Bioinformatics, 36(3), 2020, 735–741
doi: 10.1093/bioinformatics/btz670
Advance Access Publication Date: 28 August 2019
Original Paper

OXFORD

Sequence analysis

ChaperISM: improved chaperone binding prediction using position-independent scoring matrices

M. B. B. Gutierres[1], C. B. C. Bonorino[1,2] and M. M. Rigo[3,*]

[1]Laboratório de Imunoterapia, Departamento de Ciências Básicas da Saúde, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, Brazil, [2]School of Medicine, Department of Surgery, University of California San Diego, La Jolla, CA 92037, USA and [3]Escola de Medicina, Laboratório de Imunologia Clínica e Experimental, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on January 25, 2019; revised on August 20, 2019; editorial decision on August 21, 2019; accepted on August 26, 2019

Abstract

Motivation: Understanding the mechanisms of client protein interaction with Hsp70 chaperones is essential to analyze the complex dynamics in the context of normal or dysregulated metabolism. Because Hsp70 can bind millions of proteins, including key molecules involved in processes of stemness, tumorigenesis and survival, *in silico* prediction of Hsp70 interactions has great value in validating possible new clients. Currently, two algorithms are available to predict binding to DnaK—the bacterial Hsp70—but both are based on amino acid sequence and energy calculations of qualitative information—binders and non-binders.

Results: We introduce a new algorithm to identify Hsp70 binding sequences in proteins—ChaperISM—a position-independent scoring matrix trained on either qualitative or quantitative chemiluminescence data previously published, which were obtained from the interaction between DnaK and different ligands. Both versions of ChaperISM, qualitative or quantitative, resulted in an improved performance in comparison to other state-of-the-art chaperone binding predictors.

Availability and implementation: ChaperISM is implemented in Python version 3. The source code of ChaperISM is freely available for download at https://github.com/BioinfLab/ChaperISM.

Contact: mauricio.rigo@pucrs.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The 70 kDa Heat Shock Proteins (Hsp70) comprise an extremely conserved family of molecular chaperones found in prokaryotes and eukaryotes that assists both folding and degradation of proteins. All Hsp70 members encompass the same structural organization: an N-terminal Nucleotide Binding Domain (NBD) connected by a short, flexible linker to a Substrate Binding Domain (SBD), which is followed by an intrinsically disordered C-terminal portion (Fig. 1) (Goloubinoff, 2017). Hsp70s are highly dynamic entities, and their folding function depends on ATPase cycles fine-tuned by proteins called co-chaperones. Co-chaperones aid Hsp70 either delivering substrates and catalyzing ATP hydrolysis (e.g. J-domain Proteins or Hsp40), or releasing ADP from NBD (Nucleotide Exchange Factors, or NEFs) (Mayer and Bukau, 2005). The NBD and SBD allosterically communicate each other to control conversion from an SBD with low substrate affinity (when ATP is bound to NBD, also called open conformation) to a high substrate affinity (when NBD is ADP-bound, also called closed conformation). In both situations, the SBD interacts with short segments of hydrophobic and positively charged exposed residues (Mayer and Kityk, 2015). This profile of substrate recognition has been established from early peptide library and phage display studies. Arrays of cellulose-bound peptides screened with *Escherichia coli* DnaK, a bacterial member of the Hsp70 family, further confirmed this broad capability of substrate interaction. In fact, most of the current understanding of how Hsp70 and substrates interact came from studies on DnaK itself (Clerico *et al.*, 2015). Because of the pivotal role of Hsp70 chaperone—which basically regulates all aspects involving protein folding, translocation, disassembly of oligomeric complex and aggregates—the prediction of chaperone binding is crucial to understand both (i) how biologically relevant substrates interact and (ii) how mutations that affect chaperone binding affect diseases (El-Kasaby *et al.*, 2014; Goswami, 2015; Gowda *et al.*, 2018; Halder *et al.*, 2011; Lee *et al.*, 2015; Moreira *et al.*, 2013; Panda and Suresh, 2014; Rauch *et al.*, 2016; Rosam *et al.*, 2018; Solayman *et al.*, 2017).

Rüdiger *et al.* built the first published DnaK binding predictor in an elegant study screening cellulose membrane-bound peptides to

Downloaded from https://academic.oup.com/bioinformatics/article/36/3/735/5555810 by guest on 10 April 2024

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

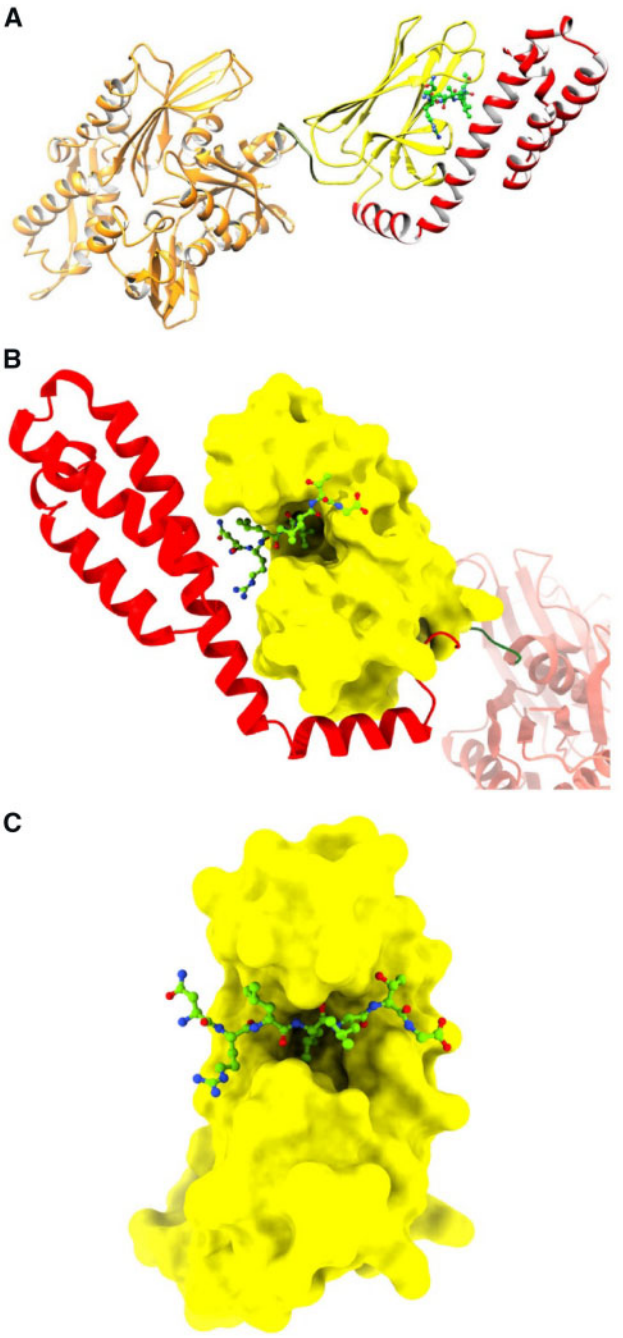

Fig. 1. DnaK domains. (**A**) DnaK in ribbon representation (PDB ID: 2KHO) in the closed state with NBD, linker region, SBD, and C-terminal 'lid'. The peptide NRLLLTG (PDB ID: 1DKX) is complexed to SBD and represented as ball and sticks. (**B**) SBD region with surface representation in complex with NRLLLTG peptide, with the 'lid' region in the closed state. (**C**) SBD surface forms a pocket that accommodates the peptide

understand the molecular basis of DnaK-substrate interaction. Binding and non-binding sequences were exploited to build a position-specific scoring matrix (PSSM). The algorithm computed the statistical energy contribution from the sequence alignment of DnaK binders and non-binders, and the relative occurrence of each amino acid was converted in a score for each position in a 20x13 PSSM for both flanking regions and hydrophobic core (Rüdiger *et al.*, 1997).

Posteriorly, Van Durme *et al.* employed the same screening methodology to develop a sequence-based PSSM. Additionally, this PSSM was combined with a structure-based PSSM derived from interaction energy calculations from a peptide-bound SBD crystallographic structure from DnaK (Van Durme *et al.*, 2009; Zhu *et al.*, 1996). First, the SBD-bound peptide NRLLLTG was converted to a polyalanine sequence and mutated in each position by the 19 remaining amino acids. The structure-based matrix was determined by computing the ΔG difference from poly-alanine reference to all possible amino acid substitutions using the FoldX force field (Schymkowitz *et al.*, 2005). The final PSSM was called LIMBO and made available through a web server currently on http://limbo.switchlab.org.

Despite similarities in the data collection step between Rüdiger and LIMBO PSSMs, there are differences regarding their development. Rüdiger PSSM derives from alignments involving a hydrophobic core of five amino acids and two flanking regions of four residues each. LIMBO, on the other hand, results from an alignment of heptamers. However, the most profound difference relies on the combination of energy calculations to the sequence-based model of LIMBO, which is an approach that usually improves the model's capability to predict unseen examples. In terms of predictive performance, LIMBO is a superior predictor when validations are based on the area under Receiver Operating Characteristic curves (auROC) on the benchmark dataset (Van Durme *et al.*, 2009). In an independent validation set, however, LIMBO superiority is less evident, because both models dominate different regions of the ROC space (Provost *et al.*, 1998).

Both Rüdiger and LIMBO PSSMs are methods to predict the binding of peptides to the bacterial DnaK chaperone. However, a predictor called BiPPred that profiles binding to the BiP chaperone, the endoplasmic member of the Hsp70 family, is also available. Similarly to LIMBO and Rüdiger PSSMs, BiPPred relies on a PSSM to perform predictions, although this matrix was built upon the analysis of BiP binding sites, energy estimations and molecular dynamics simulations. This structure-based PSSM presented a good predictive performance, showing a small improvement when fitted to experimental data (Schneider *et al.*, 2016).

All of these chaperone binding prediction methods have in common the use of PSSMs. However, other matrices could be used to evaluate the prediction capacity of a classifier, such as the position-independent scoring matrix (PISM). The PISM was already used aiming the binding prediction of epitopes to MHC-I proteins (Antes *et al.*, 2006); however, results were not satisfactory, possibly because MHC-I binding properties are very stringent (e.g. different anchor pockets, and different auxiliary residues in different allotypes). Hsp70 chaperones, however, are much less restrictive in binding to its clients, suggesting a potential situation where PISMs could be more appropriate than PSSMs.

Several applications benefit from quantitative measurements to improve predictions, instead of considering only the frequency information derived from a sequence alignment (Peters *et al.*, 2003; Peters and Sette, 2005; Tenzer *et al.*, 2005). Both available DnaK binding predictors are based on the sequence alignment of two qualitatively distinct classes of sequences: binders and non-binders. Thus, we investigated the following hypothesis: is it possible to improve predictions by considering the quantitative DnaK binding information? To answer this question, we collected raw DnaK detection data from the western blot chemiluminescence reaction of cellulose membrane-bound peptides from Van Durme *et al.* Here we present a new DnaK binding predictor—ChaperISM—based on a position-independent scoring matrix (PISM) trained on either qualitative or quantitative DnaK detection data. Both versions of our method achieved superior performance concerning chaperone binding prediction tools on two independent validation sets that benchmark distinct Hsp70 chaperones.

2 Materials and methods

2.1 Data collection and normalization

We obtained peptide sequences and raw information for DnaK binding detection from Van Durme *et al.*, summarized in Supplementary

Downloaded from https://academic.oup.com/bioinformatics/article/36/3/735/5555810 by guest on 10 April 2024

Table S1. Value distribution is demonstrated in Supplementary Figures S1–S5. Our final dataset of peptides consisted of experiments from Membrane A (Group 1 and Group 2) and Membrane B (Group 1 and Group 2), totalized 268 peptides (7-mer to 20-mer). We log 10 converted raw values after being normalized while preserving the original ratio (Equation 1). Each membrane was processed independently as:

$$NewValue = \log\left(NewMin + \frac{(OldValue - OldMin) \times (NewMax - NewMin)}{OldMax - OldMin}\right) \tag{1}$$

OldValue is the quantitative raw information of DnaK binding, *NewValue* is the transformed value, *OldMin* and *OldMax* are the minima and maxima values from the old distribution, and *NewMax* and *NewMin* are 10 000 and 1, respectively. This assures that *NewValue* distribution was comprised inside the range of 0 and 4.

2.2 FoldX calculations

First, the 268 peptides were split into all possible heptamers, generating a total of 1488 sequences. In the case of true binders, longer peptides could bear subsequences that differently contribute to its detection. For that reason, we performed energy calculations using a 0.5 kcal mol^{-1} cut-off, as previously described (Van Durme *et al.*, 2009), to find different frames from the same peptide with the highest binding affinity among the true positives. Because the original training and validation sets from LIMBO PSSM development are not available, we took the effort to follow the same preprocessing procedures to minimize possible discrepancies in the datasets. Thus, a NewValue cutoff $\geq$ 3.0 was arbitrarily chosen to separate peptides as binders, to assure the same dataset as (Van Durme *et al.*, 2009). Energy calculations with FoldX force field were employed to select the lowest interaction energy heptamer to DnaK SBD. To do so, we used the crystal structure of a DnaK SBD in complex to a peptide substrate, the heptamer NRLLLTG (PDB ID: 1DKX) (Zhu *et al.*, 1996). We mutated this peptide to each desired heptamer using FoldX command 'build model' and calculated its energy using 'analyse complex' command. For each true binder source peptide, we kept only the heptamer with the lowest FoldX interaction energy value. Other heptamers were maintained if their energy binding values were inside the range of 0.5 kcal mol^{-1} compared to the best scored one. Defining this interaction energy range allows more than one heptamer from the same high affinity source peptide to be included in the final dataset (as in Van Durme *et al.*, 2009). Finally, we represented heptamers with the same sequence as the average of their detection scores (Fluxogram in Supplementary Fig. S6).

2.3 Training and validation sets

After preprocessing of quantitative data, FoldX based energy filter returned a total of 892 non-redundant heptamers, each one associated to a value that represents the logarithm of the detection of chemiluminescent reaction. This dataset was divided as positive (binders) and negative (non-binders) examples, and then used to validate the predictors' performance through auROC (area under Receiver Operator Characteristic curves) and auPR (area under Precision-Recall curves) validation metrics. We sorted the heptamers according to their respective detection score, and 60 examples from each extremity (60 binders among the 100 best-scored and 60 non-binders among the 100 worst-scored) were randomly selected to compose a balanced validation set of 120 examples. This set was referred as 'internal evaluation set'. Another independent validation set was collected from a previous study in which it was employed to validate predictions of binding to BiP, a Hsp70 found in endoplasmic reticulum, structurally similar to DnaK. This set was referred to as Collected Dataset (CD) (Schneider *et al.*, 2016).

2.4 Algorithm rationale

We employed a modified version of the Stabilized Matrix Method (SMM) to build a model to predict binding to DnaK (Peters and Sette, 2005). Because overfitting was not diagnosed, regularization was not considered. So, for a seven-length vector of amino acids *res*, predictions in a 20x7 matrix *Mat* occur by summing the score of residue *res* for position *i*. A constant offset was added to this product, as described in Equation 2:

$$PredictionPSSM = \sum_{i=1}^{7} Mat(res, i) + Offset \tag{2}$$

Training was achieved from minimization of Equation 3 (ϕ) that computes the sum of squared errors of matrix predictions for each heptamer *j* in training set:

$$\phi = \sum_{j=1}^{n} (Observed - Predicted)^2 \tag{3}$$

This program and all scripts were implemented in Python 2.7, and minimization was accomplished using Nelder-Mead algorithm from 'minimize' function in scipy.optim python's library. To evaluate the effect of removing the quantitative information from the training, instead of considering the observed value for each peptide, we set to 0 the heptamers with threshold equal or lower than 3.0, while those with threshold above 3.0 were set to 1. Finally, we evaluated the effect of a Position-independent Scoring Matrix (PISM) *Mat*, which was achieved by minimizing Equation 3 with predictions from Equation 4:

$$PredictionPISM = \sum_{i=1}^{7} Mat(res) + Offset \tag{4}$$

In the end, we constructed six models: a 20x7 PSSM, a 20x1 PISM and a 20x7 PISM—each one based on either quantitative or simulated qualitative values. The 20x7 PISM is an explicitly defined 20x7 matrix that is converted to a 20x1 matrix by summing column values.

2.5 Performance evaluation

We obtained the PSSMs from LIMBO and Rüdiger (modified to accept heptamers) from literature (Van Durme *et al.*, 2009). BiPPred predictions were achieved using BiPPred web server (Schneider *et al.*, 2016). BiPPred analyses were performed using either maximum or minimum predicted score. We performed Receiver Operating Characteristic (ROC) and Precision-Recall (PR) analyses to compare predictive performance, where all possibilities of threshold to discriminate positives from negatives are explored to graphically visualize validation set predictions. In ROC-space, False Positive Rate (FPR, also known as 1—specificity) is plotted in the x-axis and True Positive Rate (TPR, also known as sensitivity or recall) is plotted in the y-axis. Both are calculated as follows:

$$TPR = TP/TP + FN \tag{5}$$

$$FPR = \frac{FP}{TN + FP} \tag{6}$$

where *TP*, *FP*, *TN* and *FN* stand for true positives, false positives, true negatives and false negatives, respectively. For PR-analysis, recall is plotted in the x-axis, while Precision is plotted in the y-axis:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Either ROC-analysis as PR-analysis use outline 'metrics' from Sklearn library to calculate the area under the curve.

We performed the sequential training of classifiers with a randomly selected fraction of the training set to evaluate if overfitting was occurring. More data was added each iteration until the total training set was used, with increments of 25 heptamers. Bias, or mean squared error (MSE) of the training set, and the MSE of the validation set were plotted for every iteration. All plots in this work were performed using python's library matplotlib.

Downloaded from https://academic.oup.com/bioinformatics/article/36/3/735/5555810 by guest on 10 April 2024

3 Results

3.1 *In silico* energy predictions of heptamers binding to DnaK correlates with *in vitro* data

Supplementary Table S1 presents the interaction values obtained from Van Durme peptides. Despite the wide range, the value distribution was essentially the same, with most peptides having a very low detection score that decays to very few good binders (Supplementary Figs S1–S4). Because of the difference between values even in the same membrane we normalized and log converted each experiment, keeping the original ratio. We split the 268 peptides into all 1488 heptamers, where each sub-heptamer inherited its peptide source detection value. To work on the same data distribution as Van Durme *et al.*, we set a cutoff of 3.0 to separate binders (736 heptamers) from non-binders (752 heptamers). Next, we mutated the NRLLLTG peptide *in silico* from a DnaK SBD crystal structure (PDB ID: 1DKX) to correspond to each heptamer of the set of binders using the FoldX force field. We kept only heptamers with the lowest interaction energy from each peptide source (see Section 2). We also allowed more heptamers if their interaction energies were within the limit of 0.5 kcal mol^{-1} to the best interactor of that source peptide. This resulted in the removal of 574 heptamers. We merged the repeated heptamers and DnaK interaction values were averaged. In the end, considering binders and non-binders, there was a total of 892 non-redundant heptamers, each one associated with their respective log-transformed normalized detection value. Although some non-binders were predicted to interact well with DnaK, the true binders were primarily enriched with heptamers with lower interaction energies (the lower the energies, the better the interaction). Interestingly, *in silico* energy predictions based on the FoldX force field were inversely correlated to *in vitro* data obtained from DnaK binding scores (Supplementary Fig. S7).

3.2 Use of a PISM outperforms previously published DnaK binding prediction methods

We ranked the 892 non-redundant heptamers according to their DnaK binding score (0–4) and extracted two independent sets (with 120 peptides) from both extremities for the algorithm's performance validation (see Section 2). The remaining 772 heptamers were separated into a training set. We built two types of matrices: a set of 20x7 PSSMs, similar to Rüdiger and Van Durme prediction methods, and a 20x1 PISM that assign the same score for a given residue, independent of its position. Because the conversion of Van Durme and Rüdiger PSSMs to PISMs did not impair the performance evaluation metrics of auROC and auPR, and also showed a very similar composition (Supplementary Fig. S8), we decided to use a third variation of scoring matrix: a 20x7 PISM, trained as an explicitly defined 20x7 matrix where predictions were obtained from the sum of columns. We presented the results for five independent replicates of each algorithm. Training PSSMs produced predictors with reasonable performance (mean auROC of 0.769 and mean auPR of 0.770).

The performance was strikingly improved using a 20x1 PISM (mean auROC of 0.874 and mean auPR of 0.872). However, training 20x7 PISM generated even more robust models (mean auROC of 0.982 and mean auPR of 0.969). These data are represented in Table 1 and Figure 2A. We also evaluated the influence of removing quantitative information from training by setting a simulated qualitative label. We observed a small improvement in PSSMs predictions (mean auROC of 0.802 and auPR of 0.816). Both 20x1 PISMs and 20x7 PISMs were superior to PSSMs, but the removal of quantitative information did not affect the quality of the generated models; thus, models derived from either quantitative as well the qualitative information can be considered as equivalents (Table 1). Both quantitative and qualitative 20x7 PISMs presented better performance if compared to previously published chaperone binding predictors. These models achieved the auROC values of 0.972 (qualitative) or 0.983 (quantitative) and auPR values of 0.974 (qualitative) or 0.970 (quantitative). Because the highest possible value for performance score is 1, we considered the auROC value of 0.983 obtained from the quantitative 20x7 PISM as an outstanding result, showing an improvement of 5.9% compared to the PSSM from Rüdiger and of 9.8% compared to LIMBO.

3.3 DnaK-trained PISM outperforms a BiP binding predictor in BiP binding validation set

Because ROC curves can be misleading and provide an over-optimistic scenario in balanced datasets, we also validated the performance of the classifiers in an imbalanced set (Saito and Rehmsmeier, 2015). For that, we chosen the best predictors—20x7 PISM, qualitative and quantitative—and selected a previously published dataset (CD, Collected Data from Schneider *et al.*). This set is composed of BiP-related peptides, an endoplasmic chaperone structurally similar do DnaK that belongs to the Hsp70 family (Pobre *et al.*, 2018). In this CD set, 44 peptides are examples of true BiP binders, while other 88 are examples of non-binders. This set of heptamers was originally employed to validate BiPPred, a BiP binding predictor (Schneider *et al.*, 2016). The good performance of BiPPred (maximum and minimum predicted score) in our validation set of DnaK binders and non-binders evidences that the binding profile of both proteins, BiP and DnaK, is highly superimposed (Table 1). Interestingly, the BiPPred performance in the CD decreased in terms of auROC and auPR (Table 2 and Fig. 2B).

Both LIMBO and Rüdiger PSSMs also presented a poorer auROC and auPR performance. Our 20x7 PISM classifier, with qualitative- or quantitative-based information, outperformed all tested predictors in terms of auROC and auPR in both balanced (internal evaluation set) and imbalanced datasets (CD set). The only exception was quantitative-based 20x7 PISM auROC (0.838), which performed equivalently to BiPPred Max (0.839). Surprisingly, with an auROC of 0.859 and auPR of 0.786, qualitative-based 20x7 PISM trained on information derived from DnaK shows an improvement of, respectively, 2 and 15% in comparison with BiPPred Max on a benchmark set for chaperone BiP.

3.4 ChaperISM predicts true positive peptides with better accuracy

In practice, researchers interested in predicting *in silico* client binding to Hsp70 to guide their experiments focus on predictions above the threshold determining positives. Thus, we decided to compare how many correct predictions Rüdiger PSSM, LIMBO and ChaperISM (final 20x7 PISM qualitative or quantitative) obtained when discriminating positives using the recommended threshold for each predictor. From a total of 60 true positives peptides in the DnaK validation set, Rüdiger PSSM and LIMBO correctly predict 36 and 29 peptides, respectively. ChaperISM, on the other hand, correctly predicts 51 (quantitative mode) and 56 (qualitative mode). The same applies to CD set, where only 4 and 3 true positives peptides were correctly predicted by Rüdiger and LIMBO, respectively, in a total of 44 peptides. Meanwhile, ChaperISM was able to identify 17 (quantitative mode) and 31 (qualitative mode) true positive peptides (Table 3).

4 Discussion

Herein, we present a new DnaK binding predictor—ChaperISM—based on information derived from a previously published screening of cellulose membrane-bound peptides. It possesses two prediction modes: (i) qualitative and (ii) quantitative. In qualitative mode, models are trained to inform 1.0 or 0.0 for binders and non-binders, respectively. Quantitative mode predictions are directly based on the normalized DnaK chemiluminescence detection.

We employed the SMM without the regularization parameter as learning algorithm. We implemented this modification after the initial tests revealed that having L2 or L1 regularization did not produce models with superior performance than having no regularization parameter at all. The original method possesses L2 regularization to penalize model complexity and avoid overfitting (Peters and Sette, 2005). The analysis of the learning curves for

Downloaded from https://academic.oup.com/bioinformatics/article/36/3/735/5555810 by guest on 10 April 2024

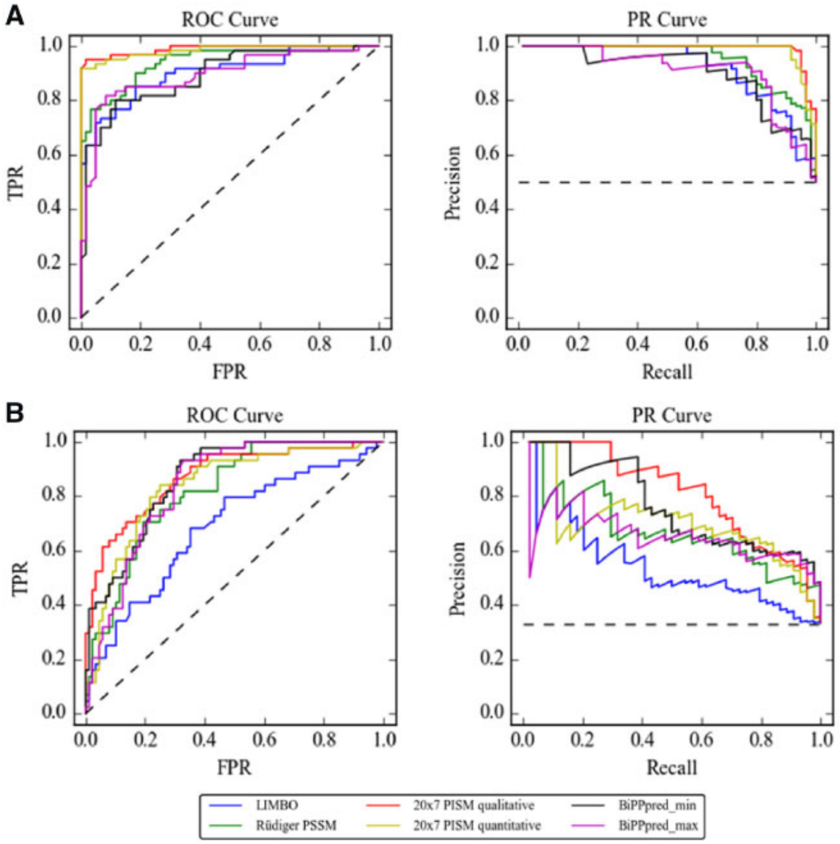

Fig. 2. (**A**) ROC and PR curves for the DnaK validation set defined in this work for LIMBO, Rüdiger, 20x7 qualitative PISM, 20x7 quantitative PISM, BiPPred_min, and BiPPred_max. (**B**) ROC and PR curves for the CD set defined by Schneider *et al.*, 2016 for LIMBO, Rüdiger, 20x7 qualitative PISM, 20x7 quantitative PISM, BiPPred_min, and BiPPred_max. See Table 2 for area under the curves

Table 1. Performance of different chaperone binding predictors for the internal evaluation set

Information type	Algorithm	Best model		Mean value	
		auROC	auPR	auROC	auPR
Quantitative	PSSM (20x7)	0.818	0.825	0.769	0.770
	PISM (20x1)	0.934	0.925	0.874	0.872
	PISM (20x7)	**0.983**	0.970	**0.982**	0.969
Qualitative	PSSM (20x7)	0.870	0.894	0.802	0.816
	PISM (20x1)	0.946	0.956	0.875	0.883
	PISM (20x7)	0.972	**0.974**	0.972	**0.974**
—	Rüdiger	0.924	0.933	NA	NA
—	Limbo	0.885	0.905	NA	NA
—	BipPred Max Score	0.894	0.892	NA	NA
—	BipPred Min Score	0.892	0.885	NA	NA

Note: Results are shown for five independent runs. Higher values are depicted in bold.

quantitative PSSM, 20x1 PISM and 20x7 PISM reveals overfitting does not occur in any of three cases, confirming the regularization parameter was not necessary for this situation (Supplementary Fig. S9). The performance of PSSMs or 20x1 PISMs considerably fluctuates as more data is added to the training set, indicating underfitting. However, 20x7 PISMs regularly show a better performance each time more data is added to the training set. Thus neither under- or overfitting are diagnosed. This reveals that training a 20x1 PISM is not equivalent to training a 20x7 PISM converted to a 20x1 matrix by summing up columns.

A unique feature of ChaperISM is the use of PISMs rather than PSSMs, which means that an amino acid is scored independently of its position. Converting PSSM from Rüdiger or LIMBO to a PISM does not impact its performance; therefore, this simplification is acceptable (Supplementary Fig. S7). The best predictive performance obtained with PISM, which is more straightforward than a PSSM, may be counterintuitive at first. To investigate that, we performed two *in silico* experiments using a true binder (NRLLLTG) and a true non-binder (EKIDNEE). We computed the interaction energy for all possible permutations of both heptamers. If the amino acids compositions were sufficient to predict DnaK binding (a PISM could capture that), we would expect the interaction energies for all permutations to be either the same or quite similar. However, this was not the case, as the energy variation was high (Supplementary Figs S16 and S17). Thus, we further analyzed sub-permutations, fixing pairs of adjacent residues (e.g. **NR**-LLLTG, **NR**-LLTGL, **NR**-TGLLL, . . .), spanning all heptamer positions. This analysis revealed that the energy variation is greatly reduced when the fixed residues are on pockets 4 and 5 (Supplementary Figs S18 and S19). The neighboring pockets (positions 1, 2, 3, 6 and 7) show a 'position-independent' behavior, which reflects the superior performance of a PISM. In summary, our results indicate that a PISM can be more suitable than a PSSM to describe promiscuous protein-protein interaction profiles.

To ensure performance improvement was due to modifications in the learning algorithm, we took the effort to work on the same data distribution as Van Durme *et al.* Thus, we performed a similar step of energy filtering with FoldX force field. Even removing this filter, though, models still show a better performance than Rüdiger or LIMBO PSSMs when validations were performed on DnaK validation set. The same models, however, showed a worsened predictive performance when validations were performed on CD set (see Supplementary Data). This reveals the increased performance in CD

Downloaded from https://academic.oup.com/bioinformatics/article/36/3/735/5555810 by guest on 10 April 2024

Table 2. Performance of different chaperone binding predictors for the internal evaluation set and the CD set

Information Type	Algorithm	Internal evaluation set		CD set	
		auROC	auPR	auROC	auPR
Quantitative	PISM (20x7)	**0.983**	0.970	0.838	0.678
Qualitative	PISM (20x7)	0.972	**0.974**	**0.859**	**0.786**
—	Rüdiger	0.924	0.933	0.806	0.648
—	Limbo	0.885	0.905	0.673	0.519
—	BipPred Max Score	0.894	0.892	0.839	0.636
—	BipPred Min Score	0.892	0.885	0.853	0.737

Note: Higher values are depicted in bold.

Table 3. Correct predictions in DnaK binding validation set and CD set for different chaperone binding predictors considering the recommended cutoff for discrimination

Predictor	Cutoff	Internal evaluation set		CD set	
		TP	TN	TP	TN
Rüdiger	5.0[a]	36/60	60/60	4/44	87/88
LIMBO	11.8[b]	29/60	60/60	3/44	87/88
Quantitative 20x7 PISM	2.7	51/60	60/60	17/44	82/88
Qualitative 20x7 PISM	0.2	56/60	59/60	31/44	75/88

[a] Original predictor outputs energy values, from which the recommended cutoff is −5.0. This modified version, however, outputs a score of DnaK binding probability, rather than energy values.

[b] Selected as default in LIMBO server.

is a direct consequence of the energy filtering step, but since LIMBO presents the worst predictive performance in this set, it does not solely explain the improvement. Removing normalization from preprocessing impacts the generated models in a negatively way (Supplementary Figs S10 and S11). Thus, the overall performance increase in ChaperISM results are a combination of factors such as normalization, energy filtering and training.

DnaK is known to bind to mostly hydrophobic moieties (Clerico *et al.*, 2015). Thus, it could be possible that our final matrices could simply be separating hydrophobic peptides from hydrophilic ones. Nevertheless, when we analyzed the matrices positions, sorted by Kyte and Doolittle hydrophobicity scale, we concluded that this was not the case (Supplementary Figs S21–S24). However, it is important to highlight that the simpler, generalized behavior of a PISM, might indicate heptamers with hydrophilic amino acids in positions 4 and 5 as false positives, which could disfavor or disrupt the binding.

Considering the default threshold to discriminate true positives of each predictor, ChaperISM is the most accurate of them. Concerning BiP binders from CD set, peptides were collected if they could stimulate BiP's ATPase activity with a stimulation factor over 1.5 (Schneider *et al.*, 2016). Many peptides that meet this criterion are exclusively predicted by ChaperISM, such as antibody heavy chain derived heptamers VAFDIWG (from VH domain), HTFPAVL (from CH1 domain), SVFPLAP (from CH1 domain) and WDFAWPW (from CH3 domain) (Knarr *et al.*, 1995). Also, Rauch and colleagues have found non-canonical interactions between Hsc70 SBD and two of its co-chaperones, Bag1 and Bag3. Using the default threshold of LIMBO, the authors were able to identify one putative binding site in the BAG domain of Bag1 and Bag3, but only one putative binding site outside this region for Bag3. ChaperISM, however, indicates several client-like binding regions outside the BAG domain, which are more likely to be the regions involved in the non-canonical interaction (Rauch *et al.*, 2016). These examples highlight the biological relevance of ChaperISM in the context of BiP and Hsc70 binding prediction.

Regarding our working hypothesis, we were able to significantly improve DnaK binding predictions using both algorithm versions, quantitative and qualitative information. The fact that qualitative-based ChaperISM performs better in CD than quantitative-based ChaperISM may appear contradictory. It is nonetheless important to note, that the learning algorithm employed consists in minimizing the observed error, which is quite different than a sequence alignment derived profile from two distinct classes of sequences. Both prediction modes are, thus, valid tools to predict DnaK binding. Still, the more accurate predictions for BiP binding indicates that general binding rules for the Hsp70 family are captured by SMM. As our proposed method outperforms all tested chaperone binding predictors, including DnaK and BiP prediction tools, we conclude it can be used to predict substrate binding other proteins of Hsp70 family.

Funding

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Financiadora de Inovação e Pesquisa (FINEP). The study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

Conflict of Interest: none declared.

References

Antes,I. *et al.* (2006) DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations. *Bioinformatics*, **22**, e16–24.

Clerico,E.M. *et al.* (2015) How hsp70 molecular machines interact with their substrates to mediate diverse physiological functions. *J. Mol. Biol.*, **427**, 1575–1588.

El-Kasaby,A. *et al.* (2014) A cytosolic relay of heat shock proteins HSP70-1A and HSP90β monitors the folding trajectory of the serotonin transporter. *J. Biol. Chem.*, **289**, 28987–29000.

Goloubinoff,P. (2017) Editorial: the HSP70 molecular chaperone machines. *Front. Mol. Biosci.*, **4**, 1.

Goswami,A.M. (2015) Structural modeling and in silico analysis of non-synonymous single nucleotide polymorphisms of human 3β-hydroxysteroid dehydrogenase type 2. *Meta Gene*, **5**, 162–172.

Gowda,N.K.C. *et al.* (2018) Nucleotide exchange factors Fes1 and HspBP1 mimic substrate to release misfolded proteins from Hsp70. *Nat. Struct. Mol. Biol.*, **25**, 83–89.

Halder,U.C. *et al.* (2011) Cell death regulation during influenza A virus infection by matrix (M1) protein: a model of viral control over the cellular survival pathway. *Cell Death Dis.*, **2**, e197.

Knarr,G. *et al.* (1995) BiP binding sequences in antibodies. *J. Biol. Chem.*, **270**, 27589–27594.

Lee,J.H. *et al.* (2015) Heterogeneous binding of the SH3 client protein to the DnaK molecular chaperone. *Proc. Natl. Acad. Sci. USA*, **112**, E4206–15.

Mayer,M.P. and Bukau,B. (2005) Hsp70 chaperones: cellular functions and molecular mechanism. *Cell Mol. Life Sci.*, **62**, 670–684.

Mayer,M.P. and Kityk,R. (2015) Insights into the molecular mechanism of allostery in Hsp70s. *Front. Mol. Biosci.*, **2**, 58.

Moreira,L.G. *et al.* (2013) Structural and functional analysis of human SOD1 in amyotrophic lateral sclerosis. *PLoS One*, **8**, e81979.

Panda,R. and Suresh,P.K. (2014) Computational identification and analysis of functional polymorphisms involved in the activation and detoxification genes implicated in endometriosis. *Gene*, **542**, 89–97.

Peters,B. *et al.* (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.*, **171**, 1741–1749.

Peters,B. and Sette,A. (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, **6**, 132.

Provost,F. *et al.* (1998) The Case Against Accuracy Estimation for Comparing Induction Algorithms. In: *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 445–453.

Pobre,K.F.R. *et al.* (2018) The endoplasmic reticulum (ER) chaperone BiP is a master regulator of ER functions: getting by with a little help from ERdj friends. *J. Biol. Chem.*, **294**, 2098–2108.

Downloaded from https://academic.oup.com/bioinformatics/article/36/3/735/5555810 by guest on 10 April 2024

Rauch,J.N. *et al.* (2016) Non-canonical interactions between Heat Shock Cognate Protein 70 (Hsc70) and Bcl2-associated Anthanogene (BAG) co-chaperones are important for client release. *J. Biol. Chem.*, **291**, 19848–19857.

Rosam,M. *et al.* (2018) Bap (Sil1) regulates the molecular chaperone BiP by coupling release of nucleotide and substrate. *Nat. Struct. Mol. Biol.*, **25**, 90–100.

Rüdiger,S. *et al.* (1997) Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J.*, **16**, 1501–1507.

Saito,T. and Rehmsmeier,M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, **10**, e0118432.

Schneider,M. *et al.* (2016) BiPPred: combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. *Proteins*, **84**, 1390–1407.

Schymkowitz,J. *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–8.

Solayman,M. *et al.* (2017) In silico analysis of nonsynonymous single nucleotide polymorphisms of the human adiponectin receptor 2 (ADIPOR2) gene. *Comput. Biol. Chem.*, **68**, 175–185.

Tenzer,S. *et al.* (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol. Life Sci.*, **62**, 1025–1037.

Van Durme,J. *et al.* (2009) Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. *PLoS Comput. Biol.*, **5**, e1000475.

Zhu,X. *et al.* (1996) Structural analysis of substrate binding by the molecular chaperone DnaK. *Science*, **272**, 1606–1614.

Downloaded from https://academic.oup.com/bioinformatics/article/36/3/735/5555810 by guest on 10 April 2024