Available online at www.sciencedirect.com
ScienceDirect
Sensors and Actuators B 125 (2007) 149–157
SENSORS AND ACTUATORS B CHEMICAL
www.elsevier.com/locate/snb

Using support vector machine regression to model the retention of peptides in immobilized metal-affinity chromatography

B.G. Kermani *, I. Kozlov, P. Melnyk, C. Zhao, J. Hachmann, D. Barker, M. Lebl

Illumina, Inc., 9885 Towne Centre Drive, San Diego, CA 92122, United States

Received 20 October 2006; received in revised form 30 January 2007; accepted 1 February 2007
Available online 11 February 2007

Abstract

Retention of histidine-containing peptides in immobilized metal-affinity chromatography (IMAC) has been studied using several hundred model peptides. Retention in a Nickel column is primarily driven by the number of histidine residues; however, the amino acid composition of the peptide also plays a significant role. A regression model based on support vector machines was used to learn and subsequently predict the relationship between the amino acid composition and the retention time on a Nickel column. The model was predominantly governed by the count of the histidine residues, and the isoelectric point of the peptide.

Keywords: Support vector machines; Regression; Peptide; Metal-affinity chromatography; Retention time

1. Introduction

Immobilized metal-affinity chromatography (IMAC) is an important tool for purification of proteins containing residues that form metal complexes, e.g., histidine, cysteine and tryptophan [1–4]. The poly-histidine tag is extremely useful in molecular biology where it serves to facilitate isolation of genetically engineered proteins from complex mixtures and can be used for targeted immobilization of these proteins [4,5].

We are developing a high-throughput method for the purification of peptide-oligonucleotide conjugates. One of our strategies is to place three histidines (His3) at the amino-terminus of the peptide, and three histidines at the 5′ end of the oligonucleotide. When joined together, the six histidines (His6) should form a tag that can be bound to a Nickel–Sepharose affinity column. After washing away unreacted components, the purified peptide–oligonucleotide conjugate can be eluted with a gradient of increasing imidazole concentration. For this strategy to be successful, the concentration of imidazole that elutes His3 must be significantly less than the concentration that elutes His6. In addition, the elution concentration for His6 should be relatively insensitive to the identity of amino acids surrounding the His6 tag. For brevity, throughout this manuscript, we use the term "imidazole concentration" in lieu of "the concentration of imidazole at which elution occurs."

Surprisingly, there is little information available in the literature on the relative affinity for Nickel–Sepharose, in the presence of imidazole gradients, of polyhistidine-containing peptides and the influence of surrounding amino acids. Consequently, we synthesized an array of model compounds containing different numbers of histidines in various sequential arrangements and in combination with various other amino acids. Here we present comprehensive information that should facilitate the design and purification of engineered peptides and proteins.

By training a model using a subset of the peptides (with measured imidazole concentration required for elution from a Nickel–Sepharose affinity column) we were able to predict the imidazole concentration needed for elution from the column for a large group of peptides. Table 1 shows the amino acid sequence for the peptides of this experiment, along with their measured imidazole concentrations. Fig. 1 shows the dependency of imidazole concentration on several parameters, including the number of histidines and the isoelectric point. As can be seen from Fig. 1a, there is a relationship between the number of histidines (*n*His, *n*H) and the imidazole concentration; however,

* Corresponding author. Tel.: +1 408 730 5700x513.
E-mail address: bkermani@completegenomics.com (B.G. Kermani).

0925-4005/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.snb.2007.02.004

Table 1
Peptides and imidazole concentration

Peptide	Im concentration (M)
GAGAHGAGAGY	0.08
GAGHHDHHGAY	0.09
GAGHHEHHGAY	0.09
HGAGAGAGAGY	0.09
HHAGAGAGAGY	0.09
HRIFLAGDEDY	0.09
GAGAHHGAGA	0.1
GAGAHHGAGA	0.1
GAGAHHGAGAY	0.1
GAGAHHGAGAY	0.1
GAGAHWGAGAY	0.1
GAGRRWGAGAY	0.1
GAHGAGAHGAY	0.1
GAHGAGHAGAY	0.1
GAHGAHGAGAY	0.1
HGAGAGAGARY	0.1
HRIFLAGDKDY	0.1
GAGHAHGAGAY	0.11
HRAGAGAGAGY	0.11
GAGAWWGAGAY	0.12
GAHHEEHHGAY	0.12
GAHHGAGAHGY	0.12
GAHHGAGHAGY	0.12
GAHHGAHGAGY	0.12
HHHGAGAGAGY	0.12
AHHSGASGASGASGHHY	0.13
ASGAHHSGASGASGHHY	0.13
ASGASGAHHSGASGHHY	0.13
ASGASGASGAHHSGHHY	0.13
ASGASGASHHGASGHHY	0.13
ASGASGHHASGASGHHY	0.13
ASGASHHGASGASGHHY	0.13
ASGHHASGASGASGHHY	0.13
ASHHGASGASGASGHHY	0.13
EEEHHHHEEEY	0.13
EEEHHHHEEEY	0.13
EEHHHHEEEY	0.13
ESEHHHHESEY	0.13
GAHHEGHHGAY	0.13
GAHHGHAGAGY	0.13
HHGAGAGAGRY	0.13
RRHHGGHHEEY	0.13
ASGASGASGASHHGHHY	0.14
ASGASGASGHHASGHHY	0.14
EEHHHHEEKY	0.14
ESHHRSHHESY	0.14
GAGAHHHGAG	0.14
GAGAHHHGAG	0.14
GAGAHHHGAGY	0.14
GAGAHHHGAGY	0.14
GAGAHWWGAGY	0.14
GAGARWWGAGY	0.14
GAGHHWGAGAY	0.14
GAHHIIHHGAY	0.14
GAHHLLHHGAY	0.14
GAQHHAHHQAY	0.14
HHAGAGAGRRY	0.14
HHASGASGASGASGHHY	0.14
GAHHERHHGAY	0.15
GAHHGAGHHAY	0.15
GAHHGAHHGAY	0.15
GAHHGAHHGAY	0.15
GAHHGGHHGAY	0.15
GAHHIGHHGAY	0.15

Table 1 (*Continued*)

Peptide	Im concentration (M)
GAHHLGHHGAY	0.15
GAHHSGHHGAY	0.15
HASGAHSGASHGASGHY	0.15
HASGHASGHASGHASGY	0.15
HASHGAHSGHASGASGY	0.15
HHASGASGASGASGHHY	0.15
HRHGAGAGAGY	0.15
KEHHHHEEEY	0.15
REEHHHHEEEY	0.15
EEKHHHHKEEY	0.16
GAGHHAHHGAY	0.16
GAGHHIHHGAY	0.16
GAGHHLHHGAY	0.16
GAGHHMHHGAY	0.16
GAGHHPHHGAY	0.16
GAGHHQHHGAY	0.16
GAHHFGHHGAY	0.16
GAHHGHHAGAY	0.16
GANHHAHHNAY	0.16
GHHAGAGHHAY	0.16
KDHHHHDDDY	0.16
AAHHHHAADY	0.17
DDDHHHHDDDY	0.17
DDHHHHDDDY	0.17
FFLHHHHESEY	0.17
GAGHHFHHGAY	0.17
GAGHHGHHGAY	0.17
GAGHHKHHGAY	0.17
GAGHHNHHGAY	0.17
GAHHAHHGAGY	0.17
GAHHFFHHGAY	0.17
HAHSHGHASGASGASGY	0.17
HASGASGASGASGHHHY	0.17
HHHHAGAGAGY	0.17
HRHRAGAGAGY	0.17
KKHHHHDDDY	0.17
KKKHHHHEEEY	0.17
KSKHHHHESEY	0.17
REEHHHHEERY	0.17
RSHHESHHRSY	0.17
AAAHHHHAAAY	0.18
AAKHHHHAADY	0.18
AAKHHHHAAEY	0.18
AAKHHHHADAY	0.18
AAKHHHHAEAY	0.18
AAKHHHHDAAY	0.18
AAKHHHHEAAY	0.18
AKAHHHHAADY	0.18
DDHHHHDDKY	0.18
DDKHHHHKDDY	0.18
EEGASGASGASGHHHHY	0.18
EKHHHHKEKY	0.18
FFLHHHHKSKY	0.18
GAGHHRHHGAY	0.18
HASGASGASGASGHHHY	0.18
KAAHHHHAAEY	0.18
KKHHHHEEEY	0.18
PPPHHHHPPPY	0.18
RDDHHHHDDDY	0.18
RREESGASGASGHHHHY	0.18
RREHHHHEEEY	0.18
SSSHHHHSSSY	0.18
AAHHHHAAKY	0.19
AARHHHHADAY	0.19
AARHHHHDAAY	0.19

Table 1 (*Continued*)

Peptide	Im concentration (M)
AKAHHHHAAEY	0.19
ASGASGASGASGHHHHY	0.19
EERHHHHREEY	0.19
GAGAHHHHGA	0.19
GAGAHHHHGAY	0.19
GAGAHHHHGAY	0.19
GAGAHHHHGAY	0.19
GAGAHHHHGAY	0.19
GAHHRRHHGAY	0.19
HHHGAGAGRRY	0.19
HHHHHSGESPSYFTA	0.19
ILVHHHHVLIY	0.19
KEHHHHEKKY	0.19
KKHHHHDDKY	0.19
KKHHHHDKDY	0.19
KKHHHHKDDY	0.19
RAAHHHHAADY	0.19
RDDHHHHDDRY	0.19
RSRHHHHESEY	0.19
AAKHHHHAAKY	0.2
AAKHHHHAKAY	0.2
AAKHHHHKAAY	0.2
AARHHHHAADY	0.2
AARHHHHAAEY	0.2
AARHHHHAEAY	0.2
AARHHHHEAAY	0.2
ARAHHHHAADY	0.2
ASGASGASGASGHHHHY	0.2
GAGHHWHHGAY	0.2
GGGHHHHGGGY	0.2
HHHHHSGEAPSYFTA	0.2
HHHHHSGESPWNFTA	0.2
HHHHHSGESPWNLTA	0.2
HHHHHSGRSPSYLTA	0.2
KAAHHHHAAKY	0.2
KKHHHHEEKY	0.2
KKHHHHEKEY	0.2
KKHHHHKEEY	0.2
RAAHHHHAAEY	0.2
RRDHHHHDDDY	0.2
RRGASGASGASGHHHHY	0.2
RRRHHHHDDDY	0.2
RRRHHHHEEEY	0.2
AAKHHHHAAKY	0.21
AAKHHHHAKAY	0.21
AAKHHHHKAAY	0.21
AKAHHHHAAKY	0.21
ARAHHHHAAEY	0.21
DDRHHHHRDDY	0.21
DKHHHHKDKY	0.21
GAGHHWWGAGY	0.21
HHHAGAGRRRY	0.21
HHHHHGAGAGY	0.21
HHHHHSGEAPSNFTA	0.21
HHHHHSGEAPWNLTA	0.21
HHHHHSGESLSNFTA	0.21
HHHHHSGESLSYFTA	0.21
HHHHHSGESLWNLTA	0.21
HHHHHSGESPRYFTA	0.21
HHHHHSGESPSNFTA	0.21
HHHHHSGESPWRLTA	0.21
HHHHHSGRSLSYLTA	0.21
KDHHHHDKKY	0.21
AKAHHHHAAKY	0.22
DKKHHHHKKDY	0.22

Table 1 (*Continued*)

Peptide	Im concentration (M)
GARRWWRRGAY	0.22
HHHHHSGEALSNFTA	0.22
HHHHHSGEAPRNLTA	0.22
HHHHHSGEAPWNFTA	0.22
HHHHHSGESLWNFTA	0.22
HHHHHSGESPRNFTA	0.22
HHHHHSGRALSYLTA	0.22
HHHHHSGRAPSNLTA	0.22
HHHHHSGRAPSYFTA	0.22
HHHHHSGRAPWNLTA	0.22
HHHHHSGRSPSYFTA	0.22
HHHHHSGRSPWNLTA	0.22
HRHRHGAGAGY	0.22
KKHHHHDKKY	0.22
RRHHGGHHRRY	0.22
RSHHRSHHRSY	0.22
AARHHHHAARY	0.23
EKKHHHHKKEY	0.23
HHHHHSGEALSYFTA	0.23
HHHHHSGEALWNLTA	0.23
HHHHHSGEAPRNFTA	0.23
HHHHHSGEAPRYFTA	0.23
HHHHHSGEAPWRLTA	0.23
HHHHHSGESLRNFTA	0.23
HHHHHSGESLRYFTA	0.23
HHHHHSGESLWRLTA	0.23
HHHHHSGESPWYFTA	0.23
HHHHHSGRAPSNFTA	0.23
HHHHHSGRSPRYLTA	0.23
HHHHHSGRSPSNFTA	0.23
KKHHHHEKKY	0.23
KSKHHHHKSKY	0.23
RRRHHHHEEEY	0.23
AARHHHHARAY	0.24
AARHHHHRAAY	0.24
ARAHHHHAARY	0.24
GAGHHHHHGA	0.24
GAGHHHHHGAY	0.24
HHHHHHAGAGY	0.24
HHHHHSGEALWNFTA	0.24
HHHHHSGEAPRRLTA	0.24
HHHHHSGRALRYLTA	0.24
HHHHHSGRALSYFTA	0.24
HHHHHSGRAPRYFTA	0.24
HHHHHSGRAPWNFTA	0.24
HHHHHSGRSLRYLTA	0.24
HHHHHSGRSPSRLTA	0.24
HHHHHSGRSPWNFTA	0.24
HRHRHRAGAGY	0.24
RAAHHHHAARY	0.24
AARHHHHAARY	0.25
AARHHHHARAY	0.25
AARHHHHRAAY	0.25
HHHHHSGEALRNFTA	0.25
HHHHHSGEALRYFTA	0.25
HHHHHSGEALWRLTA	0.25
HHHHHSGEAPWYFTA	0.25
HHHHHSGRAPRNFTA	0.25
HHHHHSGRAPSRLTA	0.25
HHHHHSGRAPWRLTA	0.25
HHHHHSGRSLRNLTA	0.25
HHHHHSGRSLRYFTA	0.25
HHHHHSGRSLWNFTA	0.25
HHHHHSGRSPRNFTA	0.25
HHHHHSGRSPSRFTA	0.25

Table 1 (*Continued*)

Peptide	Im concentration (M)
HHHHHSGRSPWRLTA	0.25
KKKHHHHKKKY	0.25
RAAHHHHAARY	0.25
RERHHHHRERY	0.25
RRRHHHHDDRY	0.25
RRRHHHHRDDY	0.25
ARAHHHHAARY	0.26
FFLHHHHRSRY	0.26
HHHAGAGHHHY	0.26
HHHAGAHHHGY	0.26
HHHAGHHHAGY	0.26
HHHAHHHGAGY	0.26
HHHHHSGESLWYFTA	0.26
HHHHHSGRALRYFTA	0.26
HHHHHSGRSLRNFTA	0.26
HHHHHSGRSLSRFTA	0.26
HHHHHSGRSLWRLTA	0.26
HHHHHSGRSPRRLTA	0.26
HHHHHSGRSPWYLTA	0.26
RREHHHHERRY	0.26
RRRHHHHDRDY	0.26
FFLHHHHLFFY	0.27
HHHGHHHGAGY	0.27
HHHHHSGRALWRLTA	0.27
RRRHHHHEERY	0.27
RRRHHHHEREY	0.27
RRRHHHHREEY	0.27
GAHHHHHHGA	0.28
GAHHHHHHGAY	0.28
HHHHHSGEALWYFTA	0.28
HHHHHSGRALSRFTA	0.28
HHHHHSGRAPWYFTA	0.28
HHHHHSGRSLWRFTA	0.28
HHHHHSGRSLWYLTA	0.28
HHHHHSGRSPRRFTA	0.28
HHHHHSGRSPWYFTA	0.28
RDRHHHHRDRY	0.28
RRDHHHHDRRY	0.28
DRRHHHHRRDY	0.29
HHHHHHHGAGY	0.29
KKKHHHHKKKY	0.29
HHHHHSGRALRRFTA	0.3
RSRHHHHRSRY	0.3
ERRHHHHRREY	0.31
HHHHHSGRALWYFTA	0.31
RRRHHHHDRRY	0.32
GHHHHHHHGAY	0.33
HHHHHHHHAGY	0.33
RRRHHHHERRY	0.34
HHHHHSGRSPWRFTA	0.35
GHHHHHHHHAY	0.37
RRRHHHHKKKY	0.37
RRRHHHHRRRY	0.39
HHHHHHHHHAY	0.41
HHHHHHHHHHY	0.46
RRRHHHHRRRY	0.46

this relationship is not simple. Fig. 1b shows that imidazole concentration is also a function of the composition of the peptide (in this case as measured by the isoelectric point). Fig. 1c shows how imidazole concentration varies as a function of the number of histidines and the number of histidine pairs (*n*HisHis, *n*HH).

2. Methods

The objective of this study is to generate a regression model that takes several parameters from the peptide sequence, and generates an estimate for the imidazole concentration. The necessary (optimal) number of the inputs is unknown (Fig. 2). Thus, the goal is to find the best model and the minimal number of informative inputs that are capable of predicting the imidazole concentration. A regression (as opposed to a classification) model is needed for this predictor, since the output, i.e., imidazole concentration, takes continuous values.

To address the problem, we started with the simplest models, i.e., a linear model. However, a simple linear model was found to be insufficient in capturing all the dependencies of the system. Therefore, we adopted various non-linear regression models for trial. The most simplistic (non-linear) model is a model with one input—number of histidines. The performance of this model was deemed insufficient in view of the alternatives. The next model that we tried was a model with two inputs—*n*His and isoelectric point. A thin-plate spline was used to model this two-input system. While the performance of this system was better than the previous one, it was found that in this model, a linear combination of *n*His and *n*HisHis is more informative than *n*His by itself. Therefore, a third model with the following parameters was made: (1) isoelectric point, (2) *n*His + Beta × *n*HisHis (Fig. 3). The parameter Beta was found to have an optimal value of approximately 2 (Fig. 4). The Beta value of 2 has a physical interpretation, i.e., it matches the hypothesis that in a Nickel column, each Nickel atom interacts with two histidine residues.

While the performance of the thin-plate spline is rather satisfactory, it has the following shortcomings:

(1) It is only capable of handling 2 inputs. Note that in order to handle the *n*HisHis situation, we had to sum it with the *n*His, in a linear fashion, in order to avoid having an extra input.
(2) The surface of the thin-plate spline had local modulations that are aggravated by local noise or data sparsity. These modulations are expected and are due to the "local" (as opposed to global) nature of a spline-based model.
(3) Finding the optimal smoothing factor for the thin-plate spline is non-trivial, and requires human supervision.

Due to the above limitations, we opted to use support vector machines. By doing so, we were able to easily extend the model to a larger number of inputs (than 2).

Support vector machines have been in the forefront of the machine learning algorithms and have gained great popularity in the last decade. This popularity is mostly stemmed from the following [6,7]:

(1) SVMs are based on the structural risk minimization, and thus have a built-in mechanism for regularization, i.e., they are robust to over-training.
(2) Soft-margin SVMs can combine the empirical risk minimization with the structural risk minimization, in order to

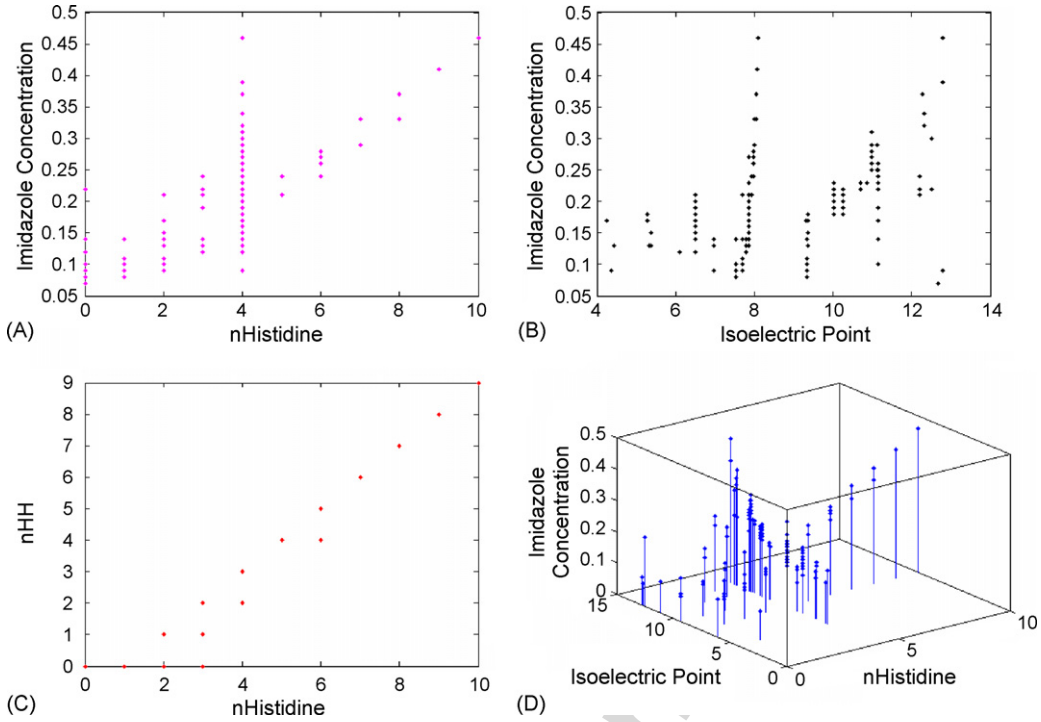

Fig. 1. (a: top-left) relationship between imidazole concentration and *n*His; (b: top-right) relationship between imidazole concentration and isoelectric point; (c: bottom-left) the number of pairs of histidine vs. the number of single histidines in the set; (d: bottom-right) relationship between imidazole concentration as a function of *n*His and isoelectric point.

find an adequate trade-off between the model complexity and the prediction error.

(3) By using non-linear kernels, non-linear systems can be modeled using a linear system in the feature space, and yet without having a need to project the solution into the feature space.

(4) The solution to the underlying optimization problem in SVMs is amenable to classical optimization techniques, namely quadratic programming and linear programming.

The most popular application of SVM is in binary classification [8,9]. However, it has been shown that regression problems can also be modeled using SVM [6,7,10]. The SVM regression (SVM-R) models are based on the epsilon-insensitive error models [6,7]. In an epsilon-insensitive model, the (absolute) error values of epsilon or lower are mapped to zero, and the other error values are mapped either linearly or quadratically. The linear error models are generally more resilient to outliers (with high values). Therefore, in this study, we considered only linear epsilon-sensitive regression SVMs.

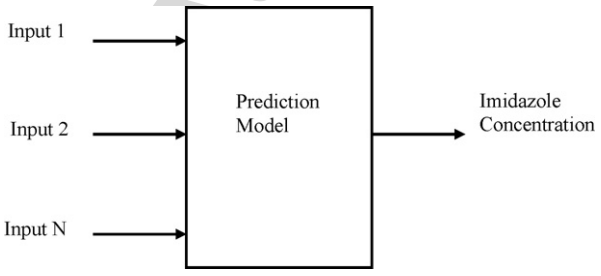

Fig. 2. The general diagram of the regression problem.

2.1. Parameter selection

Table 2 shows the specifications of the SVM-R that was used for this study. The SVM-R models (similar to most binary classifier SVMs) use a regularization parameter C [6] in order to provide a balance between the prediction error and model complexity. The optimal value of C is problem-dependent. Various methods (including cross-validation) are usually used to set this

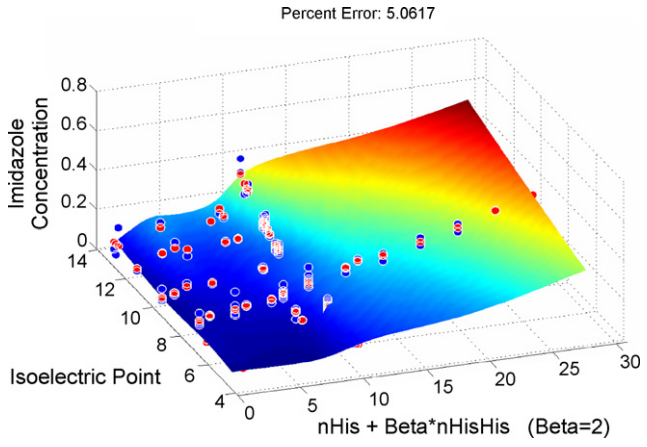

Fig. 3. The prediction results of the thin-plate spline model with the following inputs: (1) isoelectric point, (2) a linear combination of *n*His and *n*HisHis. The points that are on the surface are the predicted. The separation between the actual and the predicted points shows the error in prediction.

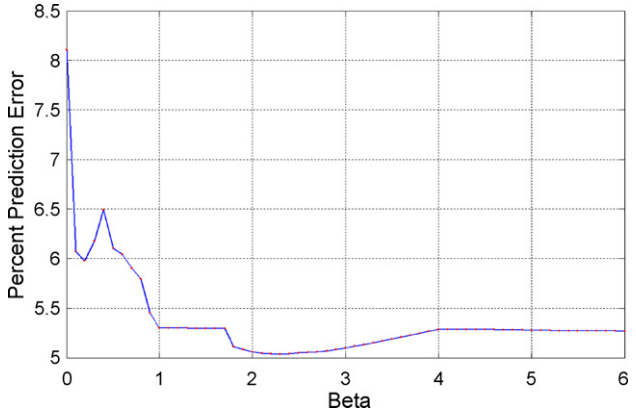

Fig. 4. The error in the prediction of imidazole concentration while using the above thin-plate spline model, as a function of the parameter Beta. The optimal value of Beta appears to be close to 2.

value. In this study, based on several trial-and-errors, the value of 10 was found to be appropriate for the parameter C, although the results were not too sensitive to the specific choice of this parameter.

In order to be able to define the parameter epsilon, we defined an alternate error model. In this model, first for each experiment, the correlation coefficient of the target values and the output values was found. Then, the error was defined as one minus this correlation coefficient. An error which was based on correlation coefficient had a more tangible physical meaning for our scientists, and thus was preferred over the alternatives, e.g., mean-squared error, mean absolute deviation. With this definition, the error values of 0.01 or lower were deemed clinically insignificant, and thus the epsilon value was set to 0.01.

$$\text{Training error} = 1 - \rho(\text{training output, training target}) \quad (1a)$$

$$\text{Test error} = 1 - \rho(\text{test output, test target}) \quad (1b)$$

$$\text{Bootstrap error} = 0.632 \times \text{test error} + 0.368 \times \text{training error} \quad (1c)$$

$$\text{Bootstrap correlation} = 1 - \text{bootstrap error} \quad (1d)$$

$$\text{Bootstrap } R\text{-squared} = (\text{bootstrap correlation})^2 \quad (1e)$$

A polynomial SVM [11] was chosen for this study. An SVM with order 1 was unable to perform as well as an SVM with order 2. However, beyond order 2, the SVM did not significantly improve the performance, and only resulted in longer training sessions. Therefore, the order was fixed at 2.

Table 2
Specifications of the SVM

Type	Polynomial
Order	2
Margin	Soft
C	10
Epsilon	0.01

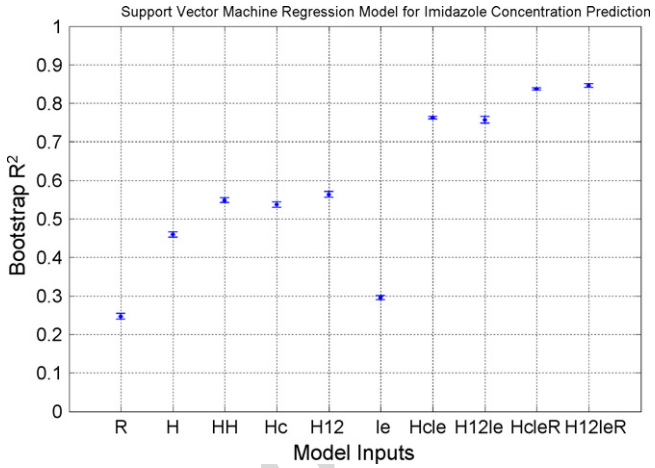

Fig. 5. The Box Plot of the Bootstrap *R*-squared values for the described SVM-R system, with the following inputs: (1) number of argenines (R) (one input); (2) number of histidines (H) (one input); (3) number of HisHis pairs; (4) compound histidine count (Hc), i.e., number of histidines plus twice the number of HisHis pairs (one input); (5) *n*His and *n*HisHis (two inputs); (6) isoelectric point (one input); (7) Hc and isoelectric point (two inputs); (8) *n*His, *n*HisHis and isoelectric point (three inputs); (9) Hc, isoelectric point, and number of argenines (three inputs); (10) *n*His, *n*HisHis, isoelectric point, and the number of argenines.

2.2. *Model validation*

In order to validate the model, a bootstrapping technique [12] was used. In this technique, a total of 100 runs were executed for each scenario. For each run, a subset of the total data was chosen at random (with substitution). This set was called the training set. The difference between this set and the original set was defined as the test set. For each run, the training error and the test error were defined as given in Eqs. (1a) and (1b). Subsequently, a composite 0.632 bootstrap error [13] was computed (Eq. (1c)); and this error was used to find a bootstrap *R*-squared value (Eq. (1e)). This *R*-squared value was used as a representation for the appropriateness of the model.

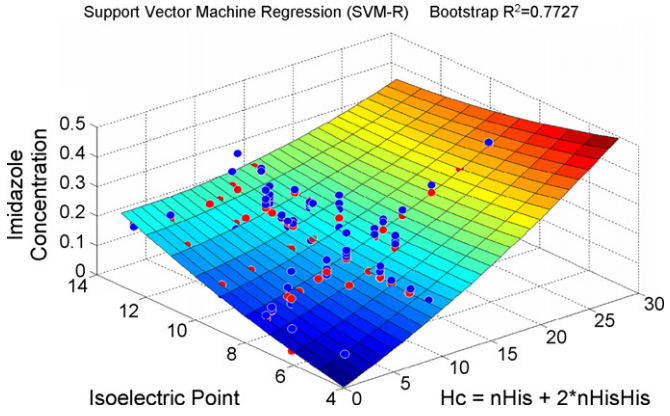

Fig. 6. Surface plot of the imidazole concentration, for a two input systems, with isoelectric point and Hc as the two inputs. The blue points are the expected (experimental) measurements, and the red points are the SVM-predicted points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article).

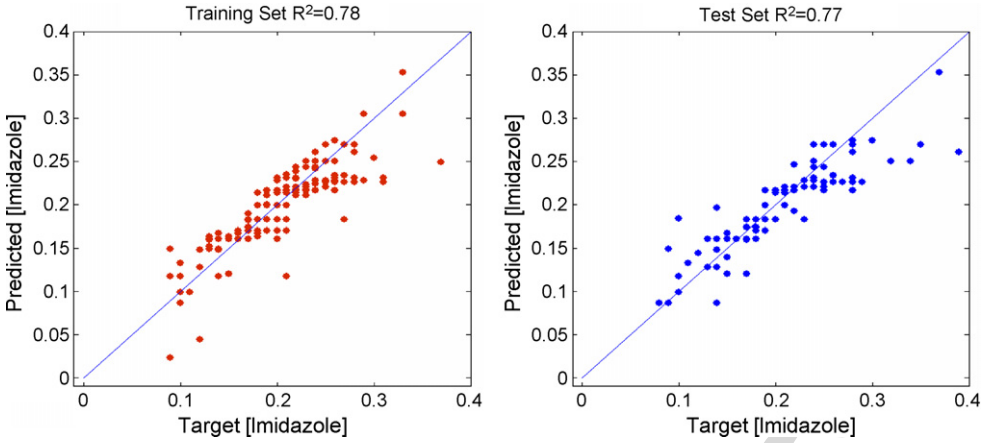

Fig. 7. The results of a single training event. The left plot shows the correlation of the predicted values for imidazole concentration vs. the target (experimentally measured) values.

3. Results

Synthesized model peptides and the concentration of imidazole needed for their elution (extrapolated from the retention time in gradient elution), are given in Table 1. All sequences were synthesized on tyrosine-modified resin to simplify UV detection (280 nm) of peptides eluted with increasing imidazole concentrations. Fig. 5 shows the results of running SVM-R on systems with the following inputs:

(1) number of arginines (R) (one input);
(2) number of histidines (H) (one input);
(3) number of histidine pairs (one input);
(4) compound histidine count (Hc), i.e., number of histidines plus twice the number of histidine pairs (one input);
(5) *n*His and *n*HisHis (two inputs);
(6) isoelectric point (one input);
(7) Hc and isoelectric point (two inputs);
(8) *n*His, *n*HisHis and isoelectric point (three inputs);
(9) Hc, isoelectric point, and number of arginines (three inputs);
(10) *n*His, *n*HisHis, isoelectric point, and the number of arginines (four inputs).

By comparing the bootstrap *R*-squared values from the tested systems, one would make the following observations:

(1) The Nickel column is also responsive to arginine, although it is more sensitive to histidine.
(2) The Nickel column is more sensitive to pairs of histidines, as opposed to single histidines.
(3) Isoelectric point by itself is informative in the predictive system.
(4) The performance of the system improves significantly if the isoelectric point and histidine (or histidine pair) information are supplied simultaneously.
(5) The best performance is achieved in a system with four inputs—number of histidines, number of histidine pairs, isoelectric point, and the number of arginines.
(6) The performance of a three-input system with inputs—Hc, isoelectric point and number of arginines is not too different from the optimal system.

Visualizing systems with more than 2 inputs is difficult. Fig. 6 shows the performance of a 2-input (Hc and isoelectric point) system. The quadratic shape of the surface is due to the specific form of the SVM-R kernel, i.e., second-order polynomial. The proximity of the actual and predicted points is a testimony to the appropriateness of the fit.

Fig. 7 illustrates the differences between the target and predicted values for one of the 100 bootstrap runs. The residuals have patterns, which indicate that the fit has not been optimal. This is partially due to the small sample size, and partially due to the low order (i.e., 2) of the non-linear kernel. However, due to the small sample size, the order of the system was not increased (beyond 2) in order to keep the model parsimonious, and thus more resilient to overtraining.

4. Conclusions

We performed an exhaustive study of the affinity of histidine-rich peptide sequences for a Nickel–Sepharose column, using elution by a gradient of increasing imadazole concentration as a measure of affinity. Retention of histidine-containing peptides depends on the arrangement of histidines as well as the composition of other amino acids in the sequence. The highest performing regression model was obtained with a 4-input system, with the following inputs: (1) number of histidines, (2) number of histidine pairs, (3) number of arginines, and (4) isoelectric point of the peptide sequence. We used the *R*-squared metric for gauging the performance of the learning algorithm. This metric was a function of the agreement between the predicted and the observed values of imidazole concentration in our model. The system rendered a bootstrap *R*-squared value of approximately 0.85, while trained with a second-degree polynomial soft-kernel epsilon-insensitive regression SVM.

Acknowledgement

The project was partially supported by NIH SBIR grant R44 AI056869-02.

Appendix A. Experimental setup

Fmoc amino acids, BOP reagent and Rink resin (0.42 mmol/g) were purchased from Novabiochem (EMD Biosciences, Inc., San Diego, CA, USA). Solvents were from VWR International, Inc. (West Chester, PA, USA). 4-Methylpiperidine was from Sigma–Aldrich (Milwaukee, WI, USA).

Rink resin (300 mg) was added into a mixture of DMF and DCM (10 ml total) to form a non-sedimenting suspension which was distributed into the wells of flat bottom polypropylene microtiterplates (Evergreen Scientific, Los Angeles, CA, USA). The plates were placed into a centrifugal synthesizer [14,15]. An additional 100 μl of DMF was added into the plate wells (beads sedimented) and the plate was centrifuged with a tilt of 6°. A standard protocol was used for the synthesis to remove the Fmoc protecting group; 4-methylpiperidine was used instead piperidine [16]. Individual Fmoc-protected amino acids (0.3 M solution in 0.3 M HOBt in DMF) were pipetted to the wells, and a solution of BOP (0.6 M in DMF) and 1.2 M DIEA in DMF was delivered to each well. Plates were oscillated five times and allowed to rest for 50 s. (During oscillation, the plates are rotated at a speed at which the liquid does not overflow the wall of the well and solid support moves towards the outer side of the well. When the rotation is stopped, liquid returns to the horizontal position and beads distribute at the well bottom, thus mixing the well content.) This procedure was repeated 30 times. The plate was centrifuged and the addition of amino acids and reagents was repeated. After another 30 cycles of oscillation and pausing, the reagents were removed by centrifugation and washing and de-protection was repeated to prepare the plate for the next cycle of synthesis.

At the end of the synthesis the plate was dried in vacuo and 150 μl of mixture K [17] (TFA/thioanisol/water/phenol/EDT: 82.5:5:5:5:2.5 v/v/v/v/v) was added. The plate was capped and shaken on the plate shaker for 3 h. The suspension was transferred by multi-channel pipettor to a filter plate (Orochem Technologies, Lombard, IL, USA). The filtrate was collected in a deep well plate (VWR) and precipitated with ether (600 μl), and after standing in the refrigerator for 2 h, a pellet was formed by centrifugation. The supernatant was removed by a surface suction device and the pellet was re-suspended in ether (600 μl) and centrifuged again. The process of supernatant removal and re-suspension was repeated three times. The product was dried in a Speedvac (ThermoSavant, Waltham, MA, USA), dissolved in 200 μl of H_2O, or 50% dimethylsulfoxide (DMSO)–50% H_2O and samples of 20 μl were taken into 180 μl of water. Twenty microliters were injected onto an HPLC column (Waters, Milford, MA, USA, μBondapak, C18, 10 μ particle, 125 Å pore, 3.9 mm × 150 mm, gradient 0.05% TFA in H_2O to 70% acetonitrile, 0.05% TFA in 15 min, flow rate, 1.5 ml/min, detection by UV at 217 nm). MS was performed at HT-Labs (San Diego, CA, USA).

The peptides were analyzed using HPLC containing a 1 ml volume HisTrap column (Amersham Biosciences, Piscataway, NJ, USA) with the detection at 260 nm. The peptides were injected in 0.02 M sodium phosphate buffer pH 7.4 containing 0.5 M NaCl. The concentration of imidazole was increased linearly from 0 to 0.5 M during 20 min.

References

[1] J. Porath, J. Carlsson, I. Olsson, G. Belfrange, Metal chelate affinity chromatography, a new approach to protein fractionation, Nature 258 (1975) 598–599.
[2] E. Hochuli, H. Dobeli, A. Schacher, New metal chelate adsorbent selective for proteins and peptides containing neighboring histidine residues, J. Chromatogr. 411 (1987) 177–184.
[3] E. Sulkowski, Purification of proteins by IMAC, Trends Biotechnol. 3 (1985) 1–7.
[4] M.C. Smith, T.C. Furman, T.D. Ingolia, C. Pidgeon, Chelating peptide-immobilized metal ion affinity chromatography, J. Biol. Chem. 263 (1988) 7211–7215.
[5] E. Hochuli, W. Bannwarth, H. Dobeli, R. Gentz, D. Stuber, Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate absorbent, Bio/Technology 6 (1988) 1321–1325.
[6] N. Cristianini, J. Shawe-Taylor, Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, CB2 2RU, UK, 2000.
[7] B. Scholkopf, A.J. Smola, Learning with Kernels, The MIT Press, 2002.
[8] M. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, M. Ares, D. Haussler, Genetics knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 262–267.
[9] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, T.R. Golub, Multiclass cancer diagnosis using tumor gene expression signatures, Proc. Natl. Acad. Sci. U.S.A. 98 (26) (2001) 15149–15154.
[10] K.P. Bennett, C. Campbell, Support vector machines: hype of hallelujah? SIGKDD Explor. 2 (2000) 1–13.
[11] SVM and Kernel Methods MATLAB Toolbox, Insa de Rouen, http://asi.insa-rouen.fr/~arakotom/toolbox/index.html.
[12] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall/CRC, Boca Raton, FL, USA, 1994.
[13] B. Efron, Estimating the error rate of a prediction rule: improvement on cross-validation, J. Am. Stat. Assoc. 78 (1983) 316–331.
[14] M. Lebl, Centrifugation based automated synthesis technologies, J. Assoc. Lab. Autom. 8 (2003) 30–36.
[15] M. Lebl, New technique for high-throughput synthesis, Bioorg. Med. Chem. Lett. 9 (1999) 1305–1310.
[16] J.P. Hachmann, M. Lebl, Alternative to piperidine in Fmoc solid phase synthesis, J. Comb. Chem. 8 (2006) 149.
[17] D.S. King, C.G. Fields, G.B. Fields, A cleavage method which minimizes side reactions following Fmoc solid phase peptide synthesis, Int. J. Pept. Prot. Res. 36 (1990) 255–266.

Biographies

Dr. Bahram Ghaffarzadeh Kermani received his BS degree (1989) in electronics engineering from Amirkabir University of Technology (Tehran Polytechnique), Tehran, Iran, MS degree (1992) in electrical engineering from North Carolina State University, Raleigh, NC, and PhD (1996) from North Carolina State University, majoring in electrical engineering, and minoring in biomedical engineering (dissertation topic: "On Using Artificial Neural Networks and Genetic Algorithms to Optimize Performance of an Electronic Nose"). During his graduate studies, he participated in multi-disciplinary, multi-institutional research with Duke University Medical Center, and the University of North Carolina at Chapel Hill, Departments of Cardiology and Biomedical Engineering.

He also consulted with Rex Radiation Oncology, and AromaScan, plc. After graduation, Dr. Kermani was employed at Duke University Medical Center as a postdoctoral fellow. Shortly after, he joined the Integrated Circuits division of Lucent Technologies (AT&T Bell Laboratories). In January 2000, he joined the Bioinformatics team of Illumina, Inc., where he led activities for various genomic and proteomics applications. In February 2007, he joined Complete Genomics, Inc. as the Senior Director of Computational Biology. Dr. Kermani has several publications in the areas of electrical engineering, biomedical engineering, and genomics. He holds in excess of 40 patents, issued in U.S., Canada, France, U.K., Germany, and/or Japan. Dr. Kermani is an associate Editor of the IEEE Sensors Journal.

Dr. David L. Barker was vice president and chief scientific officer at Illumina, Inc., in San Diego, California from 2000 to 2006. Dr. Barker served from 1998 to 2000 as vice president and chief science advisor at Amersham Biosciences, now part of General Electric. From 1988 to 1998, Dr. Barker held senior positions, including vice president of Research and Business Development, at Molecular Dynamics, Inc., until the acquisition of Molecular Dynamics by Amersham. He currently serves on the Boards of Directors of Cell Biosciences, Excellin Life Sciences Inc., Microchip Biotechnologies, Inc, and NextBio. In his academic career, Dr. Barker conducted interdisciplinary research in neurobiology as a postdoctoral fellow at Harvard Medical School, assistant professor at the University of Oregon, and associate professor at Oregon State University. Dr. Barker holds a BS with honors in chemistry from the California Institute of Technology and a PhD in biochemistry from Brandeis University.

Michal Lebl obtained his PhD from the Institute of Organic Chemistry and Biochemistry of the Czechoslovak Academy of Sciences in Prague in 1978 (DSc in 992). In 1991 he moved to USA and worked in Selectide Corporation, the first combinatorial chemistry company in the world, in Tucson. In 1993 he formed Spyder Instruments, which in 2000 merged with the new startup biotech company Illumina Inc. Dr. Lebl works at Illumina, Inc. as senior director since then. Dr. Lebl is the recipient of several awards, the most prestigious being the Leonidas Zervas Award of the European Peptide Society and the 2003 Jouan Award for his contributions to laboratory automation and laboratory process improvement. He is a member of the editorial boards of scientific journals, and serves on the scientific advisory boards of several companies.