Published in final edited form as:

Stat Med. 2015 May 10; 34(10): 1791–1798. doi:10.1002/sim.6452.

Vaccine design via nonnegative lasso-based variable selection

Zonghui Hu¹, Dean A. Follmann¹, and Kazutoyo Miura²

Zonghui Hu: huzo@niaid.nih.gov; Dean A. Follmann: dfollmann@niaid.nih.gov; Kazutoyo Miura: kmiura@niaid.nih.gov

¹Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA

²Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland, USA

SUMMARY

There are many different strains of malaria parasites, each represented by a unique sequence of amino acids. A desirable vaccine would match the amino acid sequence of the parasite antigen. Due to the three-dimensional structure of protein, not all sites in the amino acid sequence participate in the binding between the vaccine induced antibody and the parasite antigen. Nor do all sites have equal importance. In this work, we apply a nonnegative lasso based variable selection to identify the “important” amino acid sites and evaluate their relative importance. We then define a metric, the functional coverage, to measure the “effective” matching in the amino acid sequence between the vaccine and the parasite. With the variable selection procedure, development of a vaccine needs only to target the important sites, and the potential effectiveness of a vaccine candidate is reflected by the functional coverage.

Keywords

Amino acid sequence; Malaria vaccine; Nonnegative lasso regression; Seemingly independent screening

1 Problem description

1.1 Vaccine, parasite, and amino acid sequence

Malaria is a major global health problem. There are many strains of parasites that cause infection. Vaccination is an attractive method of control. However, there has been no vaccine showing significant clinical protection. Each parasite is represented by the protein on a region of its surface, which is called an antigen in immunology. Different strains of parasites correspond to different antigens. A malaria vaccine is usually an antigen, which induces the body to produce antibody that recognizes and binds to that antigen. To vaccinate against a parasite, the vaccine candidate antigen should be similar to the parasite antigen so that the induced antibody binds to the parasite antigen and blocks infection. Due to strain variability, effective vaccine design has been a major challenge [1].

An antigen is characterized by its amino acid (AA) sequence. Ideally, the vaccine should have the same AA sequence as the parasite antigen. To target a single parasite, a homologous antigen, which is the antigen of the same AA sequence as the parasite antigen,

is the best choice. For multiple parasites with distinct amino acid sequences, it is impossible for a monoclonal vaccine to match the parasites over all AA sites.

Table 1 shows a section of the AA sequences from four parasites that differ at sites 1, 2, 6, 7, 8, 9 and 10. There is not a single antigen that can fully match the parasites over all the sites. However, not all AA sites play a role in the antibody-antigen binding. Due to the three-dimensional structure of protein molecular, there is bending and folding in the AA sequence. The sites that are hard to be accessed do not actively participate in binding. Even among the AA sites involved in binding, not all sites may have the same impact on immune response [2, 3]. Suppose the only important sites are 3, 4, 5, 11, 12 and 13, then any antigen in Table 1 is a good vaccine candidate antigen as it matches all parasites at the six important sites.

The goal of this work is to identify the important AA sites and estimate their relative importance. By “important AA sites”, we mean the AA sites that actively participate in antibody-antigen binding and have big impact on the vaccine induced immunity. We then devise a metric, the functional coverage, to measure the effective matching of a vaccine with a parasite. By “effective matching”, we mean how well a vaccine matches a parasite at the important AA sites.

1.2 Malaria vaccine study

Apical membrane antigen 1 (AMA1) is one family of antigens representing the malaria parasites, and the *Plasmodium falciparum* AMA1 antigens have been the leading malaria vaccine candidates. The AMA1 antigens, and consequently the malaria parasites, can be grouped into six populations [4]. In a malaria vaccine study, six vaccine candidate antigens, one from each population, were evaluated against six parasites [3]. Thirty-six rabbits were involved: each rabbit was immunized by one vaccine, then antibody from the rabbit's blood was purified and tested against a parasite. The immune response of a rabbit under a vaccine-parasite pair was measured by the growth inhibition assay (GIA, [5]), with higher GIA indicating better immunity against the parasite. There are a total of 36 observations, each corresponding to a vaccine-parasite pair observed from one rabbit. Table 2 shows one observation, where column 1 to 10 lists a section of the AA sequences, $\mathbf{X}$ is the vector of binary variables: $X_k = 1$ if the vaccine and the parasite have the same AA at site k and $X_k = 0$ if otherwise. The response, 0.77, is the GIA measure under the vaccine-parasite pair. The AA sequence of the AMA1 antigen family has around 600 AA sites, where 59 sites are polymorphic. The other 90% of sites are constant over the family and display no strain variability. At these unchanging sites, a vaccine candidate antigen assumes the same amino acids as the AMA1 antigens. Therefore the unchanging sites do not matter for vaccine design. Hereafter, by AA sites, we mean the polymorphic sites only. In this study, we have $\mathbf{X} = (X_1, \dots, X_{59})$, but $\mathbf{X}$ can have a much higher dimension in general.

Note that definition of the binary variable $\mathbf{X}$ is based upon two assumptions. First, at a given site, the AA's in all matched pairs are equally matched so that (D,D) counts the same as (P,P). Second, the AA's in all unmatched pairs are equally different, so an E Q mismatch counts the same as an E S. These assumptions are adopted by almost all research on the development of malaria vaccine [1], [6], [7], [8].

1.3 Objective

Our objective is to find the important AA sites and their relative importance. This information helps to ease vaccine design. To prevent infection from a single parasite, the vaccine does not need to match the parasite over all the AA sites but only at the important AA sites. To prevent infection from multiple parasites, one approach is to use a mixture of antigens. The hypothesis is that a parasite can be recognized if it is covered by the mixture at each important site; that is, at each important site k , $a_k^P \in A_k^V = \{a_k^{V_1}, \dots, a_k^{V_m}\}$ with a_k^P the amino acid from the parasite and A_k^V the set of amino acids from the antigens in the mixture. Suppose the important AA sites in Table 1 do not include site 7 and 8, then a mixture of the first two antigens is enough to block infection from all four parasites. It has been shown that a mixture of five antigens may protect against a majority of malaria parasites [3].

2 Important AA site selection

2.1 Nonnegative lasso based variable selection

With the match between vaccine and parasite defined by the vector of binary variables $\mathbf{X}$, AA site selection boils down to variable selection. There are challenges for this variable selection. First, an AA sequence can contain hundreds or even thousands of sites, and there is the issue of more variables than observations. Second, all variables in $\mathbf{X}$ are binary, variable selection methods that require normality or spherical symmetry, such as the principal component analysis based approaches, are inappropriate.

Our approach to determining the important AA sites is through regression,

$$Y_i = \sum_{k=1}^K \beta_k X_{ik} + \varepsilon_i, \beta_k \geq 0 \quad (1)$$

where $k = 1, \dots, K$ indicate the polymorphic AA sites, $Y_i = Y(V, P)$ represents the immune response observed in rabbit i under the vaccine-parasite pair (V, P) , X_i represents the vaccine-parasite matching: $X_{ik} = 1$ if the vaccine matches the parasite at site k and $X_{ik} = 0$ if otherwise, ε_i is the random error. For presentational simplicity, Y_i is centered around mean by $Y_i = Y_i - 1/n \sum_{i=1}^n Y_i$. Since only a few AA sites actually participate in the antibody-antigen binding and contributes to Y , (1) is a sparse model with $\beta_k = 0$ for most sites.

As matching at an AA site entails better binding between the vaccine-induced antibody and the parasite antigen, it leads to better immunity. There-fore, $\beta_k \geq 0$ for $k = 1, \dots, K$. With Y centered around the mean, β_k indicates the increase in immune response by matching the vaccine to the parasite at site k . The sites with $\beta_k = 0$ might be hidden or hard to access, they do not participate in the antibody-antigen binding, and thus do not impact Y whether the amino acids at those sites match or not. We define the set of important AA sites as $M = \{1 \leq k \leq K : \beta_k > 0\}$.

The objective of this work is to find M . This differs from conventional variable selection. Rather than predicting the actual value of the response, our interest is to identify the AA sites that, if a vaccine is designed to match the parasite at these sites, can improve induced

immunity. We want to exclude the sites associated with response with negative coefficients. As will be shown in section 3.1, those sites with negative coefficients are mostly related to the response through correlation with other important sites.

An effective approach for handling a sparse model is lasso regression [9]. For important AA site selection, we adopt a nonnegative lasso regression

$$\begin{cases} \text{argmin} & \sum_{i=1}^n (Y_i - \sum_{k=1}^K \beta_k X_{ik})^2 \\ \text{s.t.} & \beta_k \geq 0, k=1, \dots, K \\ & \sum_{k=1}^K \beta_k \leq t, \end{cases} \quad (2)$$

where the second inequality corresponds to lasso penalty with $t > 0$ the tuning parameter. Since $\beta_k \geq 0$, (2) differs from standard lasso regression. Take $K = 2$ for example. Instead of the square region defined by $\sum_k |\beta_k| \leq t$ in standard lasso, the optimization region for the nonnegative lasso is the triangle in the first quadrant. The nonnegative lasso is easily estimable by quadratic programming or by the R package “penalized”. Asymptotic properties of the nonnegative lasso can be found in [10] and [11]. Denote the selected important sites as $M = \{1 \leq k \leq K : \hat{\beta}_k > 0\}$, with larger $\hat{\beta}_k$ indicating greater importance.

When the dimension of $\mathbf{X}$ is very high, lasso regression may be subject to computational burden and instability. We can first screen out the variables with little marginal impact through sure independence screening (SIS, [12]). SIS treats all variables as independent and fits individual univariate regressions

$$Y_i = \omega_k X_{ik} + \eta_i, \text{ for } k=1, \dots, K. \quad (3)$$

Let $c \geq 0$ be a threshold. The SIS selected subset is $M_{\text{SIS}} = \{1 \leq k \leq K : \hat{\omega}_k \geq c\}$. It is not ensured that all variables in M_{SIS} are important. The false positivity could be due to (i) threshold c is not big enough to screen out all the unimportant variables; (ii) variables in M_{SIS} are individually but not jointly important due to correlations among variables. We then follow SIS with the nonnegative lasso regression over the variables in M_{SIS} . Rigorous selection of the tuning parameter t helps to screen out the variables falsely picked by SIS due to (i). Though both SIS and lasso are based upon regression, lasso accounts for correlations among variables and thus screens out the variables falsely selected by SIS due to (ii). It was shown in [12] that, with proper selection of c , M_{SIS} contains M with probability 1. For a good chance to include all the potentially important variables into M_{SIS} , we pick a small threshold c . Since the lasso step further screens out unimportant variables, a small c should not cause false positivity. In this study, we let $c = 0$.

The tuning parameter t can be selected through cross validation, for which [9] gives details. Since the number of AA sites that can be selected in vaccine design is usually limited, another way for choosing t is by the number of designable sites. For example, if out of K AA sites, we have to pick no more than $K/2$ important AA sites, we can pick t around $\sum_1^K \tilde{\beta}_k / 2$ where $\tilde{\beta}$ is the estimate in (2) without lasso penalty.

2.2 Functional coverage

For a population of parasites where the vaccine induced immune response follows model $Y = \sum_{k \in M} \beta_k X_k + \varepsilon$, with $\beta_k > 0$ the relative importance, we define the functional coverage of a vaccine V on a parasite P as

$$e_{v,p} = \frac{\sum_{k \in M} \beta_k X_k^{v,p}}{\sum_{k \in M} \beta_k},$$

where $X_k^{v,p} = 1$ if V and P match at site k and $X_k^{v,p} = 0$ if otherwise.

The functional coverage measures the “effective matching” of a vaccine with a parasite; that is, matching at the important AA sites. It ranges from 0 to 1. A vaccine of higher functional coverage is expected to induce better immunity. Suppose P is a parasite with M the set of important AA sites, V_1 and V_2 are two vaccines with $M_1 = \{k \in M, X_k^{v_1,p} = 1\}$ and $M_2 = \{k \in M, X_k^{v_2,p} = 1\}$. If V_1 matches P at more sites than V_2 in the sense $M_1 \supset M_2$, then $e_{v_1,p} > e_{v_2,p}$. If V_1 matches P at more important sites than V_2 , for example, $M_1 = \{1, 2\}$ and $M_2 = \{1, 5\}$ with $\beta_2 > \beta_5$, then $e_{v_1,p} > e_{v_2,p}$.

For a group of parasites $\mathcal{P} = \{P_1, \dots, P_J\}$, we define the functional coverage of vaccine V as

$e_{v,\mathcal{P}} = \sum_{j=1}^J e_{v,p_j} / J$, the average coverage of the vaccine over the parasites. We can also define $e_{v,\mathcal{P}}$ as the minimum or maximum coverage over the parasites depending on the vaccination goal. Whether targeting a single parasite or a group of parasites, a vaccine with higher functional coverage is a better choice. In estimating the function coverage, we replace M and β_k by their estimates $\hat{M}$ and $\hat{\beta}_k$.

3 Malaria vaccine study

3.1 Important AA sites

As introduced in section 1.2, the malaria vaccine study measures the immune response of six study vaccines against six parasites. Table 3 presents the results from the nonnegative lasso-based selection. The tuning parameter in the non-negative lasso regression is selected through cross validation, which is shown in the supplementary document. We see that the important AA sites are mostly also individually important. Out of the 13 important sites, nine have the individual impacts among the top 10 magnitudes of individual impacts.

We evaluate the sensitivity and specificity of the nonnegative lasso-based selection, where sensitivity is defined as the probability of an important variable being selected and specificity the probability of an unimportant variable not being selected. For this purpose, we regenerate the response from

$$Y_i^* = \sum_{k \in \hat{M}} \hat{\beta}_k X_{ik} + \varepsilon_i^*$$

where X_{ik} , $i = 1, \dots, 36$, $k = 1, \dots, 59$ are the same as in the study, $\hat{M}$ and $\hat{\beta}_k$ are the nonnegative lasso estimates, and ε_i the centered regression residuals. Letting $\{\varepsilon_i^*: i=1, \dots, 36\}$ be the bootstrapping residuals, $\{Y_i^*, i=1, \dots, 36\}$ mimics the possible immune response from another set of 36 rabbits. The boot-strap procedure indicates a sensitivity of 0.93 and a specificity of 0.86 for the nonnegative lasso-based selection based on 2000 replications.

We also apply one conventional variable selection, the best subset selection [13]. Like most variable selection methods, best subset selection does not exclude variables with negative regression coefficients. It is a discrete decision where a variable is either selected or not, and can be sensitive to perturbation in data. The best subset selection selects sixteen AA sites, eight of them not selected by the nonnegative lasso. All those eight sites have low individual impacts, and five of them correspond to negative coefficients in multivariate regression. A correlation analysis, with outcomes presented in the supplementary document, indicates that these eight sites are highly correlated with the other more important sites.

3.2 Vaccine functional coverage

In addition to the six parasites involved in section 3.1, there are another six parasites whose AA sequences are known. These six additional parasites also fall into the six parasite populations, and the results in Table 3 apply. Let Vac.1, $\dots$, Vac.6 stand for the six study vaccines, Par.1, $\dots$, Par.12 the twelve parasites, where the first six are the parasites studied in section 3.1 and the last six the additional parasites. We compute $e_{v,p}$, the functional coverage of each vaccine on each parasite, based on the nonnegative lasso results in Table 3. The functional coverages are presented in Figure 1, along with the 2.5% and 97.5% bootstrap percentiles. These percentiles roughly reflect the variability of functional coverage estimates but cannot be interpreted as 95% confidence intervals [14].

Among the six study vaccines, Vac.1, Vac.2, Vac.4, and Vac.5 are homologous to Par.1, Par.2, Par.4 and Par.5 with identical AA sequence. Therefore, the four vaccines have functional coverage of 1 on the four respective parasites. Vac.3 and Vac.6, are not homologous to any parasites but from the same population as Par.3 and Par.6, respectively [4]. We see in the Figure that, among all the six vaccines, Vac.3 shows the highest functional coverage on Par.3 and Vac.6 the highest functional coverage on Par.6. Meanwhile, among the twelve parasites, Par.3 responds the best to Vac.3 and Par.6 the best to Vac.6. These observations support the theory of [4].

The functional coverage can guide vaccine selection without the actual testing of a vaccine against the parasites. Due to limited sample size in this study, the difference in functional coverage among vaccines is not statistical significant. However, Figure 1 is informative about the trend. Take Par.12 as an example. Since Vac.6 shows the highest functional coverage, it is probably the best vaccine against Par.12 among the six vaccine candidates. In Figure 1, the last column in each plot shows $e_{v,\mathcal{P}}$, the average functional coverage over the twelve parasites. To target all the twelve parasites as a whole, Vac.1 is probably the best choice as it has the highest average coverage over these parasites.

4 Simulation study

Since the goal is identification of the important AA sites, we use sensitivity and specificity as performance metrics. Recall that sensitivity is $\text{pr}(k \in M \mid k \in M)$ and specificity is $\text{pr}(k \notin M \mid K \notin M)$. We compare the nonnegative lasso-based selection with the best subset selection.

We first evaluate the scenario where variables in $\mathbf{X}$ are dependent. We mimic the scenario of malaria vaccine study by letting $\mathbf{X}_i$ be the same as in the study and $Y_i = \sum_{k \in M} \beta_k X_{ik} + \varepsilon_i$ for $i = 1, \dots, 36$. We consider three cases for data generation. In case I, M and β_k are the nonnegative lasso estimates. In case II, M and β_k are the best subset estimates where variables with negative coefficient estimates are removed. In case III, M and β_k are the SIS estimates from (3) where variables with small univariate regression coefficients are removed. The size of M is 13, 11, and 33 in the three cases, respectively. Simulation results are presented in the top portion of Table 4. The nonnegative lasso-based selection shows good sensitivity and specificity in both case I and II. In case III, some variables in M are correlated and not jointly important. The nonnegative lasso-based selection shows a reasonably low sensitivity but a good specificity. The best subset selection also shows a low sensitivity. Among the selected important sites from best subset, around 7% correspond to negative multivariate regression coefficients due to correlations among the variables in M .

We next evaluate the scenario where variables in $\mathbf{X}$ are all independent. Let $\mathbf{X} = (X_1, \dots, X_K)$, where X_k , $k = 1, \dots, K$, come from independent Bernoulli (p_k), and $Y_i = \sum_{k \in M} \beta_k X_{ik} + \varepsilon_i$. Without loss of generality, the first m variables are important with positive coefficients and the rest have coefficients zero. We consider three cases for β_k . In case I, $\beta_k = 10$ for $k \leq m$. In case II, $\beta_k = 2$ for $k \leq m/2$ and $\beta_k = 1$ for $m/2 < k \leq m$. In case III, $\beta_k = 1$ for $k \leq m/3$, $\beta_k = 0.5$ for $m/3 < k \leq 2m/3$ and $\beta_k = 0.3$ for $2m/3 < k \leq m$. The bottom portion of Table 4 presents the simulation results at $n = 100$, $K = 90$ and $m = 24$ with p_k from Beta distribution (2, 5). Other distributions for p_k do not show much difference. The lasso-based selection shows consistently good sensitivity and specificity. The best subset selection shows poor specificity, and its sensitivity can be low when β_k are small. Some additional results can be found in the supplementary document.

5 Discussion

In this work, we adopt a nonnegative lasso-based variable selection for identifying the important AA sites and their relative importance. We then estimate the functional coverage of a vaccine on a parasite or a group of parasites. The task of vaccine selection and vaccine design is simplified to the search for a vaccine with the highest functional coverage. As AA sequences are easily obtained, the proposed approach can facilitate vaccine selection and design without the costly production of each vaccine candidate followed by vaccination and measuring of the actual immune response.

The proposed variable selection is based upon a nonnegative lasso regression, which imposes the restriction of nonnegative regression coefficients. The nonnegativity constraint reflects vaccine science, where each coefficient reflects the improvement in vaccine-induced

immunity when the vaccine matches the parasite antigen at the respective AA site. An alternate to the nonnegative lasso is the nonnegative least square regression, which takes the same form as (2) but without the L_1 constraint $\beta_k \geq 0$. It has been shown in [11] that the sign constraint alone acts comparable to an L_1 constraint for high-dimensional regression and can lead to as good estimation as lasso. In fact, the two restricted regressions lead to the same results in the malaria vaccine example.

In this work, we focus on the situation that the response follows a linear model. The proposed selection method may be extended to nonlinear response model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Thera MA, Doumbo OK, Coulibaly D, Laurens MB, Ouattara A, Kone AK, Guindo AB, Traore K, Traore I. A field trial to assess a blood-stage malaria vaccine. *New England Journal of Medicine*. 2011; 365:1004–1013. [PubMed: 21916638]
2. Dutta S, Lee SY, Batchelor AH, Lanar DE. Structure basis of antigenic escape of a malaria vaccine candidate. *Proceedings of the National Academy of Sciences*. 2007; 104:12 488–12 493.
3. Miura K, Herrera R, Diouf A, Zhou H, Mu J, Hu Z, MacDonald NJ, Reiter K, Nguyen V, Shimp RL, et al. Overcoming allelic-specificity by immunization with 5-allelic forms of plasmodium falciparum apical membrane antigen 1. *Infection and Immunity*. 2013; 81:1491–1501. [PubMed: 23429537]
4. Duan J, Mu J, Thera MA, Joy D, Kosakovsky P, Diemert D, Long C, Zhou H, Miura K, Dolo A, et al. Population structure of the genes encoding the polymorphic plasmodium falciparum apical membrane antigen: implications for vaccine design. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105:7857–7862. [PubMed: 18515425]
5. Malkin EM, Diemert DJ, McArthur JH, Perreault JR, Miles AP, Giersing BK, Mullen GE, Orcutt A, Muratova O, Awkal M, et al. Phase 1 clinical trial of apical membrane antigen: an asexual blood-stage vaccine for plasmodium falciparum malaria. *Infection and Immunity*. 2005; 73:3677–3685. [PubMed: 15908397]
6. Kusi KA, Remarque EJ, Riasat V, Walraven V, Thomas AW, Faber BW, Kochen CHM. Safety and immunogenicity of multi-antigen ama1-based vaccines formulated with covaccine ht and montanide isa 51 in rhesus macaques. *Malaria Journal*. 2011; 10:182. [PubMed: 21726452]
7. Drew DR, Hodder AN, Wilson DW, Foley M, Mueller I, Siba PM, Dent AE, Cowman AF, Beeson JG. Defining the antigenic diversity of plasmodium falciparum apical membrane antigen 1 and the requirements for a multi-allele vaccine against malaria. *PLoS One*. 2012; 7:e51 023.
8. Srinivasan P, Ekanem E, Diouf A, Tonkin ML, Miura K, Boulanger MJ, Long CA, Narum DL, Miller LH. Immunization with a functional protein complex required for erythrocyte invasion protects against lethal malaria. *New England Journal of Medicine*. 2014; 111:10 311–10 316.
9. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
10. Yuan M, Lin Y. On the non-negative garrote estimator. *Journal of the Royal Statistical Society, Series B*. 2007; 69:143–161.
11. Meinshausen N. Sign-constrained least squares estimation for high dimensional regression. *Electronic Journal of Statistics*. 2013; 7:1607–1631.
12. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*. 2008; 70:849–911.
13. Breiman L, Spector P. Submodel selection and evaluation in regression: the x-random case. *International Statistical Review*. 1992; 60:291–319.

14. Chatterjee A, Lahiri SN. Bootstrapping lasso estimators. *Journal of the American Statistical Association*. 2011; 106:608–625.

Figure 1.

Functional coverage of each vaccine over the twelve parasites with the 2.5% and 97.5% bootstrap percentiles. The last column in each plot is the average functional coverage over the twelve parasites.

Table 1

A section of the AA sequence of four parasite antigens. The first line indicates the amino acid sites, the last four lines the AA sequences.

Amino acid sequence													
Antigen	1	2	3	4	5	6	7	8	9	10	11	12	13
1	E	G	D	D	E	M	E	P	Q	N	N	T	D
2	Q	E	D	D	E	I	E	P	E	K	N	T	D
3	Q	E	D	D	E	M	Q	P	Q	N	N	T	D
4	Q	E	D	D	E	M	E	S	E	N	N	T	D

Table 2

Observation under a vaccine-parasite pair. On the first line are the indices of AA sites, on the second line the AA sequence of the vaccine, and on the third line the AA sequence of the parasite.

Amino acid sequence											
	1	2	3	4	5	6	7	8	9	10	Response
Vaccine	K	D	H	P	Q	E	Q	N	P	E	...
Parasite	K	D	H	P	E	Q	P	N	P	S	...
X	1	1	1	1	0	0	0	1	1	0	0.77

Table 3

Important AA sites from nonnegative lasso-based selection: on the first line are the 13 important AA sites; on the second line the relative importance $\hat{\beta}_k$ from (2); on the third line the individual impact $\hat{\omega}_k$ from (3).

$\hat{M}$	19	9	24	8	33	7	6	14	14	34	15	10	16
$\hat{\beta}_k$	8.59	4.05	3.64	3.19	2.45	2.02	1.56	1.39	1.39	0.71	0.65	0.30	0.10
$\hat{\omega}_k$	11.21	9.10	7.41	6.43	6.68	2.81	6.00	4.23	4.23	2.68	10.12	1.22	3.13

Table 4

Simulation outcomes

Scenario of malaria vaccine study				
Case	Best subset selection		nonnegative lasso-based selection	
	Sensitivity	Specificity	Sensitivity	Specificity
I	0.83	0.71	0.93	0.89
II	1	0.79	1	0.77
III	0.41	0.58	0.55	0.85
Scenario of independent variables in X				
Case	Best subset selection		nonnegative lasso-based selection	
	Sensitivity	Specificity	Sensitivity	Specificity
I	1	0.26	0.97	0.8
II	1	0.26	0.94	0.76
III	0.84	0.25	0.91	0.76