Machine learning predicts immunoglobulin light chain toxicity through somatic mutations

Maura Garofalo[1], Luca Piccoli[1], Sara Ravasio[1,2], Mathilde Foglierini[1,3], Milos Matkovic[1], Jacopo Sgrignani[1], Marco Prunotto[4], Olivier Michielin[5,6], Antonio Lanzavecchia[1], and Andrea Cavalli[1,3*]

Brief Communication

[1]Institute for Research in Biomedicine, Università della Svizzera italiana, Bellinzona, Switzerland

[2]Institute of Microbiology, ETH Zurich, Zurich, Switzerland

[3] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4]School of Pharmaceutical Sciences, Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland

[5]Molecular Modeling Group, SIB Swiss Institute of Bioinformatics, University of Lausanne, Quartier UNIL-Sorge, Bâtiment Amphipôle, Lausanne, Switzerland

[6]Department of Oncology, University Hospital of Lausanne, Ludwig Cancer Research - Lausanne Branch, Lausanne, Switzerland

Corresponding author

Andrea Cavalli

Institute for Research in Biomedicine

Università della Svizzera italiana

Via Vincenzo Vela 6

CH-6500 Bellinzona, Switzerland

Email: andrea.cavalli@irb.usi.ch

Abstract

In light chain (AL) amyloidosis, pathogenic monoclonal light chains (LCs) deposit as amyloid fibrils in target organs. Molecular determinants of LC pathogenicity are currently unknown. Here, we present LICTOR, a method to predict LC toxicity based on the distribution of somatic mutations acquired during clonal selection. LICTOR achieves specificity and sensitivity of 0.82 and 0.76, respectively, with an AUC of 0.87, making it a valuable tool for early AL diagnosis.

Main Text

Light chain amyloidosis is a monoclonal gammopathy characterized by the abnormal proliferation of a plasma cell clone producing high amounts of pathogenic immunoglobulin free LCs. These LCs, mainly secreted as homodimers[1], misfold and accumulate in target tissues, mostly heart or kidney, forming toxic oligomers and amyloid fibrils that lead to fatal organ dysfunction and death[2].

Pre-existing monoclonal gammopathy of undetermined significance (MGUS) is a known risk factor for AL, with 9% of MGUS patients progressing to AL[3,4]. However, early AL diagnosis is still difficult since reliable diagnostic tests predicting whether MGUS patients are likely to develop AL are currently missing[5,6]. Predicting the onset of AL is problematic as each patient carries a different pathogenic LC sequence, which is composed by a unique rearrangement of variable (V) and joining (J) immunoglobulin genes, and by a unique set of somatic mutations (SMs) acquired during B cell affinity maturation[7] (Fig. 1a). Therefore, the development of specific prediction tools would be a crucial step to anticipate AL diagnosis and improve patients' prognosis.

Here, we present LICTOR (λ-LIght-Chain TOxicity predictoR), a machine learning approach to classify lambda (λ) LCs, from their amino acid sequences, as either toxic or non-toxic,

depending on their likelihood to form toxic species inducing AL. LICTOR uses SMs as predictor variables based on the hypothesis that SMs are the main discriminating factor of LC toxicity. Predictions are currently restricted to λ LCs, since this isotype is more prevalent than the kappa (κ) in AL patients (λ/κ=3:1 compared to healthy individuals λ/κ=1:2)[8].

To validate SMs as predictor variables and parametrize LICTOR, we collected a database of 1,075 λ LC sequences, including 428 "toxic" sequences from AL patients (here referred as *tox*) and 647 "non-toxic" ones (*nox*) comprising LC sequences from healthy donors' repertoires or other autoimmune and cancer diseases[9]. All LCs were aligned to the corresponding germline (GL) sequence obtained using the IMGT database[10] to identify SMs. Furthermore, LCs were numbered according to the Kabat-Chothia scheme allowing the structural comparison of LCs with a different sequence length (Methods and Fig. 1b). Then, we counted the number of mutated (M) and non-mutated (NM) residues at each position *i* in *tox* and *nox* sequences (tox_M^i and nox_M^i, tox_{NM}^i and nox_{NM}^i, respectively) and used the Fisher exact test[11] to assess their statistical difference ($p < 0.05$). Finally, the odds ratio (OR)[11] was used as a measure of the magnitude of the different probability of observing a mutation at position *i* in *tox* and *nox* sequences (Methods and Fig. 1c). Interestingly, 48 out of 53 positions with a statistically significant difference ($p < 0.05$) between the two groups (Fig. 1c) showed a higher rate of mutation in the *tox* group (OR >1), while only 5 positions displayed higher mutation rate in *nox* group (OR <1). To exclude a bias induced by the use of a group of *nox* sequences having a low level of SMs, we randomly selected 1000 LC sequences form a healthy donor repertoire (*hdnox*)[12] and compared the probability distribution of the number of SMs (PDSM) between the three groups. We observed similar PDSM between *nox* and *hdnox* groups, while the PDSM of *tox* and *hdnox*, as well as *tox* and *nox*, were significantly different. This result supports *nox* sequences as a bona fide group of LCs (Supplementary Fig. 1). Overall, these findings suggest that SMs are key determinants for the toxicity of LCs and can,

thus, be used as features to develop AL prediction tools. Therefore, as a next step, we combined the information from SMs together with the knowledge of the 3D structure of LC homodimers[13,14] to create three families of predictor variables used to train LICTOR. The first family, named AMP (Amino acid in each Mutated Position), identifies the presence or the absence of a SM at each position of the LC sequences. The second family, named MAP (Monomeric Amino acid Pairs) identifies the presence or the absence of mutations in residues in close contact in the LC monomeric 3D structure (distance <7.5Å). The third family, named DAP (Dimeric Amino acid Pairs) identifies the presence or the absence of mutations at positions in close contact but belonging to different chains. Next, four machine learning algorithms (Bayesian network, logistic regression, J48 and random forest)[15] were evaluated for their ability to solve the classification problem of toxic and non-toxic LC sequences, using our database as input. To assess the importance of the different classes of predictor variables, we performed 28 prediction experiments including all possible combinations of AMP, MAP and DAP families. In addition, to avoid class-unbalancing problems, i.e., the tendency of the machine learning algorithm to assign sequences to the largest class, each of the 28 experiments was performed with and without balancing of the training set using a SMOTE (Synthetic Minority Over-sampling TEchnique) filter[16]. We found that for all tested machine learners the best combination of predictor variables families resulted in an area under the receiver operating characteristic (AUC) that substantially differed from a random classifier (0.50), with random forest as the best classifier (0.87) and J48 the worst (0.75) (Fig. 1d and Supplementary Table 1). Furthermore, all four classifiers relied on SMs recapitulated by AMP to predict LC toxicity, while only random forest used all the three families of predictor variables (AMP+MAP+DAP). Overall, these findings highlight the importance of the structural context of somatic mutations to define the toxicity of a LC and identify random forest using AMP, MAP, and DAP, as the best machine leaner and, thus, is the one implemented in LICTOR.

To further underscore the key role of SMs as discriminants between toxic and non-toxic LCs, we trained the same machine learners using the LC germline VJ rearrangements as a unique predictor variable, given the well-documented overrepresentation of certain LC rearrangements in AL[17,18]. All the resulting *germline-based* classifiers achieved an AUC of 0.77 in their best configuration (Fig. 1e and Supplementary Table 2), a value substantially better than a random classifier, although much lower than LICTOR's one (0.87). Interestingly, adding the LC germline VJ rearrangements in LICTOR did not improve the prediction performance (Supplementary Table 3).

Next we computed the specificity and sensitivity of the two random forest predictors maximizing the Youden index (J)[19], as a function of the *confidence level* of the random forest predictions, i.e. the probability that a sequence belongs to the predicted group (Fig. 1f). LICTOR achieves a specificity of 0.82 and a sensitivity of 0.76 (J=0.58, threshold=0.46 in identifying *tox*), while the *germline-based* classifier shows a 0.69 specificity and a 0.73 sensitivity (J=0.43, threshold=0.48 in identifying *tox*).

To further validate the robustness of the method, LICTOR was used to classify 100 randomly selected LCs from the *hdnox* repertoire, which were not used in the development of the predictor. In this experiment LICTOR correctly classifies 80% of non-toxic sequences (Supplementary Table 4), confirming LICTOR as a suitable tool able to accurately predict LC toxicity on previously unseen LC sequences.

As a final test to assess the strength of LICTOR and further verify the absence of overfitting, the *tox* and *nox* sequences were randomly divided in two groups of the same size (*tox1* and *tox2*, respectively, *nox1* and *nox2*) and two classifiers were trained using these synthetic stets. Both predictors, the first trained using *tox1* and *tox2,* the second trained with *nox1* and *nox2*, obtained an AUC of 0.5 (Supplementary Table 5 and Supplementary Table 6). This result

further underlines that *tox* and *nox* sequences have distinctive features allowing their discrimination, thus reinforcing LICTOR as suitable tool to predict LC toxicity.

Finally, to identify the key features leading to LC toxicity in AL, we ranked the predictor variables of LICTOR according to their "information gain", a value representing the importance of the information carried by each predictor variable for the classification[20]. We found that, among the top-10 most important features of the three families of predictor variables, feature 49-A, which denotes a SM to alanine at position 49, obtained the highest score in the AMP family ranking, as well as in the general ranking (Fig. 2a and Supplementary Table 7). Indeed, feature 49-A was present in 54 *tox* sequences, while only in 8 *nox* sequences. Furthermore, the 49-A mutation, which is located at the dimeric interface of LCs (Fig. 2b), is also ranked among the top-10 features in the DAP family in combination with no substitutions at other residue position (Fig. 2a). Moreover, among the best-ranked predictor variables of the three families, those describing mutated positions are more frequent in *tox* sequences compared to *nox* (Fig. 2a). Interestingly, all these mutations are located at the LC homodimer interface (Fig. 2b), suggesting that mutations in these positions may affect the structural integrity of the dimeric interface and/or induce a local instability of the monomer, thus leading to LCs misfolding and aggregation. A similar trend is also visible for other top-ranked features, where unmutated positions are, conversely, more frequent in *nox* sequences than in *tox* (Fig. 2a, b). Taken together, these findings show that the presence or the absence of specific mutations at specific positions are key features used by LICTOR to classify LCs into toxic and non-toxic sequences, which further underlines the pivotal role of SMs in the development of LC toxicity in AL.

In conclusion, LICTOR represents the first method able to accurately predict LC toxicity. Hence, LICTOR may allow a timely identification of high-risk patients, such as MGUS patients likely to progress to AL, paving the way for early treatment and higher survival rates.

Furthermore, our approach may guide the development of novel predictive tools useful for other diseases, such cancer, in which the prognosis may depend on SMs of specific tumor-linked proteins. LICTOR is available as webservice at http:///lictor.irb.usi.ch.

FIGURE 1

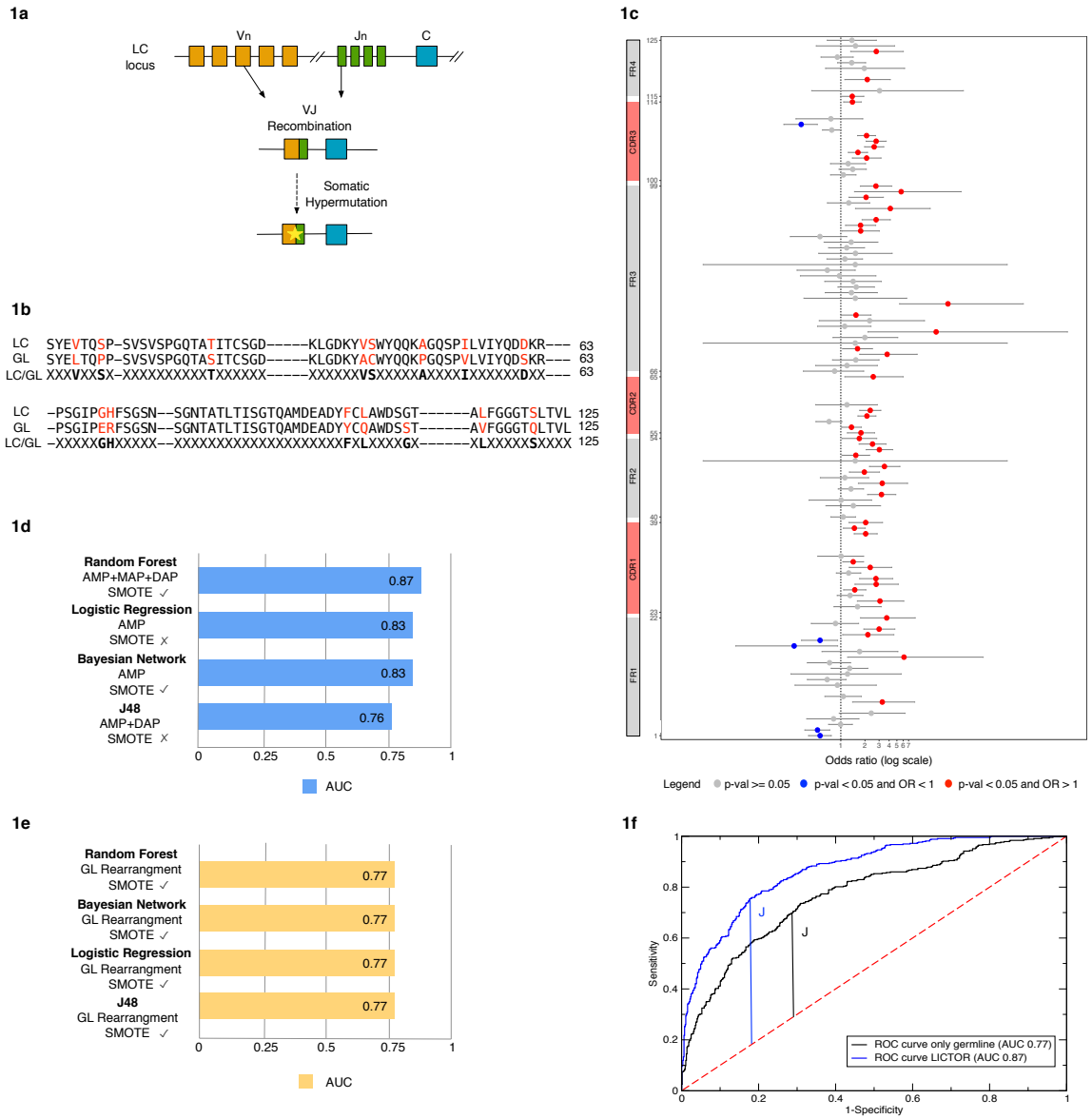

Fig. 1 | The presence of somatic mutations differentiates toxic and nontoxic LC sequences. a, Schematic representation of the generation of LC diversity through the processes of VJ recombination and somatic hypermutation. **b**, Alignment of a LC sequence with the corresponding germline, using a progressive Kabat-Chothia numbering scheme with a total of 125 positions. Residues in red depict somatic mutations. The third line shows the encoding scheme used by the classifier with somatic mutations (displayed in bold) and unmutated positions represented by a 'X'. **c**, OR for all 125 positions of the LC sequences (y-axis). Structural elements of immunoglobulin light chains are shown on the left. ORs for positions with no statistically significant difference between *tox* and *nox* sequences ($p \geq 0.05$) are represented as grey dots. Positions with statistically significant differences ($p < 0.05$) are depicted as either red (OR>1) or blue (OR<1) dots. Grey horizontal error-bars are the OR 0.95 confidence interval. **d**, AUC of the best configuration for each of the considered machine learner (blue bars). Different combinations of three families of predictor variables were tested, with (✓) or without (✗) the SMOTE balancing technique. **e**, The yellow bars show the best AUC value obtained by each machine learner using only the LC germline VJ rearrangements as predictor variable. **f**, ROC curve for LICTOR (i.e., random forest using AMP + MAP + DAP) compared with a predictor (random forest) using only the LC germline VJ rearrangements as predictor variable.

FIGURE 2

2a

Ranking	AMP	MAP	DAP
1	**49A** 54/8[a] I 2.8e-15[b] I 1[c]	**104X-108X** 178/410 I 2.9e-12 I 9	**49A-116X** 53/8 I 6.6e-15 I 2
2	**107X** 265/529 I 8.3e-13 I 8	**44X-98X** 336/599 I 6.2e-11 I 13	**65X-107X** 250/520 I 9.8e-15 I 3
3	**106X** 247/505 I 1.8e-12 I 11	**52X-65X** 323/585 I 6.2e-11 I 13	**4X-49A** 53/8 I 6.6e-15 I 4
4	**108H** 34/5 I 6.4e-10 I 19	**44X-99X** 310/571 I 8.9e-11 I 14	**49X-99X** 296/573 I 5.7e-15 I 5
5	**78X** 401/645 I 1.5e-09 I 22	**54X-58X** 193/419 I 2.6e-10 I 16	**52X-108X** 161/400 I 7.6e-15 I 6
6	**106N** 65/28 I 1.4e-09 I 24	**1X-108H** 28/2 I 6.3e-10 I 17	**49A-99X** 51/8 I 6.9e-14 I 7
7	**49X** 353/610 I 1.3e-09 I 28	**44H-96X** 61/24 I 6.1e-10 I 20	**3X-49A** 43/6 I 1.5e-12 I 10
8	**44X** 345/602 I 1.3e-09 I 29	**44X-97X** 317/575 I 5.2e-10 I 21	**66X-107X** 263/516 I 7.0e-11 I 15
9	**52X** 337/594 I 1.5e-09 I 31	**42X-52X** 323/580 I 1.0e-09 I 25	**49X-116X** 350/609 I 3.5e-10 I 18
10	**44H** 64/28 I 2.4e-09 I 35	**56X-59X** 211/493 I 2.2e-16 I 23	**52X-116X** 334/592 I 8.2e-10 I 23

2b

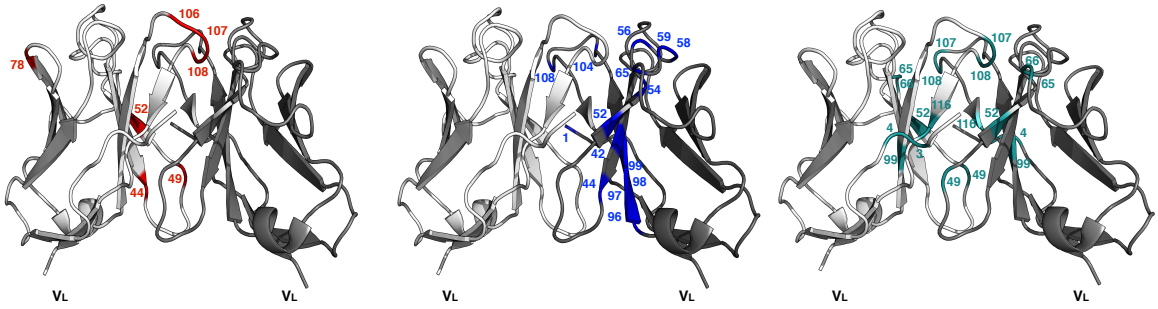

Fig. 2 | Top ranked features used to classify LC sequences. a, Top-10 features of each family ranked by information gain. Below each predictor variable are shown the occurrence in tox/nox sequences ([a]), the p-value ([b]) and the Feature Selection General Ranking ([c]) (red= AMP features, blue= MAP features, green=DAP features). **b,** Mapping of the top-10 features of each family on the variable domains of a LC homodimeric structure (PDB ID: 2OLD, in white and grey represented in cartoon). AMP features are shown in red in the left image, MAP features in blue in the middle image, while DAP in green in

the right one. The color code used in the table to represent the three feature families, is maintained in their structural representation in **b**.

Methods

Dataset. The database used in the training was composed of 428 *tox* and 590 *nox* sequences of λ isotype collected from the Amyloid Light-chain Database (ALBase) (http://albase.bumc.bu.edu). Furthermore, it contained 57 *nox* λ light chain sequences that we collected at the Institute for Research in Biomedicine (IRB-DB), known to be non-toxic in the context of AL. The 1,075 sequences were automatically aligned using a progressive Kabat-Chothia numbering scheme (http://www.bioinf.org.uk/abs/). According to this scheme, for example, the CDR1 of a given LC with Kabat-Chothia numbering 30A, 30B, 30C, 30D, 30E, and 30F, will be assigned to which 31, 32, 33, 34, 35, 36 and so on. For the ALBase's sequences the germlines' information were taken from the database, while for IRB-DB LCs germline were assessed with an in-house script. Next, germline sequences (GL) were reconstructed using the IMGT database[10].

The GL sequences were aligned with the same numbering scheme used for the LCs. Next, each light chain of the dataset was compared with the corresponding GL to identify all somatic mutations and the differences encoded using an X, for unmutated positions, and the LC amino acid for somatic mutations: this sequence was referred as S_{mut}. For example, for an LC with sequence *SYELTQPP* and its corresponding GL with the sequence *SYVLTQPP*, would be encoded as *XXEXXXXX*, since there is a somatic mutation $V \rightarrow E$ in position 3. To compare the presence of somatic mutations in S_{mut} at each position i in the Kabat-Chothia numbering scheme, the following four quantities were computed:

- tox^{i}_{NM} - the number of toxic sequences without somatic mutation in position *i;*

- tox^{i}_{M} - the number of toxic sequences with a somatic mutation in position *i;*
- nox^{i}_{NM} - the number of non-toxic sequences without somatic mutation in position *i;*
- nox^{i}_{M} - the number of non-toxic sequences with a somatic mutation in position *i;*

Statistical analysis. The *fisher.test* function from R version 3.5.1 with arguments *conf.int=TRUE* and *conf.level*=0.95 was used to assess the significant difference of somatic mutations in toxic and non-toxic sequences. The OR between tox^{i}_{M}/tox^{i}_{NM} and nox^{i}_{M}/nox^{i}_{NM} is computed as:

$$OR^{i}_{tox\text{-}nox} = \frac{tox^{i}_{M}/tox^{i}_{NM}}{nox^{i}_{M}/nox^{i}_{NM}}$$

OR=1 indicates that the event under study (i.e., the frequency of mutations at position *i*) is equally likely in the two groups (e.g., *tox* vs nox). OR>1 indicates that the event is more likely in the first group (*tox*). OR<1 indicates that the event is more likely in the second group. (*nox*).

Predictor variables used by the machine learners. Given a sequence, the following features were extracted:

Amino acid in each Mutated Position (AMP). From a sequence S_{mut}, a list of predictor variables was extracted, each one describing the type of amino acid added by the somatic mutation in a given position or the absence of mutation in the position. Thus, each of these variables is a pair (*position*, *amino acid*), where we use the letter "X" instead of the amino acid in the positions for which no somatic mutation was present.

Monomeric amino acid pairs (MAP). LCs share a conserved 3D structure. Therefore, pairs of interacting residues were defined as amino acids having a distance between the respective Cβ atoms smaller that 7.5 Å in X-ray structure (PDB ID: 2OLD).

Dimeric amino acid pairs (DAP). Similarly, pairs of residues that interact at the LC-LC interface were defined using the 2OLD LC homodimeric X-ray structure. Two residues belonging to different chains, were considered as interacting if the distance between their Cβ atoms was less than 7.5 Å.

Machine learning algorithms. Weka 3.8.1[15] implementation was used for the four machine learning algorithms (Bayesian network, logistic regression, J48, and random forest) to solve the classification task. For all algorithms, the default Weka parameters were used. The algorithms were evaluated by performing a 10-fold cross-validation over the dataset. The performance of each algorithm was first assessed only using one family of features (e.g., AMP, MAP, DAP, for a total of three combinations); second, the three families were combined into pairs (e.g., AMP U MAP, for a total of three combinations); third, all three families were combined together. This led to a total of 7 (features configuration) × 4 (algorithms) = 28 prediction experiments. Moreover, each of the 28 experiments was performed with and without the balancing of the training set with SMOTE (**S**ynthetic **M**inority **O**ver-sampling **TE**chnique)[16] on the toxic sequences so that the number of toxic instances was equal to the number of non-toxic ones in the training set during each of the ten cross-validations used in the evaluation. This led to 28 × 2 (with/without SMOTE) = 56 total experiments.

Prediction performance. The various prediction algorithms were assessed computing the following classifications errors: (i) Type I misclassifications, indicating toxic sequence wrongly classified as non-toxic (false negative—FN), and (ii) Type-II misclassifications, indicating non-toxic sequences misclassified as toxic (false positive—FP). The correct classifications are instead indicated by the number of true positive—TP (a toxic sequence correctly classified) and true negative—TN (a non-toxic sequence correctly classified). Based

on TP, TN, FP, and FN, the following metrics were used to evaluate the performance of our classifiers:

- *Area under the Receiver Operating Characteristic (AUC).* AUC is used to assess the performance of a two-class classifier (such as that in our study), and it is equal to the probability that the classifier will rank a randomly chosen positive instance (in our case, a toxic sequence) higher than a randomly chosen negative instance (non-toxic sequence). A random classifier has an AUC=0.5, while the AUC is 1.0 for a perfect classifier.
- *Sensitivity.* Computed as TP/(TP+FN): this represents the percentage of toxic sequences correctly identified by the classifier.
- *Specificity.* Computed as TN/(TN+FP): this represents the percentage of non-toxic sequences correctly identified by the classifier.

Youden index.

The Youden (J) index was used to validate the effectiveness of the predictors and to find the optimal cut-off point to separate toxic LCs associated with the disease from non-toxic LCs using the following formula:

$J=\max_c[Se(c) + Sp(c) -1]$

Information gain feature selection. InfoGainAttributeEval filter implemented in Weka 3.8.1[20] was used to remove all features that do not contribute to the information available for the prediction of the sequence type. All features having an information gain less than 0.01 were removed. Given the computational cost of this procedure, this experiment was performed for the best-performing algorithm and configuration identified in the previous 56 experiments. The full list of ranked features is shown in Supplementary Table 1.

Acknowledgements

This study was supported by a grant from the Swiss National Science Foundation (31003A-166472) to A.C. We would like to acknowledge the use of the Boston University ALBase, supported by HL68705, in this work.

Author contributions

A.C and M.G. designed research. M.G. performed data acquisition, analysis and drafted the manuscript. M.F., L.P. and A.L. provided data. M.G., L.P., S.R., M.F., M.M., J.S., M.P., O.M., A.L. and A.C. edited and approved the manuscript

References

1 Lavatelli, F. *et al.* A novel approach for the purification and proteomic analysis of pathogenic immunoglobulin free light chains from serum. *Biochim Biophys Acta* **1814**, 409-419, doi:10.1016/j.bbapap.2010.12.012 (2011).

2 Merlini, G. & Bellotti, V. Molecular mechanisms of amyloidosis. *N Engl J Med* **349**, 583-596, doi:10.1056/NEJMra023144 (2003).

3 Merlini, G. *et al.* Systemic immunoglobulin light chain amyloidosis. *Nat Rev Dis Primers* **4**, 38, doi:10.1038/s41572-018-0034-3 (2018).

4 Grogan, M., Dispenzieri, A. & Gertz, M. A. Light-chain cardiac amyloidosis: strategies to promote early diagnosis and cardiac response. *Heart* **103**, 1065-1072, doi:10.1136/heartjnl-2016-310704 (2017).

5 Merlini, G. Determining the significance of MGUS. *Blood* **123**, 305-307, doi:10.1182/blood-2013-12-539940 (2014).

6 Gertz, M. A. Immunoglobulin light chain amyloidosis diagnosis and treatment algorithm 2018. *Blood Cancer J* **8**, 44, doi:10.1038/s41408-018-0080-9 (2018).

7 Blancas-Mejia, L. M. & Ramirez-Alvarado, M. Systemic amyloidoses. *Annu Rev Biochem* **82**, 745-774, doi:10.1146/annurev-biochem-072611-130030 (2013).

8 Dispenzieri, A., Gertz, M. A. & Buadi, F. What do I need to know about immunoglobulin light chain (AL) amyloidosis? *Blood Rev* **26**, 137-154, doi:10.1016/j.blre.2012.03.001 (2012).

9 Bodi, K. *et al.* AL-Base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. *Amyloid* **16**, 1-8, doi:10.1080/13506120802676781 (2009).

10 Brochet, X., Lefranc, M. P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* **36**, W503-508, doi:10.1093/nar/gkn316 (2008).

11 Sheskin, D. *Handbook of parametric and nonparametric statistical procedures*. 4th edn, (Chapman & Hall/CRC, 2007).

12 DeKosky, B. J. *et al.* Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* **113**, E2636-2645, doi:10.1073/pnas.1525510113 (2016).

13 Makino, D. L., Henschen-Edman, A. H., Larson, S. B. & McPherson, A. Bence Jones KWR protein structures determined by X-ray crystallography. *Acta Crystallogr D Biol Crystallogr* **63**, 780-792, doi:10.1107/S0907444907021981 (2007).

14 Oberti, L. *et al.* Concurrent structural and biophysical traits link with immunoglobulin light chains amyloid propensity. *Sci Rep* **7**, 16809, doi:10.1038/s41598-017-16953-7 (2017).

15 Hall M., F. E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10-18 (2009).

16 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* **16**, 321-357, doi:DOI 10.1613/jair.953 (2002).

17 Abraham, R. S. *et al.* Immunoglobulin light chain variable (V) region genes influence clinical presentation and outcome in light chain-associated amyloidosis (AL). *Blood* **101**, 3801-3808, doi:10.1182/blood-2002-09-2707 (2003).

18 Comenzo, R. L., Zhang, Y., Martinez, C., Osman, K. & Herrera, G. A. The tropism of organ involvement in primary systemic amyloidosis: contributions of Ig V(L) germ line gene use and clonal plasma cell burden. *Blood* **98**, 714-720, doi:10.1182/blood.v98.3.714 (2001).

19 Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32-35, doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3 (1950).

20 Witten, I. H., Frank, E. & Hall, M. A. *Data mining : practical machine learning tools and techniques*. 3rd edn, (Morgan Kaufmann, 2011).