table SVI). In addition, in all cases except one, the binding orientation could be predicted correctly.

These results indicate that by combining fast IE/BA-based prediction of potential BiP-binding sequences and successive DynaDock simulations for the identified binders, highly accurate prediction results can be expected.

To specifically evaluate the hierarchical pipeline as shown in Figure 2 for BiP-peptide binding predictions, we performed molecular docking simulations for three of the predicted heptamers from Table III (HP4, HP8, and HP10, highlighted in bold typeface) using the same conditions as for the DynaDock evaluation studies. For the best-scoring docked complexes 20 ns of molecular dynamics simulations and successive MMPBSA/MMGBSA calculations were performed based on the last 5 ns of the simulations (Table IV). For the latter calculations the AMBER14 software package was used together with the AMBER ff99SBildn force field instead of the OPLS-all-atom force field used in the DynaDock studies, as in other studies performed in our group we obtained the best MMPBSA results by using the AMBER 14 suite of programs.[40] The results are provided in Table IV. For this analysis, also K_d values were measured experimentally for these peptides using titration experiments.

For all three true binders, reasonable K_d values for BiP recognition could be obtained with 12.0 μM for peptide HP4, 17.9 μM for peptide HP8, and 9.7 μM for peptide HP10 (Fig. 3, Table IV). Compared to HTFPAVL ($K_d = 11.1$ μM[14]), the affinities for BiP binding of all three true binders were in the same order of magnitude as expected for heptapeptides. For HP10, the affinity also correlated with the peptide's displacement efficiency since this peptide with the highest BiP affinity also represented the most efficient competitor for HTFPAVLGSC.

Regarding the molecular docking results (Table IV, FFscore), HP8 and HP10 should bind in a forward direction and HP4 in reverse. These binding modes are confirmed by follow-up MMPBSA/MMGBSA calculations, proving the robustness of the DynaDock-based prediction of the peptide orientation, especially considering that not only different methods, but also different force fields were used. Further analyzing the MMGBSA and MMPBSA results of the free energy of binding, qualitative agreement with the measured K_d values can be observed for both methods. The MMPBSA approach provides the same binding affinity-based ranking for all peptides as the experiment (if HP4 is considered to bind in reverse mode), whereas the MMGBSA method leads to the ranking HP4 < HP10 < HP8, which are both very good results considering the rather small differences in the measured K_d values. However, only the MMPBSA values are quantitatively in the same range as the measured K_d values (i.e., to the corresponding ΔG values, which vary around −6 kcal/mol), whereas the MMGBSA data are ten times larger. The overestimation of binding affinities by the MMGBSA method agrees with all previous MMGBSA studies on protein–ligand binding affinity estimation and is most likely caused by use of the simple SASA (solvent accessible surface area) approximation of the non-polar term.[41] For the MMPBSA implementation a more advanced non-polar treatment is available and thus used in this study (see Materials and Methods), which allowed a quasi-quantitative estimation of experimental values by the MMPBSA method.

DISCUSSION

In the first part of this work, we evaluated the overall accuracy and usefulness of peptide array measurements for the identification of binding peptides to the Hsp70 chaperone BiP. The inclusion of known binding sequences into the measured data set showed that the BiP-peptide affinity does not correlate well with the signal intensity on the chip. The well-studied C_H1 peptide HTFPAVL with demonstrated high affinity for BiP[14] was not present in the 10 C_H1 peptides with the highest intensity, while SVFPLAP with its lower BiP affinity is contained in three high-ranking peptides. To further evaluate this observation, several 15mer and heptamer peptides, which showed high intensities in the peptide array, were reinvestigated by fluorescence anisotropy spectroscopy experiments in solution (Tables I and III and Supporting Information Table SII). The results did not show a clear correlation between the peptide array intensities and the displacement efficiencies of the corresponding peptides in solution. This discrepancy indicates that the environment in solution and on the chip varies greatly, possibly due to the immobilization of the peptides and the accessible peptide concentration on each spot. Another potential reason might be the different lengths of the tested peptides, as 15mer peptides do have the potential to form short secondary structure elements, whereas heptamers do not.[42] This might alter the accessibility and thus the binding properties of a 7mer binding stretch in a 15mer peptide compared to individually tested 7mers, which could also be a potential explanation for the differences in the displacement efficiency values in Tables I and III, as the predicted heptapeptides show a tendency toward higher displacement values but at slower kinetics compared to the evaluated 15mers (Supporting Information Table SII). However, the number of peptides analyzed here is small and a comprehensive, systematic study would be necessary to draw final, unbiased conclusions. Overall, these results show that the peptide array data allows only a qualitative binary distinction between potential binding and nonbinding sequences. A subsequent BiP interaction analysis of the binding heptapeptides either experimentally in solution or computationally via docking calculations is still necessary.

As previous sequence-based prediction models for Hsp70-peptide binding are mainly using peptide array data, this might explain the limited accuracy of these models. Interestingly, in our case the use of the PA-based training and evaluation sets for position weight fitting of the different SB-PSSM matrices led to models which feature a similar accuracy as the previously published models (AUC values of 0.65 to 0.70 for the training and 0.51–0.74 for the validation sets) independent of the SB-PSSM used. PA_{train}-based position weight fitting to the IE/BA SB-PSSM even led to a decrease in accuracy. All PA_{train}-based models show an especially low AUC of around 0.50–0.60 if evaluated on the PA-independent CD data set, which predominantly contains peptide binding data determined by solution-based studies. In addition, in the second application study, the evaluation of the 7mer peptide binding properties (Table III), the PA-trained IE/4 model agreed perfectly with the experimental PA data, but showed only a 40–60% correlation to the fluorescence anisotropy displacement efficiency measurements and the predictions by the non-fitted IE/BA model.

All these observations reflect the discussed limited correlation between the experimental peptide array data on one side and the anisotropy results on the other, which can also explain why sequence-based prediction models trained on the peptide array data show a limited accuracy of about 70%. In addition, these models might be biased toward the PA data, thus explaining their even lower capability for predicting peptide binding in solution. Similar results were obtained by van Durme *et al.* who showed that although the performance of their sequence-based prediction model could be increased by the inclusion of structure-based data for the benchmark set (Matthews Correlation Coefficient (MCC) = 0.756), a

strong decrease was observed for the validation set (MCC = 0.375) compared to the performance of the structure-based model alone (MCC = 0.593).

In addition, next to the intrinsic inaccuracies in the peptide array data, which are not Hsp70 specific, there are two more major reasons, why predictions based solely on sequence data from peptide (non)binders might not perform as well for Hsp70-peptide binding as for other protein–peptide systems. The predominant problem regarding the experimental peptide data is the selection of the exact binding heptamer stretches in the longer peptides measured. This computational selection was found to impair the prediction accuracy considerably also for other protein-peptide systems with experimentally undefined peptide binding core sequences (see MHC class II-peptide binding predictions). A previous study on DnaK binding could indeed show that the performance of the bacteriophage-based prediction model could be improved by the inclusion of specific heptamer binding data.[17,24] Van Durme *et al.*[25] relied on threading experiments to identify the correct binding heptamer within the 15mer peptides. As in our study successive peptides were tested, "average" heptamer-specific intensities could be calculated directly from the measured data, which should be more consistent with the measured data (see Material and Methods). Nevertheless, both approaches introduce an additional error, which is difficult to estimate. The second crucial issue is based on the recent observation that peptides can bind in both backbone directions, as the Hsp70 binding site is nearly symmetrical. For the training of a prediction model on peptide sequence data, a consistent binding direction is normally assumed, as experimental binding assays do not provide any information about the peptide's binding direction. This could lead to an additional error.

To minimize the influence of the issues discussed above, a structure-based position-specific scoring (SB-PSSM) was developed based on structural modeling and analysis, using the experimental data only to fit general weights for the individual residue positions in the peptide. As these weights are predominantly determined by the properties and shape of the binding site, they are less critically dependent on the correct orientation of all peptides in the data set. However, we still need to assume that the majority of the peptides bind in the forward direction as suggested by all fitted models.[15]

The evaluation of our first prediction model (IE) demonstrated that, due to the special geometry of the Hsp70 binding site, the system-independent straightforward mutation-based strategy needed to be adapted manually. Thus, we developed an improved SB-PSSM (IE/BA), which is not only based on interaction energy data, but also on MD simulations and on a static analysis of the binding site properties of BiP. Our final prediction matrix shows a very high selectivity and accuracy with an AUC value of 0.83 even without any parameterization on experimental data (Model 3, Table II). Position-weight fitting using the CD data set led to a further increase in accuracy of about 2% (Table II, Model 6 vs. Model 3). This small increase is presumably due to the limited size of the CD data set. It demonstrates, however, that the accuracy of the final model can still be improved by the use of accurate experimental binding data, indicating that with a large data set of certified binding and nonbinding heptamer sequences, even higher performances are possible.

Similar observations were made in several previous studies, in which it was demonstrated that the use of structure or interaction energy-based data can improve the performance and the robustness of a prediction model considerably, and lessen its dependence on the experimental data used.[25,27,39]

Comparing our approach to the previously published prediction models, the predominant difference is that all previous models are exclusively based on the sequence and binding data of the peptide substrates, but do not explicitly include binding site features. In contrast, we did not use the sequence information from peptide binding experiments for the PSSM, but instead included information of the structural features of the protein's binding site obtained by structural analysis and calculations together with interaction energies of the individual peptide residues. Thus our approach is *a priori* independent of any experimental binding information. Nevertheless, we showed that its performance can be further improved by an additional position-weight fitting to such data. With that strategy we were able to develop a prediction model, which shows a higher performance on an independent data set than the previous approaches based on peptide sequence data. However, the procedure cannot be automated yet and must be performed individually for each new protein-peptide system. Thus its development is much more time-consuming than the training of a standard prediction approach solely based on peptide sequence data. However, to our knowledge, no other comparable model, which performs equally well, exists at the moment. Therefore, if no peptide binding data is available, the current standard procedure is to perform molecular docking calculations. Although such calculations can lead to good results, they are too time-consuming for the screening of whole protein sequences and can only be performed for a preselected set of peptides, which needs to be obtained either experimentally or by non-system specific sequence analysis studies.

The practical evaluation of Model 3 (i.e., non-fitted IE/BA-based model) on the 7mer peptides from Table III showed an excellent correlation between the measured anisotropy data and the prediction results. Using Model 3, the binding properties of 9 out of 10 peptides could be predicted correctly. This independent evaluation study demonstrates impressively the power of "*ab initio*" structure-based prediction models. Therefore the new

model can be used for a robust and solid identification of potential peptide binding sequences of BiP. As the selectivity of the model is solely based on structural and energetic data from the binding site analyses, it does not contain any bias with respect to the actual binding orientation of the peptides.

Using subsequent molecular docking calculations, very accurate bound peptide conformations could be obtained. In addition, it was possible to identify the correct binding direction of the peptides by their interaction energies. MMPBSA binding affinity estimations based on the energetically best docking solutions allowed to rank potential binding peptides with good accuracy and to obtain values that are quantitatively in the same range as the experimental values.

CONCLUSIONS

Peptide arrays are a valuable tool for a fast, first screening of protein sequences and provide a general idea whether a peptide should be considered as binder or nonbinder. However, the BiP-binding properties of these candidates can differ considerably if measured in solution and thus the peptide array results needs to be verified in this case.

For the development of a SB-PSSM-based prediction model for BiP-peptide binding, the peptide array data was of limited value. This is in agreement with the limited performance of previous sequence- and structure-based prediction models of Hsp70-peptide binding, which rely mainly on experimental peptide binding data. However, very good prediction performances could be obtained with our final "*ab initio*" structure-based IE/BA prediction model. The corresponding SB-PSSM, optimized by careful analysis of the binding site properties of BiP, already showed high selectivity (AUC = 0.83) without any fitting to experimental data. By additional parameterization of position weights on the basis of solution-based, verified heptamer binding data, the accuracy of the model could be further improved. This is an encouraging result, as it demonstrates that it is possible to obtain highly predictive models for protein-peptide binding by system-specific structural modeling and analysis studies alone. This general concept allows the development of sequence-based prediction models for protein–peptide systems for which no or only few experimental data sets is available and for which no sequence-based prediction models exist currently.

REFERENCES

1. Karlin S, Brocchieri L. Heat shock protein 70 family: multiple sequence comparisons, function, and evolution. J Mol Evol 1998;47:565–577.
2. Gething M-J, Blond-Elguindi S, Buchner J, Fourie A, Knarr G, Modrow S, Nanu L, Segal M, Sambrook J. Binding sites for Hsp70 molecular chaperones in natural proteins. Cold Spring Harbor Symp Quantitative Biol 1995;60:417–428.
3. Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ER. Solution conformation of wild-type *E. coli* Hsp70 (DnaK) chaperone complexed with ADP and substrate. Proc Natl Acad Sci USA 2009;106:8471–8476.
4. Zhu X, Zhao X, Burkholder WF, Gragerov A, Ogata CM, Gottesman ME, Hendrickson WA. Structural analysis of substrate binding by the molecular chaperone DnaK. Science 1996;272:1606–1614.
5. Goloubinoff P, De Los Rios P. The mechanism of Hsp70 chaperones: (entropic) pulling the models together. Trends Biochem Sci 2007;32:372–380.
6. Swain JF, Dinler G, Sivendran R, Montgomery DL, Stotz M, Gierasch LM. Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker. Mol Cell 2007;26:27–39.
7. Takeda S, McKay DB. Kinetics of peptide binding to the bovine 70 kDa heat shock cognate protein, a molecular chaperone. Biochemistry 1996;35:4636–4644.
8. Marcinowski M, Höller M, Feige MJ, Baerend D, Lamb DC, Buchner J. Substrate discrimination of the chaperone BiP by autonomous and cochaperone-regulated conformational transitions. Nat Struct Mol Biol 2011;18:150–158.
9. Gragerov A, Gottesman ME. Different peptide binding specificities of hsp70 family members. J Mol Biol 1994;241:133–135.
10. Hageman J, van Waarde MA, Zylicz A, Walerych D, Kampinga HH. The diverse members of the mammalian HSP70 machine show distinct chaperone-like activities. Biochem J 2011;435:127–142.
11. Wiech H, Buchner J, Zimmermann M, Zimmermann R, Jakob U. Hsc70, immunoglobulin heavy chain binding protein, and Hsp90 differ in their ability to stimulate transport of precursor proteins into mammalian microsomes. J Biol Chem 1993;268:7414–7421.
12. Flynn GC, Pohl J, Flocco MT, Rothman JE. Peptide-binding specificity of the molecular chaperone BiP. Nature 1991;353:726–730.
13. Rüdiger S, Buchberger A, Bukau B. Interaction of Hsp70 chaperones with substrates. Nat Struct Mol Biol 1997;4:342–349.
14. Marcinowski M, Rosam M, Seitz C, Elferich J, Behnke J, Bello C, Feige MJ, Becker CFW, Antes I, Buchner J. Conformational selection in substrate recognition by Hsp70 chaperones. J Mol Biol 2013;425:466–474.
15. Zahn M, Berthold N, Kieslich B, Knappe D, Hoffmann R, Strater N. Structural studies on the forward and reverse binding modes of peptides to the chaperone DnaK. J Mol Biol 2013;425:2463–2479.
16. Rudiger S, Mayer MP, Schneider-Mergener J, Bukau B. Modulation of substrate specificity of the DnaK chaperone by alteration of a hydrophobic arch. J Mol Biol 2000;304:245–251.
17. Knarr G, Modrow S, Todd A, Gething MJ, Buchner J. BiP-binding sequences in HIV gp160. Implications for the binding specificity of BiP. J Biol Chem 1999;274:29850–29857.
18. Rüdiger S, Germeroth L, Schneider-Mergener J, Bukau B. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. EMBO J 1997;16:1501–1507.
19. Munro S, Pelham HR. An Hsp70-like protein in the ER: identity with the 78 kd glucose-regulated protein and immunoglobulin heavy chain binding protein. Cell 1986;46:291–300.
20. Dudek J, Greiner M, Muller A, Hendershot LM, Kopsch K, Nastainczyk W, Zimmermann R. ERj1p has a basic role in protein biogenesis at the endoplasmic reticulum. Nat Struct Mol Biol 2005;12:1008–1014.
21. Kabani M, Kelley SS, Morrow MW, Montgomery DL, Sivendran R, Rose MD, Gierasch LM, Brodsky JL. Dependence of endoplasmic reticulum-associated degradation on the peptide binding domain and concentration of BiP. Mol Biol Cell 2003;14:3437–3448.
22. Kassenbrock CK, Garcia PD, Walter P, Kelly RB. Heavy-chain binding protein recognizes aberrant polypeptides translocated in vitro. Nature 1988;333:90–93.

23. Alder NN, Shen Y, Brodsky JL, Hendershot LM, Johnson AE. The molecular mechanisms underlying BiP-mediated gating of the Sec61 translocon of the endoplasmic reticulum. J Cell Biol 2005;168: 389–399.
24. Blond-Elguindi S, Cwirla SE, Dower WJ, Lipshutz RJ, Sprang SR, Sambrook JF, Gething M-JH. Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. Cell 1993;75:717–728.
25. Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, Schymkowitz J. Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. PLoS Computat Biol 2009;5:e1000475.
26. Knarr G, Gething M-J, Modrow S, Buchner J. BiP binding sequences in antibodies. J Biol Chem 1995;270:27589–27594.
27. Roomp K, Antes I, Lengauer T. Predicting MHC class I epitopes in large datasets. BMC Bioinform 2010;11:90.
28. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
29. Hartmann C, Antes I, Lengauer T. IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Sci Publ Protein Soc 2007;16:1294–1307.
30. Hartmann C, Antes I, Lengauer T. Docking and scoring with alternative side-chain conformations. Proteins 2009;74:712–726.
31. Antes I. DynaDock: a new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. Proteins 2010; 78:1084–1104.
32. Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
33. Liebscher M, Roujeinikova A. Allosteric coupling between the lid and interdomain linker in DnaK revealed by inhibitor binding studies. J Bacteriol 2009;191:1456–1462.
34. Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, Cheatham TE, III, Darden TA, Duke RE, Gohlke H, Goetz AW, Gusarov S, Homeyer N, Janowski P, Kaus J, Kolossváry I, Kovalenko A, Lee TS, LeGrand S, Luchko T, Luo R, Madej B, Merz KM, Paesani F, Roe DR, Roitberg A, Sagui C, Salomon-Ferrer R, Seabra G, Simmerling CL, Smith W, Swails J, Walker RC, Wang J, Wolf RM, Wu X, Kollman PA. AMBER 14. University of California, San Francisco; 2014.
35. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE, III. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 2000;33:889–897.
36. Srinivasan J, Cheatham TE, III, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate−DNA helices. J Am Chem Soc 1998;120:9401–9409.
37. Chong LT, Duan Y, Wang L, Massova I, Kollman PA. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. Proc Natl Acad Sci US A 1999;96:14330–14335.
38. Bill R, Miller BR, III, McGee TD, Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA.py: an efficient program for end-state free energy calculations. J Chem Theory Comput 2012;8:3314–3321.
39. Antes I, Siu SW, Lengauer T. DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. Bioinformatics 2006;22:e16–e24.
40. Salomon-Ferrer R, Götz AW, Poole D, Grand SL, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. J Chem Theory Comput 2013;9:3878–3888.
41. Sun H, Li Y, Tian S, Xu L, Hou T. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. Phys Chem Chem Phys 2014;16: 16719–16729.
42. Luo P, Baldwin RL. Interaction between water and polar groups of the helix backbone: An important determinant of helix propensities. Proc Natl Acad Sci USA 1999;96:4930–4935.