Bioinformatics, 31(6), 2015, 834–840
doi: 10.1093/bioinformatics/btu768
Advance Access Publication Date: 19 November 2014
Original Paper

Sequence analysis

Evolutionary non-linear modelling for selecting vaccines against antigenically variable viruses

Tameera Rahman[1], Mana Mahapatra[2], Emma Laing[3,]* and Yaochu Jin[1,]*

[1]Department of Computing, University of Surrey, Guildford GU2 7XH, UK, [2]The Pirbright Institute, Pirbright GU24 0NF, UK and [3]Department of Microbial and Cellular Sciences, University of Surrey, Guildford GU2 7XH, UK

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 30, 2014; revised on October 27, 2014; accepted on November 13, 2014

Abstract

Motivation: *In vitro* and *in vivo* selection of vaccines is time consuming, expensive and the selected vaccines may not be able to provide protection against broad-spectrum viruses because of emerging antigenically novel disease strains. A powerful computational model that incorporates these protein/DNA or RNA level fluctuations can effectively predict antigenically variant strains, and can minimize the amount of resources spent on exclusive serological testing of vaccines and make wide spectrum vaccines possible for many diseases. However, *in silico* vaccine prediction remains a grand challenge. To address the challenge, we investigate the use of linear and non-linear regression models to predict the antigenic similarity in foot-and-mouth disease virus strains and in influenza strains, where the structure and parameters of the non-linear model are optimized using an evolutionary algorithm (EA). In addition, we examine two different scoring methods for weighting the type of amino acid substitutions in the linear and non-linear models. We also test our models with some unseen data.

Results: We achieved the best prediction results on three datasets of SAT2 (Foot-and-Mouth disease), two datasets of serotype A (Foot-and-Mouth disease) and two datasets of influenza when the scoring method based on biochemical properties of amino acids is employed in combination with a non-linear regression model. Models based on substitutions in the antigenic areas performed better than those that took the entire exposed viral capsid proteins. A majority of the non-linear regression models optimized with the **EA** performed better than the linear and non-linear models whose parameters are estimated using the least-squares method. In addition, for the best models, optimized non-linear regression models consist of more terms than their linear counterparts, implying a non-linear nature of influences of amino acid substitutions. Our models were also tested on five recently generated FMDV datasets and the best model was able to achieve an 80% agreement rate.

Contact: yaochu.jin@surrey.ac.uk or e.laing@surrey.ac.uk

1 Introduction

Vaccines can be effective in preventing viral diseases and improving the health of millions. However, there are many viral diseases for which vaccine production/deployment is a challenge because of extensive antigenic variability [Antigenic variability is the mechanism by which infectious organisms (bacteria, virus or protozoa) alters its surface protein to evade a host immune system (Daintith, 1990)]. Infection and/or vaccination by an antigenically-variable virus do not necessarily confer protection to subsequent exposure of the disease, e.g. Human Immunodeficiency Virus (HIV), Influenza and Foot and Mouth Disease Virus (FMDV) (Moxon and Siegrist, 2011). Hence, upon an outbreak of an antigenically variable virus, a rapid response to reduce the spread of the virus is needed. In other words, the rapid selection of the most effective vaccine, against the particular strain of virus, is imperative. However, the *in vitro* serological tests that measure the cross reactivity of vaccines to outbreak-strains are time consuming and expensive. An *in silico* predictor that can accurately predict vaccine efficacy would be invaluable.

With the advent of high-throughput sequencing, it is now possible to obtain the sequence of a virus within hours. The computational exploitation of this sequence information for vaccine design/selection is therefore attractive. Yet to date, most computational research on viral vaccines has focussed on developing tools for identifying epitopes (Brusic *et al.*, 1994; De Groot *et al.*, 2002; Meister *et al.*, 1995; Schafer *et al.*, 1998). From our literature search, only three previous studies attempting to predict antigenic variability and therefore vaccine efficacy, for Foot and Mouth Disease and Influenza were identified (Liao *et al.*, 2008; Reeve *et al.*, 2010, 2011). Foot and Mouth Disease and Influenza are both socioeconomically important, highly infectious diseases caused by antigenically variable viruses. As DNA sequences and associated serological test data are available for these viruses, they are obvious candidates for the development of an *in silico* vaccine selection predictor.

Thus far, *in silico* approaches for predicting vaccine efficacy have focussed on the use of linear and logistic regression techniques [Reeve *et al.* (2010) used linear mixed effects models to relate estimated antigenic differences to sequence variation in FMDV and Liao *et al.* (2008) used stepwise and logistic regression to predict antigenic variants in Influenza]. However, as vaccine efficacy relies on the interaction of amino acid residues within or outside an antigenic site (loop) (Lesk, 2001) that may result in unusual (non-linear) patterns in antigenic distance (Lee and Chen, 2004) it is likely that non-linear models, that can handle complex interdependencies between variables, will be better suited. Furthermore, as biological data can be noisy and/or viral strains partially labelled (dataset contains only sequence data but not serological data) (Bandyopadhyay, 2007; Li, 2010), a predictor that can be trained with imperfect data is needed.

Here we present a genetic algorithm (GA) optimized quadratic non-linear regression model that is able to accurately (98% accuracy) predict the vaccine efficacy for FMDV and Influenza A strains taking amino acid sequences as input. We show that our model is an improvement to the approaches previously taken. Furthermore, to examine whether there are differences in the influence of amino acid changes in different capsid proteins on the Virus Neutralization (VN) titre/Haemagglutination inhibition (HI) assay value we compare two approaches for calculating amino acid changes, one reporting the total number of changes across all capsid proteins (three capsid proteins in FMDV and one in influenza), while the other reporting specific-capsid changes. Finally, we study two methods for weighting the importance of amino acid changes, one based on the characteristics and structure of amino acids and the other based on amino acid substitution scores (PAM/BLOSUM).

2 Methods

2.1 Data

Three datasets comprising serological and amino acid sequence data for outbreak-vaccine strain pairs for three different viruses were used to train and cross-validate our models: (1) 22 FMDV SAT2 outbreak-vaccine strain pairs, taken from Reeve *et al.* (2010); (2) 52 outbreak-vaccine strain pairs for FMDV A, taken from Upadhyaya *et al.* (2013) and (3) 54 outbreak-vaccine strain pairs for Influenza A H3N2, taken from Liao *et al.* (2008).

Amino acid sequences for FMDV and Influenza A were either downloaded from GenBank or provided by The Pirbright Institute. The sequences of a particular viral serotype were harmonized to the same length [669 residues for SAT2 comprising viral proteins (VP) 1–3, 655 residues for A similarly comprising (VP) 1–3 and 329 residues for HA1 gene of H3N2] by obtaining a multiple sequence alignment using ClustalW (Thompson *et al.*, 2002). This alignment was used to identify substitutions between the pairs of sequences for which the serological data were available. We have used two different approaches to build the models. In one we have used substitutions only from specific antigenic regions, two regions in VP1 and one in VP2 (Reeve *et al.*, 2010) for FMDV (SAT2 and A) and three regions in H3N2 HA1 (Lee *et al.*, 2007). In the other setup, we have used substitutions in all three capsid proteins (VP1, VP2 and VP3) for SAT2 and A, and the entire HA1 sequence for H3N2.

2.2 Scoring methods

3Scoring methods are used to determine the difference between a pair of aligned amino acid sequences. Using our models we compared three different scoring methods: (1) ‘No weighting’: the basic approach of counting the number of substitutions between a pair of aligned sequences, as previously applied in studies on FMDV (Reeve *et al.*, 2010). (2) ‘Grouping’: a weighting method applied in studies on Influenza A (Liao *et al.*, 2008). The 20 amino acids are divided into seven groups, namely, non-polar aliphatic, non-polar aromatic, polar, positively charged and negatively charged and substitutions between groups are assigned 1 while those within a group are assigned 0.

(3) ‘Matrix Scoring’: Use of the PAM-120 matrix, a matrix scoring the probability with which one amino acid is substituted by another in closely related species (Dayhoff, 1978). The matrix was modified by setting the diagonal values to 0 so that identical amino acids in the two sequences are not scored. The positive values were changed to 0 and the negative values to positive so that frequently occurring substitutions are not scored while the rare substitutions are scored. In addition, all elements in the matrix were normalized between 0 and 0.5. The scores using this approach were calculated by totalling the values of all elements corresponding to substitutions between pairs of sequences.

2.3 Linear and non-linear prediction

Linear and non-linear models were constructed to predict the antigenic variability. The scoring methods described above were used as the continuous dependent variables for prediction using regression techniques.

2.3.1 Linear regression model

Linear regression is the most widely used regression technique. It aims to find the best line that fits the data. When the amino acid changes in all capsid proteins are totalled, the linear model has only one dependent variable x, and the predicted VN titre or HI assay value (y) can be predicted as follows:

$$y = \beta_0 + \beta_1 x. \tag{1}$$

When changes are calculated for specific antigenic regions there are three dependent variables (x_1, x_2, x_3), and the linear regression model for predicting the VN titre or HI assay value has the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \tag{2}$$

where β_0, β_1, β_2 and β_3 are the regression coefficients that need to be estimated. The parameters are estimated by minimizing the mean squared error (MSE) between the experimental and the predicted values (Friedman et al., 2001), which is known as the least squares method (LSM).

2.3.2 Non-linear regression models

In the linear regression model, it is assumed that the VN titre or HI assay value depends linearly on the amino acid changes. However, this may be overly simplistic and not adequate for describing the relationship between the VN titre or HI assay value and the amino acid changes. A non-linear regression model, however, can describe a more complex relationship between the dependent (VN titre or HI assay value) and the independent variables (amino acid changes). Here, we adopted a quadratic (second-order polynomial) non-linear regression model. Similar to that described for linear models, there is one dependent variable if the amino acid changes in the capsid protein(s) are totalled, and three variables if the amino acid changes are counted specifically for each of the antigenic regions (capsid proteins). In the former case, the quadratic model is in the following form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{3}$$

and in the latter case, the quadratic model has the following form:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &+ \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3, \\ &+ \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2 \end{aligned} \tag{4}$$

where $\beta_i, i = 1, 2, \ldots, 9$ are the regression coefficients to be determined. Typically, these parameters can also be estimated using the LSM.

In this study, we not only want to estimate the coefficients in the quadratic regression model, but also to examine if any of the terms in the non-linear model in Equation (4) can be removed to reduce the complexity of the model. This is potentially beneficial due to the small number of experimental data available for determining the coefficients. In the following, we describe in more detail a method for optimizing the structure and coefficients of the quadratic regression model using an evolutionary algorithm (EA), specifically a hybrid GA.

2.3.3 A hybrid GA

EAs were first developed in the early 1960s as models that simulate biological evolutionary processes for parameter optimization (Back *et al.*, 1997), among other purposes. GAs (Goldberg, 1989) and evolution strategies (Schwefel, 1995) are two classes of widely used EAs for optimization.

GAs were introduced in 1975 by Holland (1992). A canonical GA uses binary (or Gray) coding for representing the decision variables to be optimized. Starting from a randomly generated parent population, in which each individual represents a candidate solution, a canonical GA performs crossover and flip mutation to generate new candidate solutions (offspring). The offspring individuals are then evaluated to determine their *fitness*, and in this study, the prediction error of each candidate model. Then, offspring individuals are selected according to their fitness value as parents for the next generation. In this study, the better the fitness value (i.e. the smaller the prediction error), the more likely an offspring individual will be selected. The selection process simulates the principle of *survival of the fittest* in natural evolution. This generations cycle continues until a termination criteria is fulfilled, e.g. when the predefined number of generations have been processed.

Many variants of the canonical GA have been developed by introducing new representation schemes, genetic variations and selection methods to enhance the search performance of the canonical GAs (see e.g. Back *etal.*, 1997). For example, in real-coded GAs (RCGAs), decision variables can be directly represented using real-values (Deb and Agrawal, 1995). Accordingly, new crossover and mutation operators, e.g. simulated binary crossover (SBX) and new mutation operators such as polynomial mutation have been designed for RCGAs (Deb and Agrawal, 1995).

SBX operation was proposed by Deb and Agrawal (1995). Given two parent solutions $x_i^{(1,t)}$ and $x_i^{(2,t)}$, a random number, u_i, between 0 and 1 is generated at first:

$$\beta_{qi} = \begin{cases} (2u_i)^{\frac{1}{\eta_c + 1}} & \text{if } u_i \le 0.5 \\ \left(\frac{1}{2(1-u_i)}\right)^{\frac{1}{\eta_c + 1}} & \text{otherwise,} \end{cases} \tag{5}$$

where η_c is a distribution index. As recommended in Deb and Agrawal (1995), η_c is typically set to 20. Two offspring solutions, $x_i^{(1,t+1)}$ and $x_i^{(2,t+1)}$, are then produced as follows:

$$x_i^{(1,t+1)} = 0.5\left[\left(1 + \beta_{qi}\right)x_i^{(1,t)} + \left(1 - \beta_{qi}\right)x_i^{(2,t)}\right], \tag{6}$$

$$x_i^{(2,t+1)} = 0.5\left[\left(1 - \beta_{qi}\right)x_i^{(1,t)} + \left(1 + \beta_{qi}\right)x_i^{(2,t)}\right], \tag{7}$$

In polynomial mutations, a probability distribution is first defined:

$$P(\beta_{mi}) = 0.5(\eta_m + 1)(1 - |\beta_{mi}|)^{\eta_m}, \tag{8}$$

where η_m is the distribution factor, and β_{mi} is given by

$$\beta_{mi} = \begin{cases} (2u_i)^{\frac{1}{\eta_m + 1}} - 1 & \text{if } u_i \le 0.5 \\ 1 - (2(1-u_i))^{\frac{1}{\eta_m + 1}} & \text{otherwise} \end{cases}. \tag{9}$$

If a mutation in the individual occurs, the parameter value is given as

$$x' = x + (\alpha - \delta)\beta_{mi}, \tag{10}$$

where α and δ are the upper and lower bounds for the mutation values, respectively. η_m is also set to 20 in this study. Additional details of both the SBX and polynomial mutation operations can be found in Deb and Agrawal, (1995).

We use a hybrid GA to optimize both the structure as well as the parameters of quadratic regression model in Equation (4). By structure, we mean here which of the 10 terms in the non-linear model should be kept for prediction. We aim to find a quadratic model of minimum complexity that best predicts the SAT2, A and H3N2 data. Meanwhile, we intend to optimize the coefficients of the existing terms in the non-linear regression model. In the following, we elaborate the details of the hybrid GA used for optimizing the non-linear regression model.

Representation: A binary chromosome and a real-valued chromosome are used to represent the structure and coefficients of the non-linear regression model, respectively. The binary chromosome has 10 bits of a value '0' or '1', where a '1' means that the corresponding term of the model in Equation (4) is kept in the model and a '0' indicates that the corresponding term is deleted. The real-valued chromosome directly encodes the coefficients. For example, the following individual

1	0	0	0	1	0	1	0	0	1
0.1	1.2	−1.5	0.5	2.2	0.3	−0.8	0.1	1.8	3.2

defines the following non-linear regression model:

$$y = 0.1 + 2.2x_1x_2 - 0.8x_2x_3 + 3.2x_3^2. \tag{11}$$

Population: In this study, the population size is set to 100, with each individual representing a candidate non-linear regression model specified by two chromosomes. The initial parent population is generated randomly, the value of the binary chromosome is set to '0' or '1' with a probability of 0.5, respectively, whereas the value of the real-valued chromosome is set to a uniformly distributed random number between $[-0.1, 0.1]$.

Fitness function: The fitness of the individuals is determined by calculating the MSE between the predicted and the experimental VN titre or HI assay values:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^d)^2, \tag{12}$$

where y_i and y_i^d are the VN titre or HI assay value predicted by the model and the experimental value of the *i*th training data pair, respectively, $i = 1, 2, ..., N$, N is the number of training samples.

The selection strategy adopted in this study is an *elitism* strategy. Before selection, the offspring population is combined with the parent population, forming a population of 200 individuals. Then, the best 100 individuals are selected as the parents for the next generation.

Reproduction: The *binary tournament selection* is adopted to select individuals from the parent population to generate offspring, one at a time. At each time, two individuals in the parent population are randomly selected and the fitter is chosen as a parent. This is repeated to choose a second parent individual.

Crossover and mutation are then applied to the two parents to generate two offspring individuals. As each individual has two different types of chromosomes, different crossover and mutation operators are applied to the binary and real-valued chromosomes. For the binary chromosome, uniform crossover and flip mutations are employed. For the real-valued chromosome, SBX and polynomial mutation are used. The probability for crossover and mutation are specified to be 0.9 and 1, respectively. This process repeats until 100 offspring are generated.

Selection: Parent populations are combined with the offspring population before selection. The 200 individuals are sorted based first on the individuals' front number in a decreasing order and then on the crowding distance in an ascending order. Finally, the first 100 individuals are selected as the parents of the next generation. In this study, 500 generations are run before the evolutionary optimization is terminated.

2.4 Model performance comparison

The specificity, sensitivity and agreement rates for pairs of predicted and true VN titre or HI assay values were calculated for all models. For FMDV serotype SAT1 and A r_1 values between 0.3 and 1.0 are indicative of reasonable levels of cross protection, whilst values below 0.3 indicate the need to acquire or develop a new/different vaccine strain (Barnett *et al.*, 2001). Specificity is the ratio of the predicted similar viruses (predicted VN titre or HI assay value >0.3) to the true similar viruses (actual VN titre or HI assay value >0.3) as a percentage. Sensitivity is the ratio of predicted variants (predicted VN titre or HI assay value <0.3) to true variants (actual VN titre or HI assay value <0.3) represented as a percentage. The agreement rate is the ratio of all truly predicted pairs to the number of all virus pairs. In addition to the above performance parameters, the MSEs were also compared for all models.

3 Results and discussions

Figure 1 provides an overview of our approach. Here we assume that the efficacy (cross-reactivity) of a vaccine is accurately measured by a serological test (HI for Influenza or VN for Foot and Mouth disease) and that this can be predicted from the alignment of amino acid sequences for the heterologous virus (e.g. outbreak strain) and homologous virus (e.g. the vaccine strain) pair. Thus, to create our training and validation datasets we use serological test results (r_1/HI assay values) and the sequence data from the alignment files. The serological data results are directly used as the dependent variables for regression analysis while the sequence data from pairwise comparisons are used with three types of scoring methods (refer to Table 1 for details), outputs of which are used as the independent variables for regression analysis.

In the following, we compare the prediction performance of the linear and non-linear regression models, where in one case the amino acid changes on the capsid protein(s) are totalled (herein referred to as 'no loops'), and in the other case, the changes in the three antigenic regions are counted separately (referred to as 'loops'). The agreement rates for the two FMDV serotypes SAT2, A and influenza A are shown in Table 2. The MSEs for the same datasets are presented in Figures 2–4, respectively. From the figures, we can see that better prediction results can be achieved when the amino acid changes in different capsid proteins are treated as different independent variables. The agreement rates for models with immunodominant positions (a total of 39) or 'loops' were at least similar or better than those containing amino acid positions for the capsid protein(s). In addition, the MSEs for the models with antigenic loops are lower than those without loops except the 'scoring model', whose MSEs are almost the same, refer to Figures 2–4. It is noteworthy that the non-linear *grouped model* optimized using the GA obtained the best performance. The results from the three datasets for FMDV SAT 2, two datasets for FMDV A and two datasets for influenza are consistent across serotypes. SAT2 results presented here are for SAT2 sequence RWA/2/01, A sequence TUR02 and influenza Phyllipines/02/82/01.

Even though the capsid protein(s) are directly involved in determining the antigenicity of the virus, from the modelling point of view, it is only the specific antigenic areas (loops) that are most relevant. However, when the changes in capsid(s) are summarized, useful information for prediction purposes may be lost, resulting in poor prediction performance. Indeed, certain positions within the capsid proteins, not included in the antigenic loops, may be essential for improving the prediction quality. Thus, a way to weigh each position to extract the most significant ones will greatly improve the model based on antigenically relevant residues.

When we compare scoring methods, the grouping method performs the best out of the three. The grouping method had agreement rates between 73 and 100% while the matrix scoring method between 26 and 85%. For the non-weighted approach, it was between 62 and 90% across all datasets. The lower rates correspond to models involving residues in the capsid protein(s). Table 2 shows the performance of the weighted and the non-weighted linear and non-linear models.

The grouping method that classifies amino acids based on their biochemical properties distinguishes between amino acid substitutions based on their relevance to the overall protein structure and thus would be expected to enhance the regression models. The 20 amino acids have different properties and not all substitutions are equally important for measuring antigenic variability. It is also known that certain amino acid substitutions are more common than others (Betts and Russell, 2003).

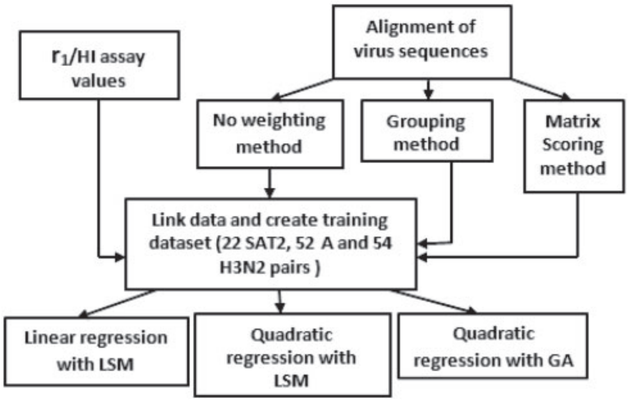
r_1/HI assay values
Alignment of virus sequences
No weighting method
Grouping method
Matrix Scoring method
Link data and create training dataset (22 SAT2, 52 A and 54 H3N2 pairs)
Linear regression with LSM
Quadratic regression with LSM
Quadratic regression with GA

Fig. 1. Flowchart of our study

Table 1. Comparison of predicted and actual r_1 values using our grouping method with the three types of regression models

Challenge Strain	Model	Predicted		Actual	
		St 1	St 2	St 1	St 2
IRN/07/13	NLR with GA	0.17	0.29	0.15	0.25
	NLR with LS	0.11	0.15		
	LR	0.09	0.16		
IRN/35/12	NLR with GA	0.15	0.39	0.18	0.45
	NLR with LS	0.18	0.22		
	LR	0.08	0.18		
TUR/05/12	NLR with GA	0.53	0.25	0.24	0.28
	NLR with LS	0.32	0.44		
	LR	0.15	0.36		
IRN/01/13	NLR with GA	0.20	0.29	0.095	0.221
	NLR with LS	0.30	0.15		
	LR	0.21	0.40		
SAU/23/86	NLR with GA	0.56	0.11	0.1	0.14
	NLR with LS	0.22	0.17		
	LR	0.46	0.02		

Key: St 1-Reference strain 1 (IRQ/24/64), St 2-Reference strain 2 (TUR/20/06).

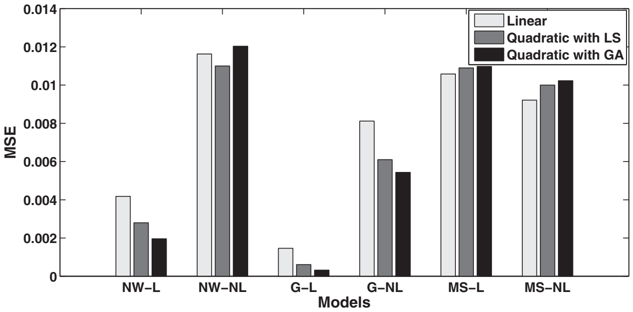

Fig. 2. The MSEs of the linear and non-linear regression models using different weighting methods for antigenic regions and whole capsid proteins for FMDV serotype SAT2. Key: L—with loops, NL—no loops, NW—non-weighted, G—grouped, MS—matrix scored

We also carried out some correlation analysis for the type of amino acid substitutions and antigenic similarity between viruses. We found that some commonly occurring substitutions have no correlation or are slightly positively correlated to antigenic similarity. Pearson's correlation coefficient for T-A/A-T and S-A/A-S substitutions is 0.179 and 0.312, respectively. The correlation obtained was weak due to the limited amount of data available for the analysis. This is contrary to the expectation that the number of amino acid substitutions must be negatively correlated to antigenic similarity. However, for some less frequently occurring substitutions the correlation was seen to be weakly negative. This is explained by the fact that, in nature, substitutions of one amino acid with another of similar properties is permissible, with the overall structure/function of the protein remaining unchanged (Chasman and Adams, 2001). The grouping method exploits this phenomenon by scoring substitutions with similar biochemical properties as 0 and substitutions that may change the structure and/or functions of the resultant proteins as 1. The non-weighted approach that considers all substitutions

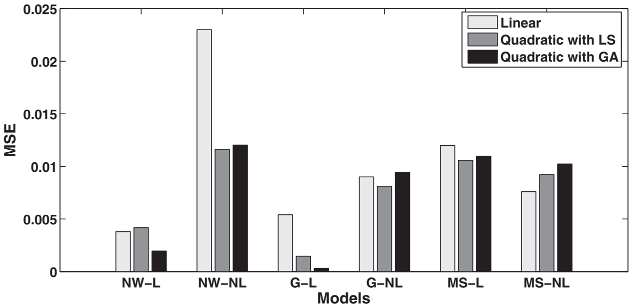

Fig. 3. The MSEs of the linear and non-linear regression models using different weighting methods for antigenic regions and whole capsid proteins for FMDV serotype A. Key: L—with loops, NL—no loops, NW—non-weighted, G—grouped, MS—matrix scored

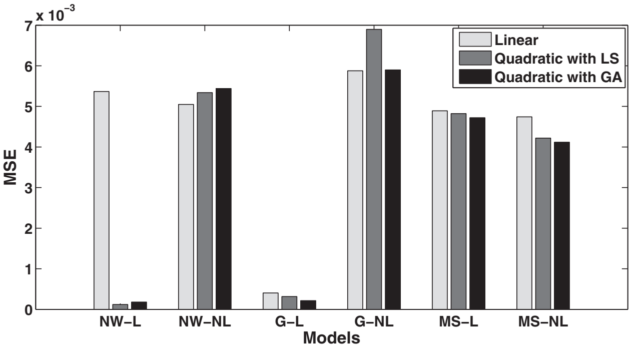

Fig. 4. The MSEs of the linear and non-linear regression models using different weighting methods for antigenic regions and whole capsid proteins for influenza A. Key: L—with loops, NL—no loops, NW—non-weighted, G—grouped, MS—matrix scored

Table 2. Performance results of different linear and non-linear models for FMDV and influenza data

Analysis	Weighting	FMDV SAT2 ($N = 22$)				FMDV A ($N = 52$)				INFLUENZA ($N = 54$)			
		No. of coef (no. sig)	Sn	Sp	Ag rate	No. of coef (no. sig)	Sn	Sp	Ag rate	No. of coef (no. sig)	Sn	Sp	Ag rate
LR	NW-L	3(2)	100	80	62	3(2)	60	70	70	3(1)	50	57	63
	NW-NL	3(3)	0	100	78	0	4	90	65	4(3)	20	76	79
	G-L	**4(3)**	**80**	**100**	**80**	**4(2)**	**100**	**80**	**75**	**4(3)**	**100**	**75**	**75**
	G-NL	2(2)	50	90	77	3(2)	20	80	60	3(2)	25	75	70
	MS-L	4(2)	25	90	80	3(1)	0	100	68	2(2)	33	77	75
	MS-NL	4(3)	25	90	30	4(2)	25	100	55	2(0)	0	11	30
NLR with LS	NW-L	10(6)	100	80	70	10(6)	80	75	84	10(7)	55	50	70
	NW-NL	10(5)	0	100	79	10(5)	0	77	70	10(6)	20	66	73
	G-L	**10(8)**	**100**	**98**	**98**	**10(6)**	**100**	**78**	**80**	**10(7)**	**100**	**75**	**95**
	G-NL	10(6)	50	75	78	10(5)	20	78	68	10(5)	44	75	68
	MS-L	10(6)	0	100	82	10(6)	0	100	75	10(5)	50	50	80
	MS-NL	10(7)	25	90	35	10(6)	25	100	70	10(4)	0	10	35
NLR with GA	NW-L	7(6)	100	80	74	8(7)	85	75	84	5(3)	50	66	75
	NW-NL	2(2)	0	100	80	4(2)	0	75	67	5(3)	20	75	73
	G-L	**6(6)**	**100**	**100**	**100**	**7(6)**	**100**	**100**	**100**	**7(7)**	**100**	**100**	**100**
	G-NL	3(3)	55	81	77	4(2)	0	100	80	5(4)	55	75	70
	MS-L	9(6)	0	100	83	8(7)	0	100	75	9(7)	53	66	85
	MS-NL	6(3)	25	90	37	7(4)	25	100	35	6(3)	0	15	37

Key: Sn-Sensitivity, Sp-Specificity, Agg. rate-Agreement rate, No. of coef(no. sig)—No. of model coefficients (no. of significant coefficients), LR—Linear regression, NLR—non-linear regression, GA—genetic algorithms, LS—least Squares, NL—non-weighted looped, N-NL—non-weighted no loop, G-L—grouped looped, G-NL—grouped no loop, MS-L—matrix scored looped, MS-NL—matrix scored no loop.

regardless of structural relevance includes frequent substitutions that are of little significance to the model; the overall result of this is a poor quality model.

Comparing the performance of the non-linear regression model with LSM and the one optimized with GA, we found from Figures 2–4 and Table 2 that the GA model perform clearly better in 12 out of 18 cases, while the GA model is worse than the LSM model only in three cases. In addition, it is noted that the agreement rates on the FMDV A dataset are slightly worse than those of the other two datasets. This might be attributed to the fact that the spread of the r_1 values in the FMDV SAT2 and INFLUENZA datasets is better than that in the FMDV A dataset.

However, the performance of the non-linear models was consistently better than the linear models across all virus datasets, scoring methods and antigenic regions (i.e. loops/non-loops). Equations (13) and (14) detail the two best performing quadratic models; a grouped with loops and a non-weighted with loops model, respectively, for SAT2.

$$f(x) = 0.09 + 0.07x_1 + 0.09x_2 + 0.03x_1x_2 - 0.07x_2x_3 \\ -0.03x_1^2 + 0.09x_2^2 - 0.06x_3^2. \tag{13}$$

$$f(x) = 0.07 - 0.01x_1 + 0.07x_2 + 0.006x_3 + 0.05x_1x_2 \\ -0.002x_1x_3 - 0.06x_2x_3 + 0.0865243x_2^2 - 0.1x_3^2. \tag{14}$$

These regression models have a much larger number of terms than their linear counterparts. In other words, the non-linear models have a much higher complexity than that of the linear models. In cases where the GA-optimized non-linear regression models have performed worse than the linear models, the non-linear models are found to be much simpler (fewer number of coefficients), refer to Table 2. We also calculated the 95% confidence interval for all model coefficients to find out the statistically significant ones which supports the above and are also reported in Table 2. This implies that the relationship between the variables may be highly non-linear and there is a complex additive effect of two or more types of amino acid substitutions, which can be better explained by quadratic terms.

We carried out a null hypothesis test to see whether our best model (non-linear regression with GA using the grouped looped scoring method) performed better than the linear, non-linear regression with LS and cubic models using the same scoring methods. For the linear model, the r^2 (coefficient of determination of the regression line) was 0.186 and P value (probability of obtaining the observed sample results when the null hypothesis is actually true) was 0.10. For the non-linear model with LS r^2 was 0.326 and P was 0.05; for the cubic model r^2 was 0.467; P was 0.02. However, for our non-linear model with GA r^2 was 0.787; P was 0.0002. The above results show that even though there are improvements in both r^2 and P values from one to the next higher degree model, our method using a quadratic model with GA performs better than a cubic model. Thus, we can confidently reject the null hypothesis that our non-linear regression with GA does not perform better than the linear and the cubic models.

We also used five additional pairwise comparisons of FMDV A sequences (provided by The Pirbright Institute) to test our best model (grouping method with non-linear regression using GA). The model was able to achieve an 80% agreement rate with the actual serological data. Both the non-linear using LS and the linear models achieved 60% agreement rate on the actual data. It was generally seen that the sequences for which the test results did not match the actual serological data, the input variables for testing were outside the range of the data that was used for training the models. Consequently, we think that the availability of more data for training would have helped us achieve better agreement rates to the actual data. Table 1 gives a summary of the results. These results are highly encouraging, considering that only very limited amount data was available for training the models. Our results clearly indicate that nature-inspired optimization techniques are competitive for solving challenging biological problems.

Although our test results indicate that a cubic regression model may not necessarily better than the quadratic model, it is still of interest to investigate whether more general models, which can be created using symbolic regression with the help of genetic programming (Schmidt and Lipson, 2009), in particular if there is more data available.

4 Conclusions

This study compared three types of regression models, namely, linear, quadratic with parameter optimization using LSM and quadratic with both the structure and parameters being optimized with a hybrid GA. All of these models were further compared based on the presence/absence of specific antigenic loops for FMDV and influenza. The looped models performed ~20% better than the unlooped models. Better prediction results are achieved when amino acid changes in the capsid(s) were treated as different independent variables. Models with substitutions in the antigenic areas performed better than ones that took the entire exposed viral capsid protein(s) i.e. only the specific antigenic regions on the capsid protein(s) are most relevant for modelling. When substitutions in the capsid(s) are summarized, important information for prediction may be overlooked resulting in poor results. MSEs for looped models are lower than their unlooped counterparts except the scoring model for which the MSEs are almost the same. Comparing the weighted and the non-weighted methods, the grouping method performed best among the three, which is probably because the grouping method classifies amino acids based on their biochemical properties that enhance the models. The non-weighted approach did not perform as well as the grouping method perhaps because it considers all substitutions regardless of structural relevance, thereby including frequent substitutions of little significance to the model. The prediction power of the scoring method was unexpectedly low which is probably because a significant amount of biological data was lost when the substitution matrix was normalized to penalize only the non-frequent substitutions. Overall, the models that combined the scoring method based on biochemical properties of amino acids with non-linear regression for the specific antigenic areas on the capsid protein(s) gave best results.

We used a GA to optimize the structure and parameters to fit the non-linear model. GA is a much more robust optimization method than LSM as LSM cannot find the optimal solution of coefficients if the initially guessed solution is not sufficiently close to the ideal solution. Hence, GAs are able to work well with noisy biological data, which is reflected in our results as the non-linear regression models using the GA performed consistently better (throughout datasets) than the linear and the non-linear models whose parameters are estimated using the LSM.

The predictive ability of any model heavily depends on the data used to build them. Thus, the quality and availability of data has a profound effect on the efficiency and the predictive ability of the method. Consequently, due to lack of serological data, the results reported in this article are still preliminary. Nevertheless, considering that we have achieved similar results across six datasets for two different viruses (FMDV and Influenza), it would not be unreasonably optimistic to say that similar models can perhaps be deployed for other viruses when working towards developing an *in silico* vaccine predictor.

Acknowledgements

This work was funded in part by the Department of Computing and Department of Microbial and Cellular Sciences, University of Surrey, UK.

Conflict of interest: none declared.

References

Back,T. *et al.* (1997) *Handbook of Evolutionary Computation.* IOP Publishing Ltd., Bristol, UK.

Bandyopadhyay,S. (2007). *Analysis of Biological Data: A Soft Computing Approach*, Vol 3. World Scientific Publishing Company, Singapore.

Barnett,P. *et al.* (2001). The suitability of the emergency foot-and-mouth disease antigens held by the international vaccine bank within a global context. *Vaccine*, **19**, 2107–2117.

Betts,M. and Russell,R. (2003). Amino acid properties and consequences of substitutions. *Bioinform. Genet.*, **317**, 289.

Brusic,V. *et al.* (1994). Prediction of MHC binding peptides using artificial neural networks. In: Stonier,R.J. and Yu,X.H. (eds) *Complex Systems: Mechanism of Adaptation*. IOS Press, Amsterdam, pp. 253–260.

Chasman,D. and Adams,R. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.

Daintith,J. (1990). *A Concise Dictionary of Chemistry*, Vol. 3. Oxford University Press, USA.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978). A model of evolutionary change in proteins. In: Dayhoff,M.O. (ed.) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

De Groot,A.S. *et al.* (2002). Use of bioinformatics to predict MHC ligands and T-cell epitopes: application to epitope-driven vaccine design. *Methods Microbiol.*, **32**, 99–123.

Deb,K. and Agrawal,R. (1995). Simulated binary crossover for continuous search space. *Complex Syst.*, **9**, 115–148.

Friedman,J. *et al.* (2001). *The Elements of Statistical Learning*, Vol. 1. Springer Series in Statistics.

Goldberg,D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Boston, MA.

Holland,J.H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* MIT Press, Cambridge, MA.

Lee,M. and Chen,J. (2004). Predicting antigenic variants of influenza a/h3n2 viruses. *Emerg. Infect. Dis.*, **10**, 1385.

Lee,M. *et al.* (2007). Identifying potential immunodominant positions and predicting antigenic variants of influenza a/h3n2 viruses. *Vaccine*, **25**, 8133–8139.

Lesk,A.M. (2001). *Introduction to Protein Architecture: The Structural Biology of Proteins.* Oxford University Press, London.

Li,Y. (2010). Exploiting noisy and incomplete biological data for prediction and knowledge discovery. Ph.D. thesis, Netherlands Bioinformatics Centre, Nijmegen, Netherlands.

Liao,Y. *et al.* (2008). Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus. *Bioinformatics*, **24**, 505–512.

Meister,G.E. *et al.* (1995). Two novel t cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from mycobacterium tuberculosis and HIV protein sequences. *Vaccine*, **13**, 581–591.

Moxon,E. and Siegrist,C. (2011). The next decade of vaccines: societal and scientific challenges. *The Lancet*, **378**, 348–359.

Reeve,R. *et al.* (2010). Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus. *PLoS Comput. Biol.*, **6**, e1001027.

Reeve,R. *et al.* (2011). Reducing animal experimentation in foot-and-mouth disease vaccine potency tests. *Vaccine*, **29**, 5467–5473.

Schafer,J.R.A. *et al.* (1998). Prediction of well-conserved HIV-1 ligands using a matrix-based algorithm, epimatrix. *Vaccine*, **16**, 1880–1884.

Schmidt,M. and Lipson,H. (2009). Distilling free-form natural laws from experimental data. *Science*, **324**, 81–85.

Schwefel,H.-P. (1995). *Evolution and Optimum Seeking.* Wiley, USA.

Thompson,J. *et al.* (2002). Multiple sequence alignment using clustalw and clustalx. *Curr. Protoc. Bioinformatics*, 2–3.