FEBS Letters xxx (2011) xxx–xxx

journal homepage: www.FEBSLetters.org

Recognition of the ligand-type specificity of classical and non-classical MHC I proteins

Eduardo Martínez-Naves [a], Esther M. Lafuente [a], Pedro A. Reche [a,b,*]

[a] Department of Microbiology I-Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N, Madrid 28040, Spain
[b] Laboratory of Immunomedicine, Department of Microbiology I-Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N, Madrid 28040, Spain

ARTICLE INFO

Article history:
Received 2 August 2011
Revised 28 September 2011
Accepted 3 October 2011
Available online xxxx

Edited by Takashi Gojobori

Keywords:
Classical MHC class I molecule
Non-classical MHC class I molecule
Machine learning
Ligand
Prediction

ABSTRACT

Functional characterization of proteins belonging to the MHC I superfamily involves knowing their cognate ligands, which can be peptides, lipids or none. However, the experimental identification of these ligands is not an easy task and generally requires some a priori knowledge of their chemical nature (ligand-type specificity). Here, we trained k-nearest neighbor and support vector machine classifiers that predict the ligand-type specificity MHC I proteins with great accuracy. Moreover, we applied these classifiers to human and mouse MHC I proteins of uncharacterized ligands, obtaining some results that can be instrumental to unravel the function of these proteins.

1. Introduction

The major histocompatibility complex class I (MHC I) protein superfamily encompasses a large number of glycoproteins including classical MHC I molecules and non-classical and MHC I-like molecules [1]. Because of their crucial role in graft rejection and antigen presentation, classical MHC I molecules (hereafter MHC Ia) were the first to be discovered and studied. In humans, classical MHC molecules are for historical reasons, known as human leukocyte antigens (HLAs).

MHC Ia molecules are cell surface expressed glycoproteins consisting of an α chain (encoded inside the MHC gene region) paired with β2-microglobulin (β2m), and their function is to present peptides for immune recognition by CD8 T cells [2]. MHC I-bound peptides are nested in the α1α2 domain of the MHC I molecule (MHC I α1α2 domain). The structural presence of this α1α2 domain is the signature that relates all the members of the MHC I superfamily.

Non-classical and MHC I-like molecules (hereafter MHC Ib molecules) were discovered later and differ in many aspects from MHC Ia molecules. First of all, MHC Ib molecules do not conform a single protein family but comprise several highly divergent protein families that are structurally related to MHC Ia molecules. MHC Ib molecules can be encoded inside or outside the MHC loci, display a wide range of functions and, in contrast to MHC Ia molecules, are either non-polymorphic or exhibit little polymorphism [3,4]. Moreover, while MHC Ia molecules can only bind peptides, there are known examples of MHC Ib molecules that bind peptides (e.g., HLA-E and HLA-G, and their mouse functional counterparts Qa1 (H2-T23) and Qa2 (H2-Q9), respectively), others that bind lipids (e.g., CD1 antigens; ZAG, zinc-binding alpha-2-glycoprotein 1; EPCR, Endothelial Protein C Receptor) and some that do not have any ligand; they have an empty groove (e.g., MICA&B, mouse TL antigens, ULBP1,2,&3, HFE and FcRn) [4].

Functional characterization of MHC I proteins can be challenging and generally requires knowing their cognate ligands. However, the identification of ligands of MHC I proteins, if any, is a difficult task, which generally requires having some a priori knowledge of the chemical nature of the ligand (ligand-type specificity) to guide the experimental efforts. Therefore, in this study, we built machine learning-based classifiers to predict the ligand-type specificity of these proteins. Subsequently, we applied such models to a number of human and mouse MHC Ib molecules of uncharacterized ligands obtaining some interesting results, such as the binding of lipids of the H2-M1 family and the lack of ligand of MR1, that could be instrumental to unravel the function of these proteins.

Abbreviations: MHC I molecules, major histocompatibility complex class I molecules; MHC Ia molecules, classical MHC I molecules; MHC Ib molecules, non-classical MHC I and MHC I-like molecules

* Corresponding author at: Department of Microbiology I-Immunology, Facultad de Medicina, Universidad Complutense de Madrid, Ave Complutense S/N, Madrid 28040, Spain. Fax: +34 91 394 1641.
E-mail address: parecheg@med.ucm.es (P.A. Reche).

0014-5793/$36.00
doi:10.1016/j.febslet.2011.10.007

2. Materials and methods

2.1. MHC I dataset

We used a dataset (MHCI^{556} dataset) consisting of 556 non-overlapping and unique sequences of MHC I proteins of known ligand-type specificity of which 355 bind peptides (P), 84 bind lipids (L) and 117 do not bind anything (N). Sequences in the MHCI^{556} dataset were collected and processed as previously described [5]. The sequences only comprise the α1α2 domain and all range between 170 and 189 amino acids and were subjected to a sequence-similarity reduction schema, setting the maximum sequence similarity allowed between two sequences to an *e*-score of 925 using a BLOSUM62 matrix [5]. Sequence identity within the P, L and N groups is 63.69 ± 17.62, 41.8 ± 15.37 and 36.37 ± 21.07, respectively.

2.2. Building and evaluation of machine learning-based models

We used the Waikato Environment for Knowledge Analysis (WEKA) [6] to built and evaluate machine learning (ML)-based classification models. As input for features for training, we used the amino acid composition of the sequences (attributes) and their known ligand-type specificity (P: Peptides, L: Lipids and N: No ligand) (classification instances). WEKA provides a large collection of ML algorithms for classification, and in this study, we selected k-Nearest Neighbor algorithm (kNN) and support vector machines (SVMs) [7,8]. Briefly, kNN classifies objects based on the majority class of their *k* nearest neighbors in the training sets. The vicinity between objects is computed as a Euclidean distance. SVMs were first introduced to classify linear data and are based on decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. For non-linear data, SVMs first use a function (kernel) to map the input data onto a higher *m*-dimensional space, where a linear model based on decision planes can then achieve an optimal separation of the data. Here we used a Gaussian Radial Basis Function (*RBF*-kernel) and a Polynomial function (*P*-kernel), as kernels for SVMs.

Model refinement was achieved in 10-fold cross-validation experiments varying the relevant parameters of the ML algorithms. In a 10-fold cross-validation, the data are randomly partitioned into 10 sets, and each set is tested using classifiers trained on the sum of the remaining sets. Thus, kNN were refined with regard to *k*, the number of neighbours, while SVMs were refined with regard to *C*, the complexity parameter – allows one to trade off training error versus model complexity – in combination with γ for the *RBF*-kernel (defines de width of the Gaussian function) and *E* for the *P*-kernel (exponent of the polynomial function).

As measures of performance to evaluate the models, we used sensitivity (*SE*), specificity (*SP*) and accuracy (*ACC*) in percentages, which for a 3-class classification can be computed using Eqs. (1)–(3), respectively,

$$ACC = 100 \times \left(\frac{P_P + L_L + N_N}{P + L + N}\right) \quad (1)$$

$$SE = 100/3 \times \left(\frac{P_P}{P} + \frac{L_L}{L} + \frac{N_N}{N}\right) \quad (2)$$

$$SP = 100/3 \times \left[\left(\frac{N + L - (N_P + L_P)}{N + L}\right) + \left(\frac{N + P - (N_L + P_L)}{N + P}\right) + \left(\frac{P + L - (P_N + L_N)}{P + L}\right)\right] \quad (3)$$

P, *L*, *N*, are the total number of instances (in our case, MHC I proteins) belonging to the corresponding class and P_i, L_i, N_i with i = (P, L, N), represents predicted instances and their class. For example, P_P numbers P instances predicted as P, while P_N and P_L, number P instances classified as N and L, respectively. Note that *SE* and *SP* are computed for each of the tree classes (P, L, N) while *ACC* corresponds to the percentage of properly predicted instances. Because we run each 10-fold cross-validation 10 times, for each model we obtained 100 different estimates of the noted parameters of performance, computing the mean and standard deviations. To compare the performance of the models, we carried our paired *t*-test in WEKA over the *ACC* [6].

2.3. Other procedures

We used MULTIPROT [9] for superimposing protein 3D-structures of MHC I proteins and STACCATO [10] for deriving a multiple sequence alignment (MSA) from the superimposed structures. MHC Ia and Ib proteins were aligned with TCOFFEE [11], using a seed structural alignment. We obtained dendrograms from the relevant MSAs using the Neighbor-joining method [12], and we addressed their reliability using bootstrapping [13] with 1000 replications. We also used bootstrapping with 1000 replications to evaluate the performance of ML-based models on holdout protein sequences with an increasing number of random mutations at random variable sites (PERL script to mutate sequences will be provided upon request). BLAST searches were executed with default settings against a BLAST-formatted database derived from the MHCI^{556} dataset in FASTA format in which each distinct MHC I sequence had a header with a label indicating the nature of its ligand (P: peptides, L: lipids and N: No ligand).

3. Results and discussion

3.1. MHC I sequence similarity space

In order to investigate whether one could define the ligand-type specificity of MHC I proteins (P, L, N) by sequence similarity approaches, we superimposed the 3D-structures of the α1α2 domain of representative MHC I proteins (Fig. 1A), obtained a structured-based alignment (Fig. 1B) and subsequently derived a sequence similarity-based dendrogram (Fig. 1C). We found that the selected MHC I proteins fail to cluster in just three distinguishable groups matching their ligand-type specificity (Fig. 1C). Thus, while some MHC I proteins group according to their ligand-type specificity (e.g., the lipid-binding CD1 antigens and EPCRs, and the peptide-binding classical and non-classical MHC I proteins), other molecules like ZAG, that bind lipids, and TLA, HFE, and FcRN, that have no ligand, appear unrelated to the molecules of the relevant groups. Also, TLA, which does not bind any ligand, is much closer to the group of peptide-binding MHC I proteins than to those that do not bind anything. Likewise, using TreeDet [14], a popular and robust program to explore sequence space, we were unable to identify key signature residues allowing the distinction of MHC I proteins according to the nature or their ligand. These results are likely due to the fact that the division between MHC I proteins by ligand-type specificity is functionally relevant but it is not phylogenetic. For example, TLA and MICA, both incapable of binding any ligand, are, beyond having the MHC I fold, completely unrelated and shared sequence identity of just 24.06%. The sequence identity between all the proteins considered in the structure-based alignment is shown in Supplementary file 1.

Since the ligand-type specificity of MHC I proteins could not be properly distinguished in the MHC I sequence similarity space, in this work, we approached the task of predicting the type of ligand

Please cite this article in press as: Martínez-Naves, E., et al. Recognition of the ligand-type specificity of classical and non-classical MHC I proteins. FEBS Lett. (2011), doi:10.1016/j.febslet.2011.10.007

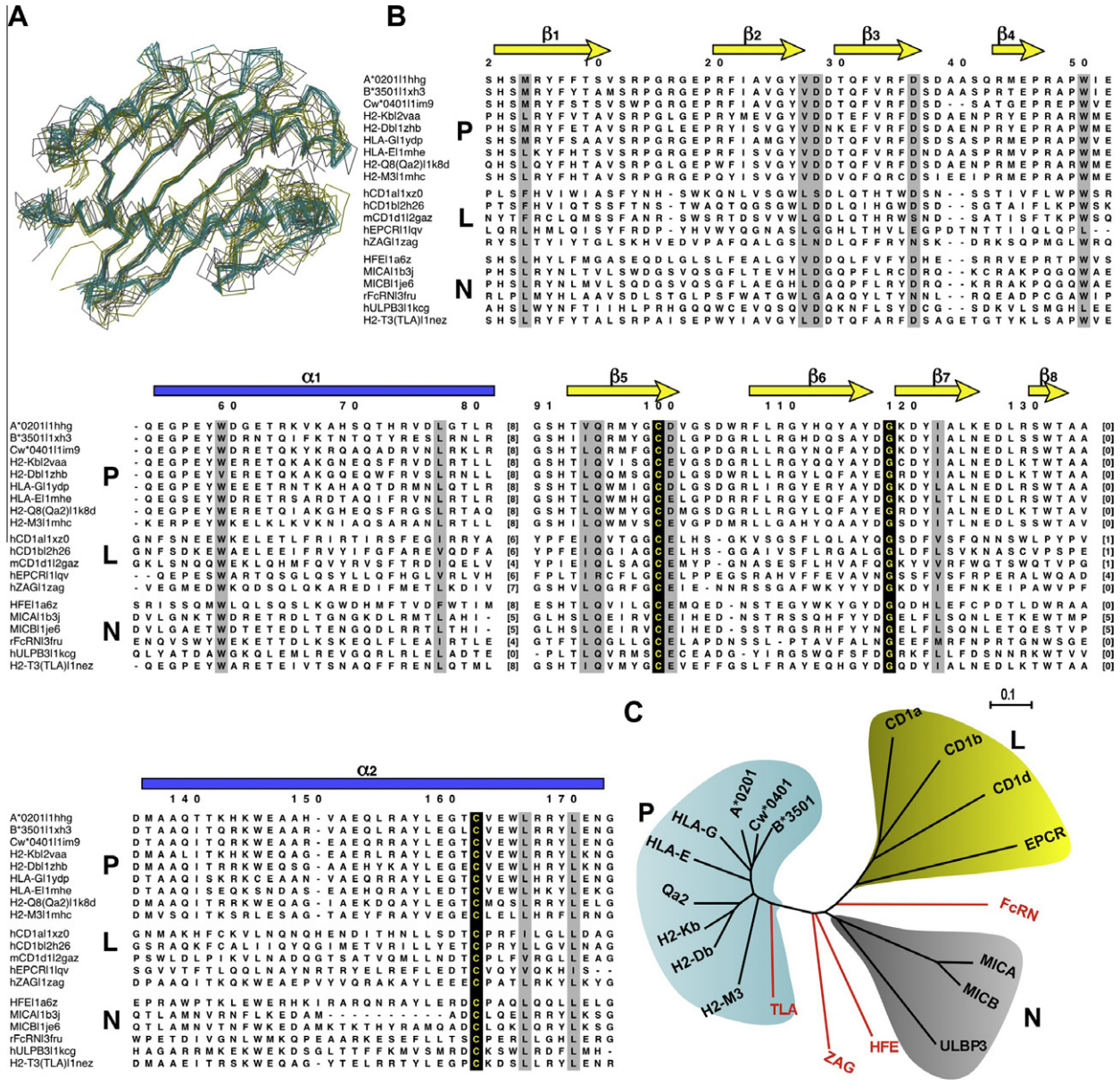

Fig. 1. MHC I sequence similarity space. (A) Structural superimposition of MHC I proteins. The figure depicts a 3D-structure superimposition of MHC I proteins (α1α2 domain) that are known to bind peptides (in blue; PDBs: 1HHG, 1XH3, 1IM9, 2VAA, 1YDP, 1MHE, 1K8D and 1MHC), lipids (in yellow; PDBs: 1XZ0, 2H26, 2GAZ, 1LQV and 1ZAG), and have no bound ligands (in grey; PDBs: 1A6Z, 1B3J, 1JE6, 3FRU, 1KCG and 1NEZ). (B) Protein MSA obtained from the structural superimposition shown in panel A. Amino acid sequence numbering and secondary structures elements match those of the human MHC I molecule HLA-A*0201 (PDB: 1HHG). For clarity, some amino acids in loop regions have been deleted (numbers shown in square brackets). Identical and conserved residues are shadowed in black and grey, respectively. (C) Neighbor-joining tree built from the structure-based alignment shown in panel B. The groups of sequences that appear shadowed cluster with a bootstrap reliability ⩾ 90%. Proteins that do not group with other proteins having the same ligand-type specificity are highlighted in red. The data to reconstruct this dendrogram is provided in Supplementary file 2.

of MHC I proteins as a classification problem for machine learning (ML).

3.2. Evaluation of ML-based classifiers in cross-validation

We built our ML-based models by training several ML algorithms on the amino acid composition of 556 non-overlapping and unique MHC I protein sequences (α1α2 domain) of know ligand-type specificity ($MHCI^{556}$ dataset): 355 bind peptides (P), 84 bind lipids (L) and 117 do not bind anything (N) (Table 1). Specifically, we trained k-Nearest Neighbor algorithm (kNN) and support vector machines (SVMs) with polynomial (SVM-Pk) and RBF-kernels (SVM-RBFk). We selected these algorithms because of their reliability, simplicity and speed [7]. We used 10-fold cross-validations to built, evaluate and optimize the models (Fig. 2).

The best performance was obtained using SVM-RBFk, which reached an *ACC* of 100.0 ± 0.0%; not a single MHC I protein was misclassified (Fig. 2B). Next to these results were those obtained using kNN (*ACC* = 99.93 ± 0.35%, *SE* = 99.98 ± 0.03%, *SP* = 99.82 ± 1.16%), which misclassified one protein (Fig. 2B). The performance achieved using SVM-Pk (*ACC* = 99.46 ± 0.87, *SE* = 99.96 ± 0.05, *SP* = 99.51 ± 1.82) was slightly lower than that obtained with kNN; three proteins were misclassified (Fig. 2B). It could be well argued that these results are conditioned by the fact that the $MHCI^{556}$ dataset is imbalanced; the peptide-binding group is much

Table 1
MHC I proteins in the MHCI[556] dataset.

MHCI	Species	Seqs.	Ligand
HLA-[ABC]	Human	111	P
DLA-88	Dog	22	P
SLA-[123]	Swine	51	P
BoLA-N	Cattle	39	P
OLA-N	Sheep	12	P
ONMY-UBA	Rainbow trout	29	P
SASA-UBA	Atlantic Salmon	27	P
RT1-A	Rat	21	P
H2-X	Mouse	26	P
HLA-E	Human and Primates	6	P
HLA-G	Human and primates	1	P
H2-T23(Qa1)	Mouse and Rat	4	P
H2-Q9	Mouse	2	P
H2-M3	Mouse and Rat	4	P
CD1[A-E]	Vertebrates	71	L
ZAG	Vertebrates	6	L
EPCR	Vertebrates	7	L
MICA&B	Vertebrates	38	N
HFE	Vertebrates	6	N
MILL1&2	Mouse and Rat	4	N
FcRN	Mammals	9	N
ULBP	Vertebrates	45	N
H2-T3(TLA)	Mouse and Rat	15	N

We only included the sequence of the α1α2 domain. The corresponding author will provide this dataset upon writing request.

larger than the other two. However, all MHC I proteins were properly classified regardless of their class (Fig. 2B). Moreover, we virtually obtained the same results (Supplementary Fig. S1) using a dataset (MHCI[334] dataset) encompassing only 133 MHC I peptide-binding proteins (P). The MHCI[334] dataset is described in Supplementary Table S1.

In sum, classifying the highly divergent MHC I protein families into the three defined groups was a surprisingly simple task for ML, which on the one hand highlights the quality of the assembled dataset and on the other suggests that the ligand-type specificity of MHC I proteins is readily imprinted in their amino acid composition. In comparison, accurate classification of biological sequences using ML often requires many more input features such as using dipeptide composition [15] However, it is true that those studies involved binary classifications of a group of related proteins (e.g., Histones) from the remaining universe of proteins (e.g., non-Histones). Instead, here we performed a multiclass classification.

3.3. Evaluation of ML-based classifiers on holdout MHC I proteins

We carried out several validations on distinct independent datasets consisting on entire groups of MHC I proteins (holdout sequences) that were drawn from the MHCI[556] dataset prior to model building. Our goal was to explore whether ML-based classifiers built and optimized in cross-validation can predict the ligand-type specificity of MHCI proteins differing entirely from those used for model building as well as to identify MHC I proteins that are critical to guarantee the robustness of the method. The results of these analyses are depicted on Table 2 and summarized next.

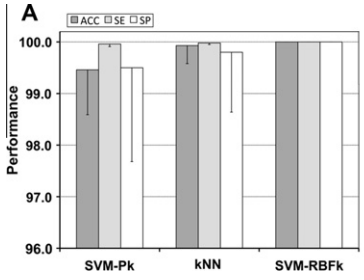

Fig. 2. Predictive performance of ML-based classifiers. (A) Graph depicting the *ACC* (dark grey bars), *SE* (light grey bars) and *SP* (white bars) in percentage achieved by ML-based classifiers (abscissas) predicting the class (P, L, N) of the molecules included in the MHCI[556] dataset in cross-validation. ML-based classifiers consisted k-Nearest Neighbor (kNN) and SVM with polynomial (SVM-Pk) and RBF (SVM-RBFk) kernels. We represent average values of *ACC*, *SE*, and *SP* with their corresponding standard deviations as negative errors bars. Note that we plot standard deviations not standard errors. Also, the y-scale starts at 96%. (B) Confusion matrix obtained in a representative 10-fold cross-validation experiment. The performance of ML-based classifiers shown in the figure was achieved using the following parameters: kNN: $K = 4$; SVM-Pk: $E = 3$, $C = 1$; SVM-RBFk: $\gamma = 6$, $C = 1$.

Table 2
Evaluation of ML-based classifiers on holdout MHC I proteins.

Holdout group			Performance in cross-validation[c]		Classification[d]		
	Lig.[a]	Seq.[b]	Algorithm	*ACC*	P	L	N
HLA-[ABC]	P	111	kNN	99.97 ± 0.22	111	0	0
			SVM-Pk	99.33 ± 1.17	111	0	0
			SVM-RBFk	100.0 ± 0.00	111	0	0
OLA-& BoLA-N (sheep & cattle)	P	51	kNN	99.96 ± 0.25	51	0	0
			SVM-Pk	99.46 ± 0.78	51	0	0
			SVM-RBFk	100.0 ± 0.00	51	0	0
SLA-[123] (Swine)	P	51	kNN	99.98 ± 0.18	51	0	0
			SVM-Pk	99.40 ± 1.12	51	0	0
			SVM-RBFk	100.0 ± 0.00	51	0	0
RT1-A & H2-X (murine)	P	47	kNN	99.90 ± 0.42	47	0	0
			SVM-Pk	99.53 ± 1.02	47	0	0
			SVM-RBFk	100.0 ± 0.00	47	0	0
SASA- & ONMY-UBA (fish)	P	56	kNN	100.0 ± 0.0	27	22	7
			SVM-Pk	99.43 ± 1.00	18	10	28
			SVM-RBFk	100.0 ± 0.00	25	10	21
CD1[A-E]	L	71	kNN	99.92 ± 0.31	6	10	55
			SVM-Pk	99.80 ± 0.66	6	28	37
			SVM-RBFk	100.0 ± 0.0	4	7	60
EPCR	L	7	kNN	99.98 ± 0.18	0	7	0
			SVM-Pk	99.41 ± 1.02	0	7	0
			SVM-RBFk	100.0 ± 0.00	0	7	0

Please cite this article in press as: Martínez-Naves, E., et al. Recognition of the ligand-type specificity of classical and non-classical MHC I proteins. FEBS Lett. (2011), doi:10.1016/j.febslet.2011.10.007

Table 2 (*continued*)

Holdout group			Performance in cross-validation[c]		Classification[d]		
	Lig.[a]	Seq.[b]	Algorithm	*ACC*	P	L	N
ZAG	L	6	kNN	99.94 ± 0.31	4	0	2
			SVM-Pk	99.75 ± 0.63	5	0	1
			SVM-RBFk	100.0 ± 0.00	2	0	4
MICA&B	N	38	kNN	99.96 ± 0.25	0	0	38
			SVM-Pk	99.54 ± 0.99	0	0	38
			SVM-RBFk	99.96 ± 0.27	0	0	38
HFE	N	6	kNN	99.96 ± 0.25	1	0	5
			SVM-Pk	99.44 ± 0.98	3	0	3
			SVM-RBFk	100.00 ± 0.00	0	0	6
ULBP	N	45	kNN	99.80 ± 0.59	3	21	21
			SVM-Pk	99.61 ± 0.97	9	11	25
			SVM-RBFk	99.88 ± 0.46	1	7	37
FcRN	N	9	kNN	99.96 ± 0.25	6	0	3
			SVM-Pk	99.44 ± 0.92	0	3	6
			SVM-RBFk	100.0 ± 0.00	0	3	6
TLA	N	15	kNN	99.94 ± 0.31	15	0	0
			SVM-Pk	99.94 ± 0.41	15	0	0
			SVM-RBFk	100.0 ± 0.00	15	0	0

[a] Known ligand-type class of holdout sequences (P: bind peptides; L: bind lipids; N: no ligand).
[b] Number of holdout sequences.
[c] Models were built and optimized on MHC I datasets without the holdout proteins in 10-fold cross-validation experiments that were repeated 10 times. *ACC* depicted in table correspond to that of the optimal model.
[d] Class assignment (predicted ligand-type specificity: P, L, N) of holdout sequences using models built and optimized on datasets lacking that same sequences. Shadowed cells point to the right outcome of the predictions.

In each distinct training dataset, the ML-based classifiers reached in cross-validation an extraordinary accuracy (*ACC* ⩾ 99.3%) that mirrored the results obtained on the full MHCI[556] dataset (*ACC* SVM-RBFk > *ACC* kNN > *ACC* SVM-Pk). In many occasions, ML-based classifiers were able to predict the right ligand-type specificity of the proteins that were removed prior to model building (Table 2). Moreover, we noted that classifiers that performed best in cross-validation classified the holdout proteins with fewer errors (see results involving HFE, ULBP, and FcRn holdout tests in Table 2). However, we also found MHC I proteins, such as CD1 antigens ZAG and TLA proteins (Table 2), that could not be classified appropriately, indicating that it is critical to include them in the training datasets. ML-based models were also unable to predict peptide binding for all MHC I proteins form fish (Table 2). Although there is not enough experimental evidence, it is reasonable, yet arguable, to think that all these fish MHC I should bind peptides; they have been classified as classical MHC I molecules in specialized databases (see http://www.ebi.ac.uk/ipd/mhc/fish/index.html).

We also investigated the potential tolerance of ML-models to mutations in the testing data. Specifically, we evaluated the ability of the three SVM-RBFk models trained on datasets lacking HLA I, EPCR and MICA&B proteins to predict the correct class of the relevant proteins modified with an increasing number of random mutations at random variable sites. The results indicate that at least these models are quite tolerant to mutations in the tested proteins (decreasing the percentage of properly predicted instances to 80% required more than 30 mutations) (see Supplementary Fig. S2).

3.4. Predicted ligand-type specificity of MHC Ib molecules of uncharacterized ligands

There are a number of mouse and human MHC Ib molecules whose ligands, if any, have not been characterized yet. In humans, the MHC Ib molecules of unverified ligands are HLA-F and MR1, whereas in mouse are MR1 and a number molecules encoded by the *H2-T*, *H2-Q* and *H2-M* loci. Here, we predicted the ligand-type specificity of these proteins (α1α2 domain) using the ML-based classifiers trained on the MHCI[556] dataset and compared the results with those resulting from a BLAST search using the corresponding α1α2 domain sequences against the MHCI[556] dataset (ligand-type specificity was assigned to that of the closest hit). To verify the generalization power of our classifiers, we also predicted the ligand-type specificity of two MHC I molecules from chicken (BF2*2101 and YF1*7.1) as well as UL18, H2-T9, H2-T10 and H2-T22. All these proteins, despite their known binding ability, were not used for model building (they were not included in the MHCI[556] dataset). UL18 is a viral MHC I-like molecule from human cytomegalovirus that is known to bind peptides [16], whereas H2-T9, H2-T10 and H2-T22 are closely related murine MHC Ib molecules, which are incapable of binding any ligand because of having a truncated α1α2 domain [17,18]. As for the chicken MHC I proteins, BF2*2101 binds peptides [19] while YF1*7.1 appears to bind some unknown non-classical ligand (perhaps a lipid) [20]. A phylogenetic tree depicting the analyzed proteins is shown Fig. 3. The predictions are summarized in Table 3 and we will next highlight some of the results.

UL18 and both chicken MHC I proteins were predicted to bind peptides using BLAST and the ML-based classifiers (Table 3). However, only ML-based classifiers predicted the lack of ligand of H2-T9, H2-T10 and H2-T22 (Table 3). In fact, BLAST predicted peptide-binding for all the MHC Ib proteins that were tested (Table 3). These results suggest that the ML approach is more suitable to predict the ligand-type specificity of MHC I proteins than the BLAST approach. Nevertheless, the two approaches are not necessarily exclusive but complementary.

ML-based classifiers revealed MHC Ib proteins with ligand-type specificities that could not have been anticipated using sequence similarity analyses. Thus, ML-based classifiers predicted that the mouse H2-M1 and H2-M10 proteins (Fig. 3), which are expressed in vomeronasal sensory neurons (VNS) [21], bind lipids and have no ligand, respectively (Table 3). The lack of ligand of H2-M10 proteins is consistent with the available crystal structure of H2-M10.5

Please cite this article in press as: Martínez-Naves, E., et al. Recognition of the ligand-type specificity of classical and non-classical MHC I proteins. FEBS Lett. (2011), doi:10.1016/j.febslet.2011.10.007

Table 3
Predicted ligand-type specificity of selected MHC Ib proteins.

Molecule	Gene ID	GB ACN[b]	Predicted ligand-type specificity			
			BLAST	SVM-Pk	kNN	SVM-RBFk
UL18	3077466	YP_081477	P	P	P	P
BF2(BF2∗2101)	425389	NP_001026509	P	P	P	P
YF1∗7.1	427746[c]	NP_001074336	P	P	P	P
HLA-F	3134	NP_001091948	P	P	P	P
H2-Q1	15006	NP_034520	P	P	P	P
H2-Q2	15013	NP_034522	P	P	P	P
H2-M2	14990	NP_032230	P	P	P	P
H2-M1	224756	NP_808304	P	N	P	L
H2-M9	14997	NP_032231	P	L	L	L
H2-M11	224754	NP_808303	P	L	L	L
H2-M10.1	14985	NP_038572	P	N	N	N
H2-M10.2	333715	NP_808591	P	N	N	N
H2-M10.3	110696	NP_963902	P	N	N	N
H2-M10.4	224753	NP_808302	P	N	P	N
H2-M10.5	224761	NP_808305	P	P	P	N
H2-M10.6	399549	NP_963905	P	N	P	N
H2-T24	15042	NP_032233	P	N	N	N
H2-T9[a]	15051	NP_034529	P	N	N	N
H2-T10[a]	15024	NP_034525	P	N	P	N
H2-T22[a]	15039	NP_034527	P	N	N	N
mMR1	15064	NP_032235	P	N	N	N
hMR1	3140	NP_001522	P	N	N	N

Predictions were carried out using only the α1α2 domains of the relevant proteins.
[a] Sequence length shorter than training sequences.
[b] Genbank accession numbers.
[c] Gene annotated in NCBI as MR1 major histocompatibility complex, class I-related, most likely by mistake.

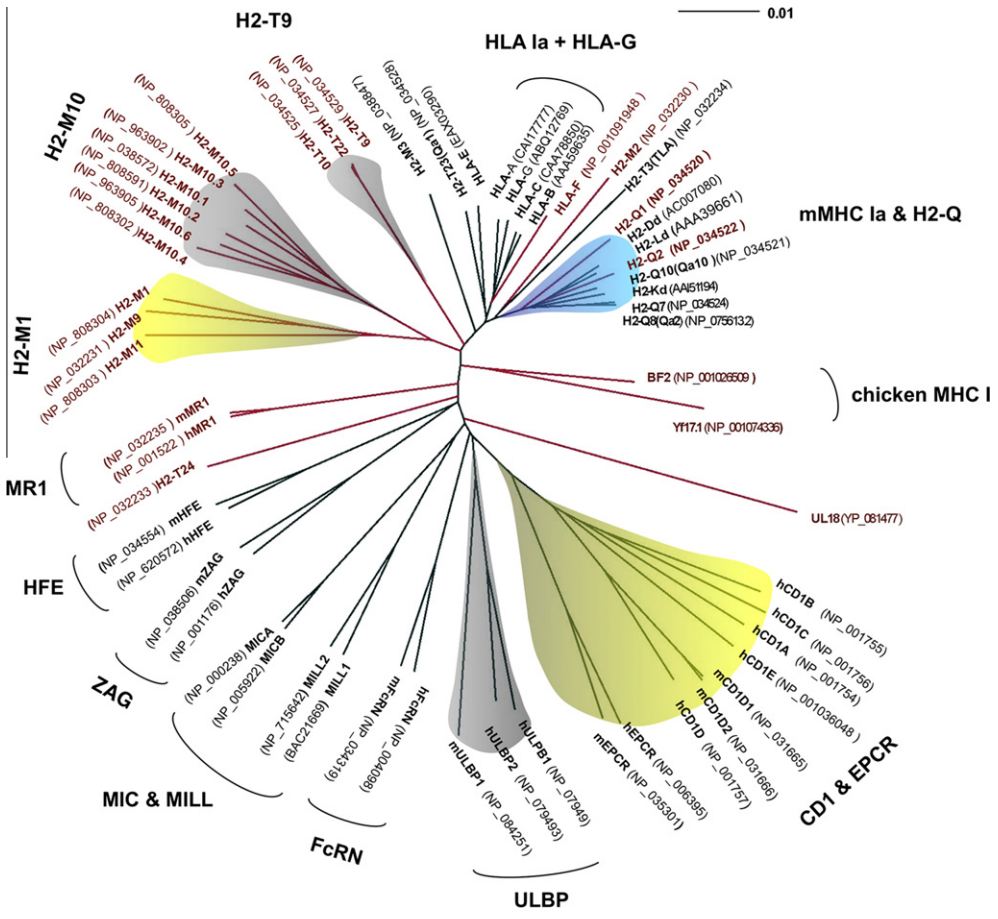

Fig. 3. Sequence relationship between human and mouse MHC I proteins. Figure depicts a Neighbor-joining dendrogram from an MSA of the α1α2 domain of various proteins, including mouse and human MHC Ia and Ib molecules of uncharacterized ligands (Genbank accession numbers shown in parenthesis). The MHC Ib proteins that were the subject of the predictions are shown in red. MHC I clusters with more than three members and a bootstrap confidence >90% are shadowed as follows: blue if they bind peptides, yellow if they bind lipids and grey in they have no ligand. The molecules labeled as HLA Ia consist of human classical MHC I proteins. The molecules labeled as mMHC Ia & H2-Q are mouse classical MHC I molecules and non-classical MHC I molecules encoded by the mouse H2-Q loci, respectively. In Supplementary file 3, we provide the data to reconstruct this dendrogram.

Please cite this article in press as: Martínez-Naves, E., et al. Recognition of the ligand-type specificity of classical and non-classical MHC I proteins. FEBS Lett. (2011), doi:10.1016/j.febslet.2011.10.007

[22]. Functional V2R pheromone receptors express concomitantly with H2-M1 proteins in VNS [21], and we speculate that likely these receptors recognize lipidic pheromones presented by H2-M1 proteins.

ML-based classifiers also predicted that MR1 does not bind any ligand. MR1 (MHCI-related, class I, molecule) is a highly conserved MHC Ib molecule, which has been shown to restrict a sub-population of αβ T cells named MAIT (Mucosal-associated invariant T cells) [23]. The expression of MR1 appears to be TAP and proteasome independent but yet some authors have found evidence supporting antigen-presentation function for MR1, possible of peptides [24–26]. Moreover, a group has indicated that MR1 presents α-mannosyl glycolipids to invariant V α19-J α33 MAIT cells [27], although others have failed to confirm these findings [28]. While it is possible that MR1 binds molecules that are different from peptides or lipids, our results indicate that MR1 function and restriction of MAIT cells might be independent of antigen presentation.

Because it is not clear whether all of the classical MHC I proteins from fish included in the $MHCI^{556}$ dataset (Table 1) do really bind peptides, we repeated all these predictions using ML-based models trained without the fish proteins ($MHCI^{500}$ dataset). The results were largely the same, as shown in Supplementary Table S2.

4. Conclusions and limitations

Currently, there is a plethora of methods to predict peptide binding to specific MHC Ia molecules [29], but surprisingly, until now no method was available to predict whether any uncharacterized member of the MHC I protein family can bind any ligand at all, and if so, the nature of such ligand (peptides or lipids). This information is key to lead the experiments that allow the identification of the relevant ligands. Upon an original multi-class classification approach, we developed here ML-based classifiers that achieve such task with great accuracy and generalization power. Prediction of the ligand-type specificity of MHC I proteins using our classifiers is available for free public use at http://imed.med.ucm.es/MHCLIG/. It is important to stress that the classifiers developed here can only predict the three known ligand-type specificities of MHC I proteins (P, L and N). If there would be MHC I proteins with other, yet to be characterized, type of ligands (e.g., sugars, nucleotides, etc.) the breath of our predictions will be clearly limited.

Acknowledgements

We wish to thank Dr. Alfonso Valencia for helpful comments. This work was supported by the Ministerio de Ciencia e Innovación of Spain (grants SAF2006-07879 and SAF2009-08103 to P.A.R. and grant SAF2007-60578 to E.M.L.), the Spanish Ministry of Health (grant PI080125 to E.M.N.), the Comunidad Autonoma de Madrid (grant CCG08-UCM/BIO-3769 to P.A.R.) and the Universidad Complutense de Madrid (grant Gr58/08 920631 to E.M.N., E.M.L. and P.A.R.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2011.10.007.

References

[1] Maenaka, K. and Jones, E.Y. (1999) MHC superfamily structure and the immune system. Curr. Opin. Struct. Biol. 9, 745–753.
[2] Townsend, A. and Bodmer, H. (1989) Antigen recognition by class I-restricted T lymphocytes. Annu. Rev. Immunol. 7, 601–624.
[3] Braud, V.M., Allan, D.S. and McMichael, A.J. (1999) Functions of nonclassical MHC and non-MHC-encoded class I molecules. Curr. Opin. Immunol. 11, 100–108.
[4] Rodgers, J.R. and Cook, R.G. (2005) MHC class Ib molecules bridge innate and acquired immunity. Nat. Rev. Immunol. 5, 459–471.
[5] Martínez-Naves, E., Lafuente, E.M. and Reche, P.A. (2011) Classification of MHC I proteins according to their ligand-type specificity in: 10th International Conference on Artificial Immune Systems (Liò, P., Nicosia, G. and Stibor, T., Eds.), pp. 55–65, Springer-Verlag Cambridge, England, UK.
[6] Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I.H. (2004) Data mining in bioinformatics using WEKA. Bioinformatics 20, 2479–2481.
[7] Wu, X. et al. (2008) Top 10 algorithms in data mining. Knowl. Inf. Syst 14, 1–37.
[8] Dasarathy, B.V. (1991) Nearest neighbor (NN) norms: NN pattern classification techniques, IEEE Computer Society Press, Los Alamitos, California.
[9] Shatsky, M., Nussinov, R. and Wolfson, H.J. (2004) A method for simultaneous alignment of multiple protein structures. Proteins 56, 143–156.
[10] Shatsky, M., Nussinov, R. and Wolfson, H.J. (2006) Optimization of multiple-sequence alignment based on multiple-structure alignment. Proteins. 62, 209–217.
[11] Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217.
[12] Firestine, S.M., Nixon, A.E. and Benkovic, S.J. (1996) Threading your way to protein function. Chem. Biol. 3, 779–783.
[13] Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39, 783–791.
[14] Carro, A., Tress, M., de Juan, D., Pazos, F., Lopez-Romero, P., del Sol, A., Valencia, A. and Rojas, A.M. (2006) TreeDet: a web server to explore sequence space. Nucleic Acids Res. 34, W110–W115.
[15] Bhasin, M., Reinherz, E.L. and Reche, P.A. (2006) Recognition and classification of histones using support vector machine. J. Comput. Biol. 13, 102–112.
[16] Yang, Z. and Bjorkman, P.J. (2008) Structure of UL18, a peptide-binding viral MHC mimic, bound to a host inhibitory receptor. Proc. Natl. Acad. Sci. USA 105, 10095–10100.
[17] Crowley, M.P., Fahrer, A.M., Baumgarth, N., Hampl, J., Gutgemann, I., Teyton, L. and Chien, Y. (2000) A population of murine γδ T cells that recognize an inducible MHC class Ib molecule. Science 287, 314–316.
[18] Shin, S., El-Diwany, R., Schaffert, S., Adams, E.J., Garcia, K.C., Pereira, P. and Chien, Y.H. (2005) Antigen recognition determinants of γδ T cell receptors. Science 308, 252–255.
[19] Koch, M. et al. (2007) Structures of an MHC class I molecule from B21 chickens illustrate promiscuous peptide binding. Immunity 27, 885–899.
[20] Hee, C.S., Gao, S., Loll, B., Miller, M.M., Uchanska-Ziegler, B., Daumke, O. and Ziegler, A. (2010) Structure of a classical MHC class I molecule that binds "non-classical" ligands. PLoS Biol. 8, e1000557.
[21] Loconto, J. et al. (2003) Functional expression of murine V2R pheromone receptors involves selective association with the M10 and M1 families of MHC class Ib molecules. Cell 112, 607–618.
[22] Olson, R., Huey-Tubman, K.E., Dulac, C. and Bjorkman, P.J. (2005) Structure of a pheromone receptor-associated MHC molecule with an open and empty groove. PLoS Biol. 3, e257.
[23] Treiner, E. et al. (2003) Selection of evolutionarily conserved mucosal-associated invariant T cells by MR1. Nature 422, 164–169.
[24] Huang, S., Gilfillan, S., Cella, M., Miley, M.J., Lantz, O., Lybarger, L., Fremont, D.H. and Hansen, T.H. (2005) Evidence for MR1 antigen presentation to mucosal-associated invariant T cells. J. Biol. Chem. 280, 21183–21193.
[25] Le Bourhis, L. et al. (2010) Antimicrobial activity of mucosal-associated invariant T cells. Nat. Immunol. 11, 701–708.
[26] Abos, B., Gomez Del Moral, M., Gozalbo-Lopez, B., Lopez-Relano, J., Viana, V. and Martinez-Naves, E. (2011) Human MR1 expression on the cell surface is acid sensitive, proteasome independent and increases after culturing at 26 degrees C. Biochem. Biophys. Res. Commun. 411, 632–636.
[27] Shimamura, M. et al. (2007) Modulation of Vα19 NKT cell immune responses by α-mannosyl ceramide derivatives consisting of a series of modified sphingosines. Eur J Immunol. 37, 1836–1844.
[28] Huang, S. et al. (2008) MR1 uses an endocytic pathway to activate mucosal-associated invariant T cells. J. Exp. Med. 205, 1201–1211.
[29] Lafuente, E.M. and Reche, P.A. (2009) Prediction of MHC-peptide binding: a systematic and comprehensive overview. Curr. Pharm. Des. 15, 3209–3220.