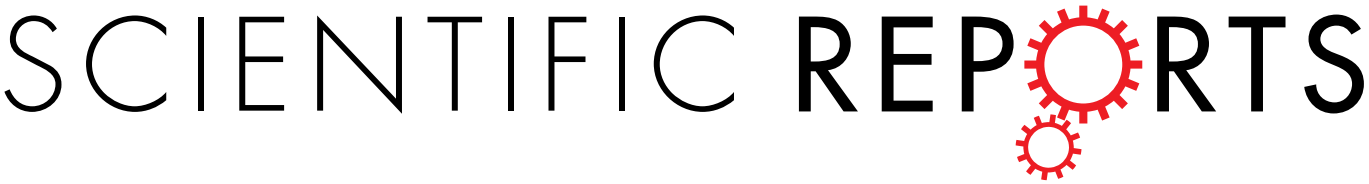

OPEN

sNebula, a network-based algorithm to predict binding between human leukocyte antigens and peptides

Received: 06 April 2016
Accepted: 02 August 2016
Published: 25 August 2016

Heng Luo[1,2], Hao Ye[1], Hui Wen Ng[1], Sugunadevi Sakkiah[1], Donna L. Mendrick[1] & Huixiao Hong[1]

Understanding the binding between human leukocyte antigens (HLAs) and peptides is important to understand the functioning of the immune system. Since it is time-consuming and costly to measure the binding between large numbers of HLAs and peptides, computational methods including machine learning models and network approaches have been developed to predict HLA-peptide binding. However, there are several limitations for the existing methods. We developed a network-based algorithm called sNebula to address these limitations. We curated qualitative Class I HLA-peptide binding data and demonstrated the prediction performance of sNebula on this dataset using leave-one-out cross-validation and five-fold cross-validations. This algorithm can predict not only peptides of different lengths and different types of HLAs, but also the peptides or HLAs that have no existing binding data. We believe sNebula is an effective method to predict HLA-peptide binding and thus improve our understanding of the immune system.

Human leukocyte antigens (HLAs) are the major histocompatibility complexes (MHCs) in humans. They are expressed on the surfaces of antigen presenting cells to recognize endogenous or foreign peptides for immunological reactions[1,2]. The genes that encode HLAs are a gene system located at the short arm of Chromosome 6. They are highly polymorphic across populations[3–5]. There are different classes of HLAs, including Class I, II and III, according to their genetic locations. Different classes of HLAs have divergent structures and functions. Both Class I and Class II HLAs have a long binding groove that can bind peptides and present them onto T-cell receptors[6–8], while Class III HLAs are a part of the complement system to help with pathogen clearance[9]. Class I HLAs capture the endogenous peptides degraded from cytosolic proteins and present them to the T-cell receptors on the surface of CD8+ T-cells for cytotoxic responses, while the Class II HLAs present exogenous peptides from extracellular sources to the CD4+ T-cells to trigger acquired responses including antibody synthesis[10,11]. The binding between Class I/II HLAs and peptides is an important process for immune responses. Studying HLA-peptide binding will help us better understand the immune system and the mechanisms of autoimmune diseases and adverse drug reactions[12,13] and will also provide important information needed in the development of vaccines and protein therapeutics[14,15].

Since HLA-peptide binding is important for immune-related applications, experimental binding assays were developed to test *in vitro* binding affinities between HLAs and peptides and the data were collected in databases such as AntiJen[16], IEDB[17], MHCBN[18] and SYFPEITHI[19]. The IMGT/HLA database recorded more than 13,000 HLA alleles by August 2015[20]. Since it is time-consuming and costly to experimentally test the binding between large numbers of HLAs and peptides, computational methods have been developed to predict HLA-peptide binding[21]. The current widely used methods are machine learning methods; however, several challenges limit their applicability. First, many machine learning methods can only predict a limited number of HLAs or peptides with a specific length. Second, an HLA-specific model would be unreliable if the training samples were not large enough[21]. Therefore, we developed the neighbor-edges based and unbiased leverage algorithm (Nebula) based on

[1]National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Rd, Jefferson, AR 72079 USA. [2]University of Arkansas at Little Rock/University of Arkansas for Medical Sciences Bioinformatics Graduate Program, 2801 S University Ave, Little Rock, Arkansas, AR 72204 USA. Correspondence and requests for materials should be addressed to H.H. (email: Huixiao.Hong@fda.hhs.gov)

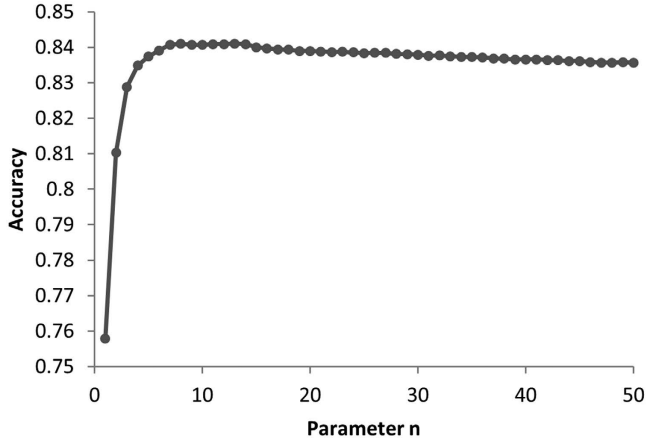

Figure 1. Determination of parameter *n* using LOO cross-validations for sNebula. The y-axis is the prediction accuracy and the x-axis indicates *n*.

network analysis to overcome the limitations of machine learning methods[22,23]. We successfully applied Nebula to predict HLA-peptide binding and found that it delivered a reasonable performance. However, Nebula is not applicable to predict the binding between a peptide and an HLA if no experimental data are available between the peptide and other HLAs or no binding assay has been developed for the HLA. Thus, Nebula is not able to predict binding for unstudied peptides and HLAs, limiting its application. Nebula is an algorithm purely based on the topology of a network; alternatively, the network is treated as a colorless graph where the nodes are not differentiated (colorless). Actually, the nodes (HLAs and peptides) in the bipartite network of HLA-peptide could be differentiated in many ways. Thus, appropriate consideration of node difference in a prediction algorithm is expected to improve its performance. In this study, we developed a new network-based prediction algorithm called similar neighbor-edges based and unbiased leverage algorithm (sNebula) by presenting the bipartite network of HLA-peptide binding data in a color graph. By introducing color to the network as additional information, sNebula can predict binding activity for peptides and HLAs that are not included the training network, overcoming the limitation of Nebula. We used the qualitative binding data between Class I HLAs and peptides as an example. We demonstrated that sNebula is a reliable algorithm for prediction of HLA-peptide binding and can be applied to HLAs or peptides with or without experimental binding data.

Results

Data curation. We curated 43,935 peptides, 135 Class I HLAs and 141,224 qualitative HLA-peptide binding data from the four databases. The binding data are given in Supplementary Table S1. Among the 43,935 distinct peptides, the peptide length varies from 6 to 30. Most of the peptides are 9-mers (65%) and 10-mers (25%), which is consistent with the experimental discovery of Class I HLA-binding peptides[24]. The distribution of peptide lengths is summarized in Supplementary Table S2. The 135 HLAs include 49 *HLA-A* alleles, 75 *HLA-B* alleles, 9 *HLA-C* alleles and 2 *HLA-E* alleles. The HLA alleles and their pseudo-sequences are listed in Supplementary Table S3. Among the 141,224 HLA-peptide binding data, 47% are bindings and 53% are non-bindings.

Leave-one-out (LOO) cross-validation. Parameter *n* indicates the maximum of neighbors from the peptide and HLA that are used for sNebula to make a prediction of the binding between the HLA and the peptide. Fifty leave-one-out (LOO) cross-validations were conducted on the HLA-peptide binding data using parameter $n = 1$ to 50. The prediction accuracy values yielded from the 50 LOO cross-validations are shown in Fig. 1. When $n = 13$, the accuracy reached the maximal value 0.841, and the corresponding sensitivity, specificity and area under the receiver operating characteristic curve (AUC) values were 0.818, 0.862 and 0.841, respectively. As *n* increases after this point, the accuracy of the model gradually dropped. Thus we used $n = 13$ for LOO cross-validations.

Five-fold cross-validations. One thousand iterations of five-fold cross-validations were conducted on the HLA-peptide binding data using sNebula. The prediction values in each of the five-fold cross-validations were compared with the experimental values and a set of accuracy, sensitivity and specificity values was calculated. The distributions of 1,000 values of these performance metrics are shown in Fig. 2. The average sensitivity, specificity and accuracy values are 0.816, 0.852 and 0.835, respectively, with the same standard deviation of 0.001.

Confidence analysis. The sNebula predictions are continuous values that not only indicate binding status of binder/non-binder but also represent the prediction confidence levels. The confidence levels of sNebula predictions from the 1,000 iterations of five-fold cross-validations were calculated and used to place the predictions into 10 groups by confidence. The performance of sNebula was assessed for each of the 10 groups of predictions. The performance in terms of accuracy, sensitivity, specificity and AUC at different confidence levels were plotted in Fig. 3. As the confidence increased, the AUC, accuracy, sensitivity and specificity (indicated by the left y-axis) improved, and the predictions in number (indicated by the right y-axis) also increased. The confidence analysis

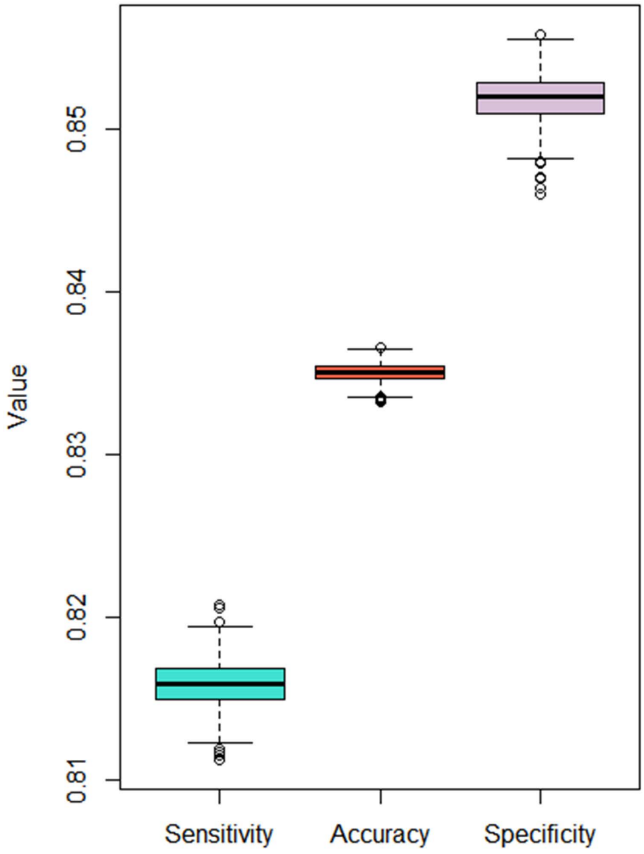

Figure 2. The distributions of sensitivity, specificity and accuracy seen in the 1,000 iterations of five-fold cross-validations.

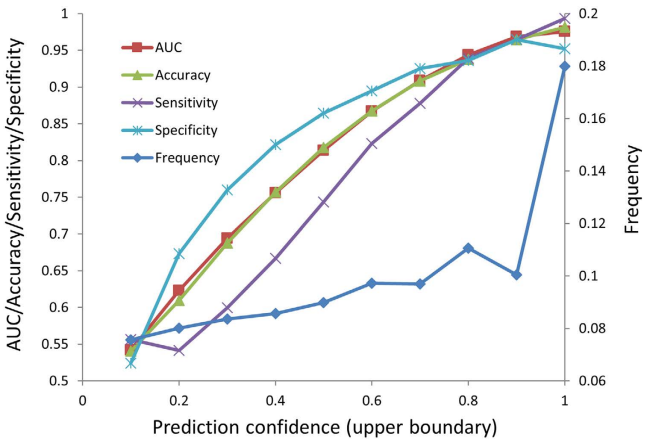

Figure 3. The relationships between prediction confidence and sNebula performance (AUC, accuracy, sensitivity, specificity) and prediction frequency. The confidence values ranging from 0 to 1 are grouped into 10 bins. The X-axis represents the upper boundary of each confidence bin. The left Y-axis indicates AUC, accuracy, sensitivity and specificity. The right Y-axis gives prediction frequency.

results revealed that the higher the prediction confidence level the better the prediction performance of sNebula. Moreover, most predictions from sNebula were at high confidence.

Benchmark. The IEDB website contains the performance comparison of various prediction methods for HLA-peptide binding (http://tools.iedb.org/auto_bench/mhci/weekly/). We used NetMHCpan[25–27] to compare with sNebula and Nebula. The performance comparison is shown in Table 1. Different methods had different performance depending on the dataset and the HLA. While NetMHCpan performed well on some datasets such

Dataset	IEDB reference	HLA	Peptide length	Peptide count	Measurement	NetMHCpan		sNebula		Nebula	
						SRCC	AUC	SRCC	AUC	SRCC	AUC
2016-05-03/2016-02-19	1029957	B*38:01	9	28	ic50	0.766	0.963	0.171	0.531	0.400	1.000
	1029824	A*02:01	9	77	binary	0.071	0.546	0.060	0.539	—	—
2015-08-07	1027131	B*15:02	9	14	binary	0.713	1.000	0.693	0.939	−0.707	0.000
	1029125	B*27:04	9	21	binary	0.717	0.939	0.133	0.582	—	—
	1029125	B*27:05	9	21	binary	0.751	0.959	0.752	0.959	—	—
	1029125	B*27:06	9	21	binary	0.421	0.750	0.421	0.750	0.500	0.750
	1029061	B*57:01	9	26	ic50	0.612	0.943	0.169	0.575	0.000	0.250
	315209	C*03:04	9	14	t1/2	0.781	0.911	0.113	0.923	—	—
	1028928	A*02:01	9	13	binary	0.570	0.955	0.539	0.909	—	—
	1028928	B*07:02	9	12	binary	0.648	1.000	0.522	0.900	—	—
	315174	B*27:03	9	11	binary	0.657	0.893	0.179	0.607	0.436	0.750
	1028790	A*02:01	9	55	ic50	0.615	0.574	0.505	0.778	0.478	0.856
	1028790	A*02:01	10	35	ic50	0.407	0.677	0.432	0.704	0.528	0.725
	1028790	A*02:02	9	55	ic50	0.582	0.713	0.372	0.680	0.427	0.668
	1028790	A*02:03	9	55	ic50	0.539	0.696	0.477	0.629	0.450	0.757
	1028790	A*02:03	10	35	ic50	0.208	0.750	0.419	0.697	0.308	0.691
	1028790	A*02:06	9	55	ic50	0.630	0.770	0.510	0.848	0.537	0.795
	1028790	A*02:06	10	35	ic50	0.572	0.768	0.525	0.680	0.502	0.700
	1028790	A*68:02	9	55	ic50	0.534	0.806	0.482	0.713	0.545	0.889
	1028790	A*68:02	10	35	ic50	0.272	0.620	0.591	0.813	0.533	1.000

Table 1. Performance comparison of NetMHCpan, sNebula and Nebula on IEDB benchmark datasets. The datasets along with the performance metrics of NetMHCpan were harvested from IEDB automatic server benchmark page (http://tools.iedb.org/auto_bench/mhci/weekly/). SRCC stands for Spearman's Rank Correlation Coefficient and AUC stands for area under the receiver operating characteristic curve. The "–" mark means not applicable.

as B*07:02, B*15:02 and B*27:04 in terms of AUC values, sNebula had better results on datasets such as C*03:04 and A*02:06. As a comparison, Nebula also had high AUC values on some datasets such as A*68:02 and B*38:01. However, because Nebula could not make predictions on HLAs and peptides that are not included in the training network, the results of Nebula were not complete for some datasets as B*27:04 and B*27:05.

Discussion

The human HLA loci are in a genomic region that is among the most polymorphic. The HLA loci have retained much variation[28–30]. Thousands of HLA alleles have been discovered, including approximately 9000 alleles of Class I HLAs[20,31]. The proteins encoded by HLAs are used by the immune system to recognize invaders such as foreign pathogens. However, the proteins themselves are not able to display biological functions. The binding groove of HLA proteins holds a peptide that can exhibit functions of HLAs such as social recognition skills[32]. It follows that knowledge of HLA-peptide binding plays a key role in understanding related biomedical questions such as autoimmune diseases and HLA-mediated adverse drug reactions[33,34]. Many *in vitro* experiments have been designed to assay HLA-peptide binding[35]. However, due to the huge number of possible binding interactions between thousands of HLAs and millions of peptides, it is difficult, if not impossible, to comprehensively ascertain the binding interactions between HLAs and peptides. Thus, computational methods can play a crucial role in the study of HLA-peptide binding. Though some computational approaches have been proposed for prediction of HLA-peptide binding[21,36], the practicability is limited since many methods do not support HLAs with few binding peptides or peptides that are diverse in length. Some recently developed methods such as NetMHCpan[25–27], NetMHC[37] and kernel functions[25,38] can predict for peptides with different lengths; however, extra processes[39,40] are usually required to identify core binding sequences within the peptides so that they can be converted to a fixed length. Though such extra processes may not necessarily reduce performance of such methods, algorithms that can overcome some restrictions of the current computational methods and handle peptides with different lengths are expected to have wider applications. Using sNebula, one can generate a comprehensive atlas of binding interactions between HLAs and peptides. Based on bipartite network analysis, sNebula has no limitations on the number of HLA molecules used or the size of the peptides utilized for training and, thus, provides a promising solution for the construction of a comprehensive atlas of HLA-peptide binding. However, different from machine learning-based methods such as NetMHCpan[25–27], sNebula is unable to directly predict HLA-peptide binding if neither the peptide nor the HLA exists in the training network.

The results of this current study suggest that sNebula can accurately predict the binding activity between HLAs and peptides, even though this is a very sparse dataset. The algorithm is useful because it can not only make predictions for untested peptides and HLAs given sequence information, but also can make more accurate predictions with a higher confidence. In addition, it does not set any limitation on the peptide length or the number of HLA alleles. With all these advantages, sNebula can help us study the binding between HLAs and peptides and

improve our understanding of the immune system. Like HLA-peptide binding data, a lot of big data are diverse and incomplete[41] such as gene expression data[42,43], drug-target binding[44] and clinical information[45]. Methods have been developed to impute the missing values for analysis including unsupervised and supervised classifications[43,46,47]. However, unlike the classification models, sNebula can deal with sparse or incomplete data without requiring the process of missing data imputation. It also accepts the diversity and flexibility of data so an assured length of features is not required. With the arising of big data era and increasing needs of big data analysis, we believe sNebula is one of the possible solutions to deal with large, diverse and incomplete data for predictions and novel discoveries.

Future applications of sNebula remain to be explored. In this study, sequences were utilized to calculate the similarity between nodes. It is possible to use sNebula in the development of similar algorithms for other applications. For example, it is possible to utilize the 2D structural fingerprints of drugs to replace sequences of peptides for similarity calculation and, thus, modify sNebula to predict drug-HLA binding or even drug-target binding that may underlie some observed genetic links to adverse drug events. Network-based inference (NBI) is a powerful network approach that can integrate a variety of data sources for a wide spectrum of applications such as drug-target predictions[48,49], drug safety assessment[50–52], driver mutations prioritization in cancer genomics[53], RNA network prediction[54] and xenobiotics gene and disease prediction[55]. Cheng *et al.* utilized Node Weighted Network-based Inference (NWNBI) to predict drug-target interactions using a node-weighted network and observed a better performance than the unweighted NBI[49,56]. They calculated the drug similarity by 2D fingerprints and target similarity by sequences. However, their method uses all the neighbors for prediction instead of selecting the top similar ones. It is possible to improve the prediction performance by selecting top similar neighbors using sNebula. Another possible application for sNebula is to predict drug-disease association for drug repurposing. Gottlieb *et al.* collected a network of drug-disease associations as well as information of drug-drug similarities and disease-disease similarities to predict novel drug-disease associations using logistic regression[57]. The machine learning method is useful; however, there are some challenges such as problems to deal with a flexible length of features[21] or missing data[45]. Since sNebula is based on similarity and does not require the completeness or an assured length of features, it is possible to extend sNebula to predict drug-disease associations while overcoming those problems.

Another potential application of sNebula is to develop new therapeutics such as tumor immunotherapy. The neoantigens are peptides in the human body that are not encoded by the normal human genome. In tumors, they are generated by the tumor-specific DNA alternations[58]. When the gene expression data for patients are available, predicting HLA-peptide binding may help to identify or filter patient-specific neoantigens, which are a major factor for clinical immunotherapy development[58,59]. As more HLA-peptide binding data and patient-specific RNA sequencing data are becoming available, we believe sNebula can potentially help with neoantigen identification and the development of immunotherapies.

In addition to predictions values, sNebula also provides the confidence values. Confidence values are estimations about how likely the result is true; therefore, users can differentiate the results using confidence values and select the most confident predictions for validation. A good method not only makes more predictions in number, but also predicts with higher accuracy at higher confidence. From the confidence analysis result of sNebula, we saw sNebula predicted more and performed better with higher confidence. We believe the confidence values are useful information that can potentially help with the selection of prediction results for experimental validation in applications such as HLA-peptide binding, drug-target binding and drug-disease associations.

Conclusion

We developed a network-based prediction algorithm, sNebula, to predict the binding potential between HLAs and peptides. We found this algorithm exhibited a good performance in both the LOO cross-validation and five-fold cross-validations using the experimental HLA-peptide binding data curated from major databases. The confidence analysis indicated its ability to make predictions with more accuracy when the confidence level is higher. This algorithm not only overcomes the limitations of the current machine learning methods on the number of HLAs and lengths of peptides, but also makes it possible to predict HLA-peptide binding for new peptides or HLAs. It could be expected that sNebula can help with the construction of a comprehensive atlas of HLA-peptide binding that, in turn, facilitates better understanding of the immune system.

Methods

Study design. The workflow of the study is shown in Fig. 4. Qualitative Class I HLA-peptide binding data were collected and curated from four databases: AntiJen[16], IEDB[17], MHCBN[18] and SYFPEITHI[19]. A bipartite network of HLA-peptide binding data was then constructed. The binding data network was used to assess the performance of sNebula using leave-one-out (LOO) cross-validation and 1,000 iterations of five-fold cross-validations. The prediction confidence analysis was conducted based on the results of five-fold cross-validations.

Data curation. The experimental Class I HLA-peptide binding data were collected from four databases (AntiJen[16], IEDB[17], MHCBN[18] and SYFPEITHI[19]) as described in our previous study[22]. The databases provide qualitative binding categories (binding versus non-binding) for each HLA-peptide pair. We merged the four databases and recorded only one qualitative datum for each of HLA-peptide pairs using the majority voting strategy[22]. Since sNebula is based on a color graph of the network in which the nodes (HLAs and peptides) are colored using their amino acid sequences, we downloaded HLA sequences from the IMGT/HLA database[20] and removed the HLAs that do not have an affirmative protein sequence such as HLA serotypes and allele groups and their binding peptides. The curated HLA-peptide binding data were used to construct a bipartite binding network, where HLAs and peptides are nodes, and the binding data between them, either binding or non-binding, are edges.

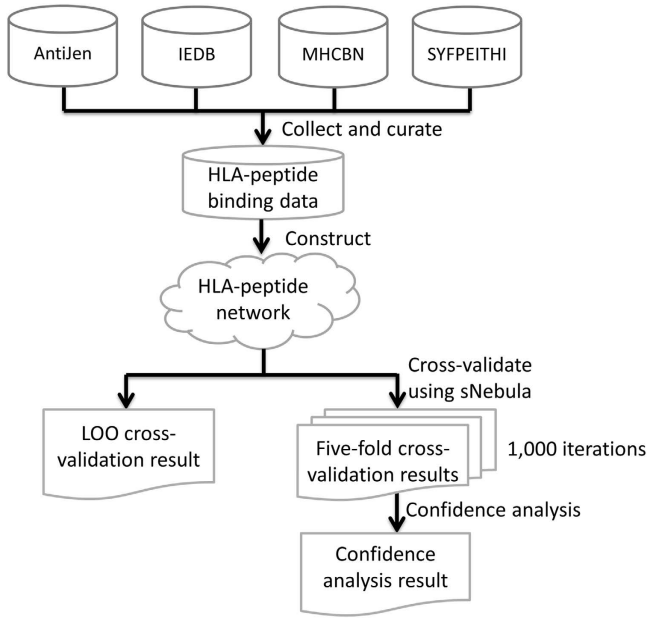

Figure 4. Study workflow. The qualitative Class I HLA-peptide binding data were curated from four major databases (AntiJen, IEDB, MHCBN and SYFPEITHI). A HLA-peptide binding data network was constructed based on the curated data. To assess the performance of sNebula, LOO cross-validation and 1,000 iterations of five-fold cross-validations were executed. Based on the results of five-fold cross-validations, confidence analysis was conducted to evaluate the relationship between the confidence levels and the prediction performance of sNebula.

sNebula. To predict the binding between peptide p_x and HLA h_y, sNebula first identifies the peptides that are connected to HLA h_y, p_i ($i = 1, 2, \ldots$). The similarity between peptide p_x and each of the connected peptides of h_y, p_i, is calculated based on their sequences. To calculate the similarity between peptide p_x and peptide p_i, sNebula first aligns the sequences of the two peptides without opening gaps and then calculates the similarity score $sp_{x,i}$ using equations (1–2).

$$sp_{x,i} = \frac{p_{x,i}}{p_{x,x}} e^{-l_{x,i}} \tag{1}$$

$$l_{x,i} = \frac{|l_i - l_x|}{l_x} \tag{2}$$

In equation (2), l_x and l_i are the lengths of peptides p_x and p_i, respectively. In equation (1), $p_{x,i}$ is the sequence similarity between peptides p_x and p_i and is calculated using BLOSUM50 matrix[60], and $p_{x,x}$ is the sequence similarity between peptide p_x and itself. BLOSUM50 matrix was used in this study because it has been used in many other methods for predicting HLA-peptide binding[26,39,61,62]. For each of possible alignments between peptide p_x and peptide p_i, a sequence similarity value, $p_{x,i}$, is calculated. The alignment with the largest sequence similarity value is then selected and its similarity, $p_{x,i}$, is used for the calculation of equation (1).

In the same way, sNebula recognizes the HLAs that are connected to peptide p_x, h_j ($j = 1, 2, \ldots$). The similarity between HLA h_y and each of the connected HLAs of peptide p_x, h_j, is then calculated. For the HLA similarity calculation, instead of using HLA full sequences, sNebula utilizes the 48 unique residues on HLAs that closely interact with peptides which were identified by Chelvanayagam[63] as HLA pseudo-sequences. The pseudo-sequence similarity score $sh_{y,j}$ between HLA h_y and HLA h_j is calculated using equation (3).

$$sh_{y,j} = \frac{h_{y,j}}{h_{y,y}} \tag{3}$$

In equation (3), $h_{y,j}$ is the pseudo-sequence similarity between HLA h_y and HLA h_j and is calculated using BLOSUM50 matrix, and $h_{y,y}$ is the pseudo-sequence similarity between HLA h_y and itself. It is noted that both $sp_{x,i}$ and $sh_{y,j}$ are directional; thus, $sp_{x,I}$ ($sh_{y,j}$) is not necessarily equal to $sp_{j,x}$ ($sh_{j,y}$).

After the similarity scores for the nodes (HLAs and peptides) of the neighbor edges are calculated, sNebula ranks the peptides and HLAs using their similarity scores $sp_{x,i}$ and $sh_{y,j}$ to select n top ranked peptides and HLAs for p_x and h_y, respectively, for the calculation of a continuous value $p_{x,y}$ using equation (4) as the prediction of binding between p_x and h_y. Here n is a parameter to be determined. The n with the best prediction of accuracy of LOO cross-validation based on the network of HLA-peptide binding data is used.

$$p_{x,y} = \frac{\frac{\sum_{i=1}^{n} sp_{x,i} e_{i,y}}{\sum_{i=1}^{n} sp_{x,i}} + \frac{\sum_{j=1}^{n} sh_{y,j} e_{x,j}}{\sum_{j=1}^{n} sh_{y,j}}}{2} \quad (4)$$

In equation (4), $e_{i,y}$ is the binding edge weight between peptide i and HLA y in the HLA-peptide binding data network given by equation (5).

$$e_{i,y} = \begin{cases} 1, & \textit{if binding between peptide i and HLA y} \\ -1, & \textit{if non-binding between peptide i and HLA y} \end{cases} \quad (5)$$

As an unbiased approach, sNebula considers the contribution from the peptides p_x and HLAs h_y of the neighbor edges equally. When peptide p_x do not contain neighbor edges, sNebula predicts the binding between peptide p_x and HLA h_y using equation (6).

$$p_{x,y} = \frac{\sum_{i=1}^{n} sp_{x,i} e_{i,y}}{\sum_{i=1}^{n} sp_{x,i}} \quad (6)$$

Therefore, if a peptide has no experiment data against any HLA in the training network, sNebula is still able to predict its binding towards the HLAs based on its sequence and the topological feature of the training network using equation (6). In the same way, if HLA h_y do not have neighbor edges, sNebula predicts the binding between peptide p_x and HLA h_y using equation (7).

$$p_{x,y} = \frac{\sum_{j=1}^{n} sh_{y,j} e_{x,j}}{\sum_{j=1}^{n} sh_{y,j}} \quad (7)$$

When the number of available connecting nodes (HLAs or peptides) for peptide p_x or HLA h_y is less than n, all of the nodes are used. When multiple connecting nodes for peptide p_x or HLA h_y have the same similarity score to be selected as the top n similar nodes, the average of their binding edge weights is used in equations (4), (6) and (7).

The binding prediction $p_{x,y}$ is a continuous value between −1 and 1 and is converted into a categorical prediction value $c_{x,y}$ using equation (8).

$$c_{x,y} = \begin{cases} \textit{binding}, & \textit{if } p_{x,y} \geq 0 \\ \textit{non-binding}, & \textit{if } p_{x,y} < 0 \end{cases} \quad (8)$$

The goodness of the prediction $c_{x,y}$ is assessed using a metric term as prediction confidence $conf_{x,y}$ that is defined in equation (9).

$$conf_{x,y} = \left| p_{x,y} \right| \quad (9)$$

LOO cross-validation. In the LOO cross-validation, each of the HLA-peptide binding data was taken out and the remaining HLA-peptide binding data were used to construct the HLA-peptide binding network for prediction of the binding value for the taken-out HLA-peptide pair using sNebula. This process was repeated until every HLA-peptide binding data were used as a test sample. The predicted values were then compared with the actual experimental binding data and sensitivity, specificity, accuracy and area under receiver operating characteristic curve (AUC) were calculated to evaluate the performance of sNebula. To determine the parameter n, we repeated the LOO cross-validation for $n = 1, 2, 3 \ldots 50$. The n value with the highest prediction accuracy of LOO cross-validation was selected to be the parameter for sNebula.

Five-fold cross-validations. We conducted 1,000 iterations of five-fold cross-validations on the HLA-peptide binding data to obtain a statistically robust estimation of sNebula performance. In a five-fold cross-validation, the HLA-peptide binding data were randomly divided into five parts as equal as possible. One part of HLA-peptide binding data was taken out to be used as test samples and the remaining four parts of HLA-peptide binding data were used as the training samples to construct a bipartite network. LOO cross-validations were conducted on training samples to determine the parameter n in sNebula. The network constructed from the training samples and the determined parameter n was used by sNebula to predict HLA-peptide binding of the test samples. This process was repeated five times so that each of the five parts of the HLA-peptide binding data was used once and only once as test samples. The categorical prediction results from all the five folds of test samples were compared to the actual experimental HLA-peptide binding data to calculate the sensitivity, specificity and accuracy to estimate the performance of sNebula.

Prediction confidence analysis. Prediction confidence has been proposed as one of the metrics to measure performance of predictive models developed in the FDA's endocrine disruptors knowledge based project[64–71] using variety of machine learning methods such as decision tree[72], Decision Forest models[73–78] based on molecular descriptors[79] that are calculated using the algorithm developed for the expert systems of structure elucidation[80–85], support vector machine[86,87] and principal component analysis based algorithm[88,89]. The continuous value output from sNebula for prediction of binding between an HLA and a peptide is the measure of likelihood

of the peptide is a binder or non-binder of the HLA and indicates the confidence for the prediction. A good prediction method should not only show an overall high prediction accuracy but also is expected to 1) predict most unknown samples with a high confidence and 2) show a higher accuracy for the predictions with a higher confidence than the predictions with a lower confidence. We examined the relationship between prediction confidence and accuracy using all predictions from the 1,000 iterations of five-fold cross-validations. The prediction confidence was calculated using equation (9) for every prediction. The predictions were then placed into 10 groups with even confidence bins according to their confidence values. For each of the 10 groups of predictions, we calculated the sensitivity, specificity, accuracy and AUC by comparing the predictions with the actual experimental HLA-peptide binding data. At last, the performance of sNebula at difference confidence levels was analyzed.

Benchmark. IEDB has an automatic server benchmark page (http://tools.iedb.org/auto_bench/mhci/weekly/) that evaluates different prediction methods for HLA-peptide binding based on new data submitted in the past three months or last week. We used the latest three datasets of 3-month period (2016-05-03, 2016-02-19 and 2015-08-07) as an example to compare sNebula with existing methods as well as its predecessor, Nebula. The parameter n was set to 13 for sNebula to make predictions. Two performance metrics, Spearman's Rank Correlation Coefficient (SRCC) and AUC, were calculated between the predicted values and experimental values using R.

References

1. Bushkin, Y., Demaria, S., Le, J. M. & Schwab, R. Physical association between the CD8 and HLA class I molecules on the surface of activated human T lymphocytes. *Proc. Natl. Acad. Sci. USA* **85,** 3985–3989 (1988).
2. Poncet, P., Arock, M. & David, B. MHC class II-dependent activation of CD4+ T cell hybridomas by human mast cells through superantigen presentation. *J. Leukoc. Biol.* **66,** 105–112 (1999).
3. Jin, P. & Wang, E. Polymorphism in clinical immunology - From HLA typing to immunogenetic profiling. *J. Transl. Med.* **1,** 8, doi: 10.1186/1479-5876-1-8 (2003).
4. Trowsdale, J. The MHC, disease and selection. *Immunol. Lett.* **137,** 1–8, doi: 10.1016/j.imlet.2011.01.002 (2011).
5. Illing, P. T., Vivian, J. P., Purcell, A. W., Rossjohn, J. & McCluskey, J. Human leukocyte antigen-associated drug hypersensitivity. *Curr. Opin. Immunol.* **25,** 81–89, doi: 10.1016/j.coi.2012.10.002 (2013).
6. Luo, H. *et al.* Molecular docking to identify associations between drugs and class I human leukocyte antigens for predicting idiosyncratic drug reactions. *Comb. Chem. High Throughput Screen.* **18,** 296–304 (2015).
7. Saper, M. A., Bjorkman, P. J. & Wiley, D. C. Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution. *J. Mol. Biol.* **219,** 277–319 (1991).
8. Stern, L. J. *et al.* Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* **368,** 215–221, doi: 10.1038/368215a0 (1994).
9. Sim, E. & Cross, S. J. Phenotyping of human complement component C4, a class-III HLA antigen. *Biochem. J.* **239,** 763–767 (1986).
10. Villadangos, J. A. & Schnorrer, P. Intrinsic and cooperative antigen-presenting functions of dendritic-cell subsets *in vivo*. *Nat. Rev. Immunol.* **7,** 543–555, doi: 10.1038/nri2103 (2007).
11. Felix, N. J. & Allen, P. M. Specificity of T-cell alloreactivity. *Nat. Rev. Immunol.* **7,** 942–953, doi: 10.1038/nri2200 (2007).
12. Gebe, J. A., Swanson, E. & Kwok, W. W. HLA class II peptide-binding and autoimmunity. *Tissue Antigens* **59,** 78–87 (2002).
13. Illing, P. T. *et al.* Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature* **486,** 554–558, doi: 10.1038/nature11147 (2012).
14. van der Burg, S. H., Bijker, M. S., Welters, M. J., Offringa, R. & Melief, C. J. Improved peptide vaccine strategies, creating synthetic artificial infections to maximize immune efficacy. *Adv Drug Deliv Rev* **58,** 916–930, doi: 10.1016/j.addr.2005.11.003 (2006).
15. Chirino, A. J., Ary, M. L. & Marshall, S. A. Minimizing the immunogenicity of protein therapeutics. *Drug Discov. Today* **9,** 82–90, doi: 10.1016/S1359-6446(03)02953-2 (2004).
16. Toseland, C. P. *et al.* AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.* **1,** 4, doi: 10.1186/1745-7580-1-4 (2005).
17. Vita, R. *et al.* The immune epitope database 2.0. *Nucleic Acids Res.* **38,** D854–D862, doi: 10.1093/nar/gkp1004 (2010).
18. Lata, S., Bhasin, M. & Raghava, G. P. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res. Notes* **2,** 61, doi: 10.1186/1756-0500-2-61 (2009).
19. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. & Stevanovic, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50,** 213–219 (1999).
20. Robinson, J. *et al.* The IMGT/HLA database. *Nucleic Acids Res.* **41,** D1222–D1227, doi: 10.1093/nar/gks949 (2013).
21. Luo, H. *et al.* Machine learning methods for predicting HLA-peptide binding activity. *Bioinform. Biol. Insights* **9,** 21–29 doi: 10.4137/BBI.S29466 (2015).
22. Luo, H. *et al.* Understanding and predicting binding between human leukocyte antigens (HLAs) and peptides by network analysis. *BMC Bioinformatics* **16** Suppl 13, S9, doi: 10.1186/1471-2105-16-S13-S9 (2015).
23. Ye, H. *et al.* Applying network analysis and Nebula (neighbor-edges based and unbiased leverage algorithm) to ToxCast data. *Environ. Int.* **89–90,** 81–92, doi: 10.1016/j.envint.2016.01.010 (2016).
24. Yewdell, J. W. & Bennink, J. R. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu. Rev. Immunol.* **17,** 51–88, doi: 10.1146/annurev.immunol.17.1.51 (1999).
25. Jacob, L. & Vert, J. P. Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics* **24,** 358–366, doi: 10.1093/bioinformatics/btm611 (2008).
26. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61,** 1–13, doi: 10.1007/s00251-008-0341-z (2009).
27. Zhang, L., Udaka, K., Mamitsuka, H. & Zhu, S. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Briefings in bioinformatics* **13,** 350–364, doi: 10.1093/bib/bbr060 (2012).
28. Parham, P. & Ohta, T. Population biology of antigen presentation by MHC class I molecules. *Science* **272,** 67–74 (1996).
29. Apanius, V., Penn, D., Slev, P. R., Ruff, L. R. & Potts, W. K. The nature of selection on the major histocompatibility complex. *Crit. Rev. Immunol.* **17,** 179–224 (1997).
30. Castro-Prieto, A., Wachter, B. & Sommer, S. Cheetah paradigm revisited: MHC diversity in the world's largest free-ranging population. *Mol. Biol. Evol.* **28,** 1455–1468, doi: 10.1093/molbev/msq330 (2011).
31. Marsh, S. G. *et al.* Nomenclature for factors of the HLA system, 2004. *Tissue Antigens* **65,** 301–369, doi: 10.1111/j.1399-0039.2005.00379.x (2005).
32. Boehm, T. & Zufall, F. MHC peptides and the sensory evaluation of genotype. *Trends Neurosci.* **29,** 100–107, doi: 10.1016/j.tins.2005.11.006 (2006).

33. Kongkaew, S. *et al.* Molecular Dynamics Simulation Reveals the Selective Binding of Human Leukocyte Antigen Alleles Associated with Behcet's Disease. *PLoS One* **10,** e0135575, doi: 10.1371/journal.pone.0135575 (2015).
34. Le Clerc, S. *et al.* A double amino-acid change in the HLA-A peptide-binding groove is associated with response to psychotropic treatment in patients with schizophrenia. *Transl Psychiatry* **5,** e608, doi: 10.1038/tp.2015.97 (2015).
35. Yamada, E. *et al.* Identification of a naturally processed HLA-Cw7-binding peptide that cross-reacts with HLA-A24-restricted ovarian cancer-specific CTLs. *Tissue Antigens* **86,** 164–171, doi: 10.1111/tan.12607 (2015).
36. Ali, M. T., Morshed, M. M. & Hassan, F. A. Computational Approach for Designing a Universal Epitope-Based Peptide Vaccine Against Nipah Virus. *Interdiscip Sci* **7,** 177–185, doi: 10.1007/s12539-015-0023-0 (2015).
37. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32,** 511–517, doi: 10.1093/bioinformatics/btv639 (2016).
38. Salomon, J. & Flower, D. R. Predicting Class II MHC-Peptide binding: a kernel based approach using similarity scores. *BMC Bioinformatics* **7,** 501, doi: 10.1186/1471-2105-7-501 (2006).
39. Andreatta, M., Schafer-Nielsen, C., Lund, O., Buus, S. & Nielsen, M. NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. *PLoS One* **6,** e26781, doi: 10.1371/journal.pone.0026781 (2011).
40. Nielsen, M. & Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* **10,** 296, doi: 10.1186/1471-2105-10-296 (2009).
41. Slavakis, K., Giannakis, G. & Mateos, G. Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. *Signal Processing Magazine, IEEE* **31,** 18–31 (2014).
42. Yu, D. *et al.* Permutation test for incomplete paired data with application to cDNA microarray data. *Comput. Stat. Data Anal.* **56,** 510–521 (2012).
43. Liew, A. W., Law, N. F. & Yan, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinform* **12,** 498–513, doi: 10.1093/bib/bbq080 (2011).
44. Chen, B., Ding, Y. & Wild, D. J. Assessing drug target association using semantic linked data. *PLoS Comput. Biol.* **8,** e1002574, doi: 10.1371/journal.pcbi.1002574 (2012).
45. Jerez, J. M. *et al.* Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif. Intell. Med.* **50,** 105–115, doi: 10.1016/j.artmed.2010.05.002 (2010).
46. Stekhoven, D. J. & Buhlmann, P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118, doi: 10.1093/bioinformatics/btr597 (2012).
47. Moorthy, K., Saberi Mohamad, M. & Deris, S. A review on missing value imputation algorithms for microarray gene expression data. *Current Bioinformatics* **9,** 18–22 (2014).
48. Wu, Z. *et al.* SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. *Briefings in bioinformatics*, doi: 10.1093/bib/bbw012 (2016).
49. Cheng, F. *et al.* Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology* **8,** e1002503, doi: 10.1371/journal.pcbi.1002503 (2012).
50. Zhang, C., Hong, H., Mendrick, D. L., Tang, Y. & Cheng, F. Biomarker-based drug safety assessment in the age of systems pharmacology: from foundational to regulatory science. *Biomarkers in medicine* **9,** 1241–1252, doi: 10.2217/bmm.15.81 (2015).
51. Cheng, F. *et al.* Adverse drug events: database construction and in silico prediction. *Journal of chemical information and modeling* **53,** 744–752, doi: 10.1021/ci4000079 (2013).
52. Cheng, F. *et al.* Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *Journal of chemical information and modeling* **53,** 753–762, doi: 10.1021/ci400010x (2013).
53. Cheng, F., Zhao, J. & Zhao, Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in bioinformatics*, doi: 10.1093/bib/bbv068 (2015).
54. Li, J. *et al.* Computational prediction of microRNA networks incorporating environmental toxicity and disease etiology. *Scientific reports* **4,** 5576, doi: 10.1038/srep05576 (2014).
55. Cheng, F. *et al.* Prediction of human genes and diseases targeted by xenobiotics using predictive toxicogenomic-derived models (PTDMs). *Molecular bioSystems* **9,** 1316–1325, doi: 10.1039/c3mb25309k (2013).
56. Cheng, F., Zhou, Y., Li, W., Liu, G. & Tang, Y. Prediction of chemical-protein interactions network with weighted network-based inference method. *PloS one* **7,** e41064, doi: 10.1371/journal.pone.0041064 (2012).
57. Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* **7,** 496, doi: 10.1038/msb.2011.26 (2011).
58. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348,** 69–74, doi: 10.1126/science.aaa4971 (2015).
59. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515,** 572–576, doi: 10.1038/nature14001 (2014).
60. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919 (1992).
61. Lundegaard, C., Lund, O. & Nielsen, M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* **24,** 1397–1398, doi: 10.1093/bioinformatics/btn128 (2008).
62. Karosiene, E. *et al.* NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* **65,** 711–724, doi: 10.1007/s00251-013-0720-y (2013).
63. Chelvanayagam, G. A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics* **45,** 15–26 (1996).
64. Hong, H. *et al.* Rat alpha-Fetoprotein binding affinities of a large set of structurally diverse chemicals elucidated the relationships between structures and binding affinities. *Chem. Res. Toxicol.* **25,** 2553–2566, doi: 10.1021/tx3003406 (2012).
65. Hong, H. *et al.* Human sex hormone-binding globulin binding affinities of 125 structurally diverse chemicals and comparison with their binding to androgen receptor, estrogen receptor, and alpha-fetoprotein. *Toxicol. Sci.* **143,** 333–348, doi: 10.1093/toxsci/kfu231 (2015).
66. Shen, J. *et al.* Homology modeling, molecular docking, and molecular dynamics simulations elucidated alpha-fetoprotein binding modes. *BMC Bioinformatics* **14** Suppl 14, S6, doi: 10.1186/1471-2105-14-S14-S6 (2013).
67. Ding, D. *et al.* The EDKB: an established knowledge base for endocrine disrupting chemicals. *BMC Bioinformatics* **11** Suppl 6, S5, doi: 10.1186/1471-2105-11-S6-S5 (2010).
68. Shen, J. *et al.* EADB: an estrogenic activity database for assessing potential endocrine activity. *Toxicol. Sci.* **135,** 277–291, doi: 10.1093/toxsci/kft164 (2013).
69. Ng, H. W., Perkins, R., Tong, W. & Hong, H. Versatility or promiscuity: the estrogen receptors, control of ligand selectivity and an update on subtype selective ligands. *Int. J. Environ. Res. Public Health* **11,** 8709–8742, doi: 10.3390/ijerph110908709 (2014).
70. Tong, W. *et al.* Assessing QSAR limitations-A regulatory perspective. *Curr. Comput. Aided Drug Des.* **1,** 195–205 (2005).
71. Ng, H. W. *et al.* Estrogenic activity data extraction and in silico prediction show the endocrine disruption potential of bisphenol A replacement compounds. *Chem. Res. Toxicol.* **28,** 1784–1795, doi: 10.1021/acs.chemrestox.5b00243 (2015).
72. Hong, H. *et al.* Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environ. Health Perspect.* **110,** 29–36 (2002).

73. Hong, H. *et al.* Multiclass decision forest-a novel pattern recognition method for multiclass classification in microarray data analysis. *DNA Cell Biol.* **23,** 685–694 (2004).
74. Hong, H., Tong, W., Xie, Q., Fang, H. & Perkins, R. An in silico ensemble method for lead discovery: decision forest. *SAR QSAR Environ. Res.* **16,** 339–347, doi: 10.1080/10659360500203022 (2005).
75. Tong, W., Hong, H., Fang, H., Xie, Q. & Perkins, R. Decision forest: combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **43,** 525–531, doi: 10.1021/ci020058s (2003).
76. Tong, W. *et al.* Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence. *Environ. Health Perspect.* **112,** 1622–1627 (2004).
77. Xie, Q. *et al.* Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. *BMC Bioinformatics* **6** Suppl 2, S4, doi: 10.1186/1471-2105-6-s2-s4 (2005).
78. Ng, H. W. *et al.* Development and Validation of Decision Forest Model for Estrogen Receptor Binding Prediction of Chemicals Using Large Data Sets. *Chemical research in toxicology* **28,** 2343–2351, doi: 10.1021/acs.chemrestox.5b00358 (2015).
79. Hong, H. *et al.* Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of chemical information and modeling* **48,** 1337–1344, doi: 10.1021/ci800038f (2008).
80. Hong, H. & Xin, X. ESSESA: an expert system for elucidation of structures from spectra. 1. Knowledge base of infrared spectra and analysis and interpretation programs. *J. Chem. Inf. Comput. Sci.* **30,** 203–210 (1990).
81. Hong, H. & Xin, X. ESSESA, an expert system for structure elucidation from spectral analysis: Part II. Novel algorithm of perception of the linear independent smallest set of smallest rings. *Anal. Chim. Acta* **262,** 179–191 (1992).
82. Hong, H. & Xin, X. ESSESA: An expert system for structure elucidation from spectra. 3. LNSCS for chemical knowledge representation. *J. Chem. Inf. Comput. Sci.* **32,** 116–120 (1992).
83. Hong, H. & Xin, X. ESSESA: An Expert System for Structure Elucidation from Spectra. 4. Canonical Representation of Structures. *J. Chem. Inf. Comput. Sci.* **34,** 730–734 (1994).
84. Hong, H. & Xin, X. ESSESA: An Expert System for Structure Elucidation from Spectra. 5. Substructure Constraints from Analysis of First-Order 1H-NMR Spectra. *J. Chem. Inf. Comput. Sci.* **34,** 1259–1266 (1994).
85. Masui, H. & Hong, H. Spec2D: a structure elucidation system based on 1H NMR and H-H COSY spectra in organic chemistry. *J. Chem. Inf. Model.* **46,** 775–787, doi: 10.1021/ci0502810 (2006).
86. Hong, H. *et al.* The accurate prediction of protein family from amino acid sequence by measuring features of sequence fragments. *J. Comput. Biol.* **16,** 1671–1688, doi: 10.1089/cmb.2008.0115 (2009).
87. Liu, J. *et al.* Predicting hepatotoxicity using ToxCast *in vitro* bioactivity and chemical structure. *Chem. Res. Toxicol.* **28,** 738–751, doi: 10.1021/tx500501h (2015).
88. Hong, H. *et al.* Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. *SAR QSAR Environ. Res.* **14,** 373–388, doi: 10.1080/10629360310001623962 (2003).
89. Su, Z. *et al.* Consensus analysis of multiple classifiers using non-repetitive variables: diagnostic application to microarray gene expression data. *Comput. Biol. Chem.* **31,** 48–56, doi: 10.1016/j.compbiolchem.2007.01.001 (2007).

Acknowledgements

This research was supported in part by an appointment to the Research Participation Program at the National Center for Toxicological Research (Heng Luo, Hao Ye, Hui Wen Ng and Sugunadevi Sakkiah) administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. This project was partially supported by grant funding from the National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) (P20 GM103429) (formerly P20RR016460). The findings and conclusions in this article have not been formally disseminated by the US Food and Drug Administration (FDA) and should not be construed to represent the FDA determination or policy.

Author Contributions

H.H. and D.L.M. designed and led the project. H.L. collected the data and implemented the methods. H.L., H.Y., H.W.N., S.S. and H.H. discussed the data analysis and the results. H.L., H.H. and D.L.M. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Luo, H. *et al.* sNebula, a network-based algorithm to predict binding between human leukocyte antigens and peptides. *Sci. Rep.* **6**, 32115; doi: 10.1038/srep32115 (2016).