Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions

Ranko Gacesa [1] , David J Barlow [1] , Paul F Long Corresp. [1, 2, 3, 4]

[1] Institute of Pharmaceutical Science, King's College London, London, United Kingdom

[2] Department of Chemistry, King's College London, London, United Kingdom

[3] Brazil Institute, King's College London, London, United Kingdom

[4] Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, São Paulo, Brazil

Corresponding Author: Paul F Long
Email address: paul.long@kcl.ac.uk

Ascribing function to sequence in the absence of biological data is an on-going challenge in bioinformatics. Differentiating the toxins of venomous animals from homologues having other physiological functions is particularly problematic as there are no universally accepted methods by which to attribute toxin function using sequence data alone. Bioinformatics tools that do exist are difficult to implement for researchers with little bioinformatics training. Here we announce a machine learning tool called 'ToxClassifier' that enables simple and consistent discrimination of toxins from non-toxin sequences with > 99% accuracy and compare it to commonly used toxin annotation methods. 'ToxClassifer' also reports the best-hit annotation allowing placement of a toxin into the most appropriate toxin protein family, or relates it to a non-toxic protein having the closest homology, giving enhanced curation of existing biological databases and new venomics projects. 'ToxClassifier' is available for free, either to download (https://github.com/rgacesa/ToxClassifier) or to use on a web-based server (http://bioserv7.bioinfo.pbf.hr/ToxClassifier/).

1   **Machine learning can differentiate venom toxins from other proteins having non-toxic**

2   **physiological functions**

4   Ranko Gacesa[1], David J. Barlow[1, 2], Paul F. Long[1-4*]

6   [1]Institute of Pharmaceutical Science, King's College London, 150 Stamford Street, London SE1

7   9NH, United Kingdom.

8   [2]Department of Chemistry & [3]Brazil Institute, King's College London, Strand, London WC2R

9   2LS, United Kingdom.

10 [4]Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, Av. Prof. Lineu Prestes,

11 580, B16, 05508-000 São Paulo, SP, Brazil.

13 *Corresponding author:

14 Prof Paul F. Long

15 email: paul.long@kcl.ac.uk

16 tel/fax: 00 44 (0)20 7484 4842

18   **Abstract**

19   Ascribing function to sequence in the absence of biological data is an on-going challenge

20   in bioinformatics. Differentiating the toxins of venomous animals from homologues having other

21   physiological functions is particularly problematic as there are no universally accepted methods

22   by which to attribute toxin function using sequence data alone. Bioinformatics tools that do exist

23   are difficult to implement for researchers with little bioinformatics training. Here we announce a

24   machine learning tool called 'ToxClassifier' that enables simple and consistent discrimination of

25   toxins from non-toxin sequences with > 99% accuracy and compare it to commonly used toxin

26   annotation methods. 'ToxClassifer' also reports the best-hit annotation allowing placement of a

27   toxin into the most appropriate toxin protein family, or relates it to a non-toxic protein having the

28   closest homology, giving enhanced curation of existing biological databases and new venomics

29   projects.   'ToxClassifier'   is   available   for   free,   either   to

30   download (https://github.com/rgacesa/ToxClassifier)   or   to   use   on   a   web-based

31   server (http://bioserv7.bioinfo.pbf.hr/ToxClassifier/).

33   **Introduction**

34   Falling costs of tandem mass spectrometry for shotgun proteomics have made generating

35   vast amounts of protein sequence data increasingly affordable, yet the gap between obtaining

36   these sequences and then assigning biological function to them continues to widen (Bromberg et

37   al. 2009). Often, most sequences are deposited into protein databases with little, if any,

38   accompanying experimental data from which biological functions can be inferred. Customarily,

39   biological function is deduced indirectly by comparing amino acid sequence similarity to other

40   proteins in large databases to calculate a ranking of proteins with respect to the query sequence.

41   Using simple pair-wise comparisons as a sequence searching procedure, the BLAST suite of

42   programs (for example, BLASTp) was first of its kind and has gained almost unprecedented

43   acceptance among scientists (Neumann et al. 2014). Variations of the BLAST algorithm (for

44   instance, PSI-BLAST (Altschul et al. 1997)) and development of probabilistic models (such as

45   hidden Markov models, HMMs (Krogh et al. 1994)) use multiple sequence alignments to detect

46   conserved sequences (also referred to as motifs). Use of these models enables detection of

47   remote homology between proteins seemingly unrelated when analysed by pairwise alignment

48   alone (Krogh et al. 1994). Conversely, development of accurate algorithms and fast software

49   tools that can automatically identify critical amino acid residues responsible for differences in

50   protein function amongst sequences having exceptionally high sequence similarity remains a

51   challenging problem for bioinformatics. In a post-genomic era, the toxins of venomous animals

52   are an emerging paradigm.

53   Animal venoms comprise predominantly toxic peptides and proteins. Duplication of

54   genes that formerly encoded peptides and proteins having non-toxic physiological functions is

55   one of the foremost evolutionary drivers that gives rise to the enormous functional diversity seen

56   in animal venom toxins (Fry 2005; Chang & Duda 2012). However, evidence is ambiguous as to

57   whether these genes were expressed in multiple body tissues, with the duplicate copy then

58   recruited into venom glands with subsequent neo-functionalisation to develop toxicity, or if there

59   was duplication with ensuing sub-functionalisation of genes encoding pre-existing but non-toxic

60   venom gland proteins (Hargreaves et al. 2014). Examples of both mechanisms have been

61   demonstrated in different venomous animals (Reyes-Velasco et al. 2015; Vonk et al. 2013;

62   Junqueira-de-Azevedo et al. 2015). Nonetheless, many toxin proteins that constitute venoms

63   share a remarkable similarity to other proteins with non-toxic physiological functions, and

64   deciphering sequence data to disentangle these functions is not a trivial task (Kaas & Craik

65   2015).

66   Previous proteomic data from our laboratory and subsequent results of others realised an

67   astonishingly high sequence similarity between cnidarian (jellyfish, coral, sea anemones etc.)

68   toxins and those of other higher venomous animals (Weston et al. 2012; Weston et al. 2013; Li et

69   al. 2012; Li et al. 2014). This suggested to us that a small number of sequences, when occurring

70   in combination, might explain this similarity and prompted the search for toxin-specific motifs

71   (Starcevic & Long 2013). An unsupervised procedure was developed that resulted in the

72   identification of motifs we called 'tox-bits', and which could describe most toxins as

73   combinations of between 2–3 'tox-bits' (Starcevic et al. 2015). The 'tox-bits' are defined as

74   HMM-profiles and were found to be superior at differentiating toxin from non-toxin sequences,

75   when compared against standard BLAST or HMM based methods (Starcevic et al. 2015).

76   However, implementation of 'tox-bits' HMM profiles is not straightforward for scientists with

77   little or no bioinformatics experience. Hence, in this paper we introduce and make freely

78   available an easy-to-use machine learning tool called the 'ToxClassifier' that employs 'tox-bits'

79   HMM profiles and other standard classifier tools running in parallel to distinguish toxins from

80   their non-toxic homologues.

81

82

83 **Methods**

84 *Datasets*

85 Data for training and testing of machine learning classifiers used in ToxClassifier ensemble was

86 obtained from UniProtKB database (Bateman et al. 2015), according to the following

87 methodology:

88 1) 'Positive' dataset representing well annotated putative animal toxins was downloaded from

89   UniProtKB/SwissProt-ToxProt (Jungo et al. 2012). Database was searched for animal toxins

90   and venoms, using search query: *taxonomy:"Metazoa [33208]" (keyword:toxin OR*

91   *annotation:(type:"tissue specificity" venom))*. All duplicate entries with identical sequence or

92   sequence identifier were removed, resulting in 8,093 sequences.

93 2) 'Easy' negative dataset representative of physiological proteins was obtained by random

94   sampling of 50,000 sequences in UniProtKB/SwissProt database (Bateman et al. 2015). All

95   entries with duplicate sequence identifier or protein sequence were removed, as were all

96   entries also occuring in *Positive dataset*; final dataset included 47,144 protein sequences.

97 3) 'Moderate difficulty' negative dataset was designed to match highly curated toxin-like

98   proteins with physiological function; it was created by BLASTp searching

99   UniProtKB/SwissProt database with *Positive dataset*, with e-value cutoff of 1.0e-10.

100   Resulting BLASTp hits were collected and all duplicates (with identical sequence or

101   sequence identifier), and all sequences also present in *Positive dataset* or *Easy dataset* were

102   removed, resulting in 8,034 proteins.

103 4) 'Hard' negative dataset was constructed from TrEMBL database (Bairoch & Apweiler 2000)

104   instead of Swiss-Prot. As with *Moderate dataset*, it was created from results of BLASTp

105   using *Positive dataset* as query and TrEMBL as target database. Duplicates and sequences

106   also occurring in *Positive, Easy* or *Moderate datasets* were removed for total of 7,403

107   sequences.

108   All datasets are available for download at https://github.com/rgacesa/ToxClassifier/datasets.

109   *Machine learning classifiers, training and testing*

110   Models describing protein sequences were constructed as follows:

111   1) *Single Amino acid frequency model* (OF): model uses length of sequence and frequency of

112   each amino acid as input features.

113   2) *Amino acid dimer frequency model* (BIF): model uses length of sequence, frequency of each

114   amino acid and of each amino acid 2-mer.

115   3) *Naive tox-bits model* (NTB): input features for this model are the number of 'tox-bits' for

116   each 'tox-bits' HMM listed in the 'tox-bits' database (Starcevic et al. 2015).

117   4) *Scored 'tox-bits' model* (STB): STB is a modification of the NTB model, with HMM bit-

118   scores replacing the number of 'tox-bits' in each 'tox-bit' HMM model.

119   5) *Tri-Blast Simple* (TBS) model: TBS uses BLASTp searches against positive

120   (UniProtKB/SwissProt-ToxProt) and two negative control databases (close non-toxins from

121   UniProtKB/SwissProt and non-toxins from TrEMBL); features include *bit-score, query*

122   *length*, *subject length*, *query/subject length ratio*, *query coverage*, *percentage of identity*,

123   *percentage of positive matches;* features also include amino-acid frequencies. Scores are

124   computed from the 'best-hit' in each database, with a BLAST e-value of 1.0e-10.

125   6) *Tri-Blast Enhanced A* (TBEa) model: TBEa model is an expanded variant of TBS, with

126   amino dimer frequencies included in the model.

127   7) *Tri-Blast Enhanced B* (TBEb): model is a variation of TBEa, trained on 80% of the input

128   dataset and with a BLAST e-value cut-off value of 1.0e+3 for the detection of similar toxin

129   or non-toxic sequences.

130   Feature extraction and vectorisation was implemented using the Python 2.7

131   (https://www.python.org/download/releases/2.7/) scripting language, NCBI BLAST+ version

132   2.2.31 (Camacho et al. 2009), HMMER 3.1b1 (Eddy 2011), the 'tox-bit' HMM profile database

133   (Starcevic et al. 2015) and a set of custom BLAST databases based on UniProtKB/SwissProt-

134   ToxProt, UniProtKB/SwissProt and TrEMBL databases. Vectorization scripts, vectorised data

135   sets and trained models are available for download at https://github.com/rgacesa/ToxClassifier.

136   Support Vector Machine (SVM), Gradient Boosted Machine (GBM) and Generalised Linear

137   Model (GLM) classifiers were trained for each of the models; Classifiers were implemented

138   using the R-programming language, Caret package (http://topepo.github.io/caret/index.html).

139   The training set for each dataset was selected by random sampling of 75% of the sequences, and

140   combined training sets were used to train the classifiers. Input features were 0-centered and

141   scaled by standard deviation and training was performed with 10 internal bootstraps. Learning

142   curves were constructed for each classifier to evaluate training efficiency and potential

143   overtraining by plotting performance versus number of sequences in training set.

144   Classifiers were tested using those sequences from *Positive, Easy, Moderate* and *Hard* datasets

145   not used in training. Performance was evaluated on each of the datasets and on the summary

146   dataset. The following performance measurements were calculated:

147   1) *Number of true positives* (TP): Number of toxic sequences in dataset correctly predicted as

148   toxins; sequence was considered a toxin if listed in UniProtKB/ToxProt database.

149   2) *Number of true negative*s (TN): Number of non-toxic sequences in dataset which are

150   correctly predicted as non-toxic (not in UniProtKB/ToxProt database).

151   3) *Number of false positive* (FP): Number of non-toxic sequences in dataset incorrectly

152   classified as toxins.

153   4) *Number of false negatives* (FN): Number of toxic sequences in dataset incorrectly classified

154   as non-toxic.

155   5) *Accuracy* (ACC): Accuracy is calculated as proportion of sequences correctly classified as

156   toxins or non-toxins; ACC = (TP+TN) / (TP+TN+FP+FN).

157   6) *Specificity* (SPEC): Also called *true negative rate*, specificity is proportion of correctly

158   predicted non-toxins (true negatives); SPEC = TN / (TN+FP).

159   7) *Sensitivity* (SENS): Also called *true positive rate* or *recall*, sensitivity is proportion of

160   correctly predicted toxins (true positives); SENS = TP / (TP+FN).

161   8) *Balanced Accuracy* (B.ACC): Balanced accuracy is a mean value of specificity and

162   sensitivity; B.ACC = (SPEC+SENS)/2.

163   9) *Negative predictive value* (NPV): Proportion of negatives that are true negatives. NPV = TN

164   / (TN+FN).

165   10) *Positive predictive value* (PPV): Also called Precision, PPV measures proportion of positives

166   which are true positives; PPV = TP / (TP+FP).

167   11) *F-Score* (F1): F-score is harmonic mean of precision and sensitivity and represents weighted

168   average of precision and recall. It is calculated as F1 = 2*TP / (2*TP+FP+FN).

169   12) Matthews' correlation coefficient (MCC): The MCC value (also known as phi-coefficient) is

170   a measure of correlation between observed and predicted. It is considered balanced measure

171   even for classes with different amounts of positive and negative values; MCC =

172    $\frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$ (Matthews 1975; Powers 2011).

173   *Conventional annotation models*

174   Annotation models simulating manual annotation were constructed based on BLAST and

175   HMMER, as follows:

176   1) Naive-BLAST: Annotation method is based on the assumption that a sequence is a toxin if a

177   BLAST search against the UniProtKB/SwissProt-ToxProt dataset returns a positive hit at a

178   certain e-value (usually between 1.0e-20 and 1.0e-5). This approach and its variations are

179   commonly encountered in the literature (Sher & Zlotkin 2009; Schwartz et al. 2015; Liu et al.

180   2015; Whittington et al. 2010).

181   2) OneBLAST: This approach classifies a sequence by a single BLAST search against

182   databases including UniProtKB/SwissProt-ToxProt and other 'toxin-like' but non-toxic

183   sequences extracted from UniProtKB/SwissProt and TrEMBL databases (combination of

184   *Positive, Moderate* and *Hard* datasets). A sequence is classified as a putative toxin if the

185   highest scoring BLAST hit at a selected e-value is from UniProtKB/SwissProt-ToxProt; it is

186   classified as non-toxic if top BLAST hit is not from UniProtKB/SwissProt-ToxProt, or if all

187   BLAST hits are above the selected e-value.

188   3) TriBLAST: Annotation is performed by a variation of OneBLAST, using separate BLAST

189   searches against UniProtKB/SwissProt-ToxProt and two non-toxin databases (for the

190   *Moderate* and *Hard* datasets); a sequence is classified as a toxin if the BLAST search against

191   UniProtKB/SwissProt-ToxProt returns a hit below a certain e-value and with a higher score

192   than the searches against the other databases. Variations of this method are commonly used

193   in toxin annotation (Gacesa et al. 2015; Rachamim et al. 2014).

194   4)   hmmerToxBits: This method is a variation of the naive-BLAST, using HMMER package

195   Hmmsearch instead of BLAST, and the database of 'tox-bits' HMM models as the target

196   database. A sequence is classified as a toxin if one or more HMMs can be detected within a

197   certain e-value cut-off.

198   5)   hmmerVenom: Modification of hmmerToxBits, the method uses HMM profiles extracted by

199   'venom' and 'toxin' text search of the Pfam database instead of 'tox-bit' HMMs.

200   6)   twinHmmerPfam: A HMMER based variant of TriBLAST approach, this method performs

201   hmmsearches against two HMM databases (toxin/positive HMMs and negative control) and

202   compares bitscores. Sequence is annotated as a toxin if bitscore for toxin HMM database is

203   higher. TwinHmmerPfam toxin HMMs were extracted from Pfam by keyword search for

204   'toxin' and 'venom', while negative control database is comprised from remainder of Pfam.

205   7)   twinHmmerToxBits: A variation of twinHmmerPfam method, method compares hmmsearch

206   against toxin HMMs and negative control. For this model, positive database is composed

207   from 'tox-bit' HMMs derived from Tox-Prot toxins (Starcevic et al. 2015), while negative

208   control is Pfam database from which HMMs containing 'toxin' or 'venom' keywords were

209   removed.

210   BLAST databases for Naive-BLAST, OneBLAST and TriBLAST were constructed from 75%

211   randomly sampled sequences in *Positive, Easy, Moderate* and *Hard* and performance was

212   measured by annotating remainder of data. HMMER based methods were tested with 25% of

213   random sequences from input datasets, using all appropriate HMM models. Database

214   construction and testing was repeated 10 times and results were averaged.

215   *ToxClassifier Meta-classifier calibration and testing*

216   ToxClassifier meta-classifier was constructed from nine annotation model and classifier

217   combinations (BIF_SVM, BIF_GBM, STB_SVM, STB_GBM, TBS_SVM, TBS_SVM,

218   TBEa_SVM, TBEb_SVM and TBEb_GBM). Each of nine classifiers reports 1 if the input

219   sequence is predicted as toxin or 0 if predicted as non-toxic, for final prediction score of 0 to 9.

220   Datasets used for calibration of meta-classifier were chosen from the set of venomous and non-

221   venomous animals (human *Homo sapiens*, the house mouse *Mus musculus,* the Burmese python

222   *Python bivittatus*, king cobra *Ophiophagus hannah*, the duck-billed platypus *Ornithorhynchus*

223   *anatinus*, the snakelocks sea anemone *Anemonia viridis*, the starlet sea anemone *Nematostella*

224   *vectensis*; and all proteins deposited in the UniProtKB/SwissProt and TrEMBL databases

225   attributed to snakes, spiders, wasps and *Conus* snails). All sequences were downloaded from

226   UniProtKB database with exception of *Python bivittatus* which was not available in UniProtKB

227   and was downloaded from NCBI protein database (Wheeler et al. 2003); all data is available at

228   https://github.com/rgacesa/ToxClassifier/datasets.   Datasets   were   split   into   training   sets

229   consisting of 75% of data and test sets including remaining 25% of sequences.

230   ToxClassifier meta-classifier was calibrated by evaluating prediction score versus performance

231   for each animal training set and for summary dataset constructed by combining animal training

232   datasets with exclusion of *Conus snail* data, which was dropped due to suspected low quality of

233   annotation. Calibrated ToxClassifier, with prediction score 5 or more as cut-off for positive

234   classification, was tested on animal data test sets, and performance measures were compared to

235   OneBLAST, NaiveBLAST models and ClanTox server. Data used for training and testing and all

236   calculated   performance   metrics   are   available   at

237   https://github.com/rgacesa/ToxClassifier/datasets/toxclassifier_calibration_test.

238   *ToxClassifier comparison to other published tools*

239   Performance of ClanTox server (Kaplan et al. 2007) was tested on *Positive, Easy, Moderate* and

240   *Hard* datasets and on *animal data test sets*. ToxinPred server (Gupta et al. 2013) was tested on

241   868 *Positive dataset* sequences of length up to 30 amino acids (ToxinPred sequence length limit)

242   and on negative dataset composed 30 amino acid or shorter protein sequences randomly selected

243   from UniProt database (5,673 non-duplicate, non-ToxProt sequences). ToxinPred was not tested

244   on animal data due to lack of short sequences available in these datasets. SpiderP server (Wong

245   et al. 2013) was not tested as service was not available at the time and PredCSF server (Fan et al.

246   2016) was not tested as it was deemed too specialized and only accepts single sequence as input.

247   *User interface*

248   The   ToxClassifier   web   service   front-end   is   implemented   using   HTML   5.1

249   (https://www.w3.org/html/),   JavaScript   (https://www.javascript.com/),   jQuery

250   (https://jquery.com/),   CSS   (https://www.w3.org/Style/CSS/),   Java   7.0

251   (https://www.oracle.com/java/index.html) and the Java Server Pages (JSP 2.1) framework

252   (http://www.oracle.com/technetwork/java/javaee/jsp/index.html). Visualisation is performed

253   using   R   (https://www.r-project.org/),   ggplot2   (http://ggplot2.org/)   and   R   markdown

254   (http://rmarkdown.rstudio.com/) packages. ToxClassifier runs on an Apache Tomcat 8.0 web

255   server (http://tomcat.apache.org/download-80.cgi), under an Ubuntu Linux 12.04 operating

256   system (http://www.ubuntu.com/). The service is hosted by the Section of Bioinformatics,

257   Faculty of Food Technology and Biotechnology, University of Zagreb, Croatia

258   (http://www.pbf.unizg.hr/en/departments/department_of_biochemical_engineering/section_for_b

259   ioinformatics).

260   **Results**

261   The accuracy of 3 individual machine-learning classifiers to predict toxins from proteins

262   having other physiological functions was assessed by training each classifier using 7 different

263   annotation models. The learning classifiers were a Support Vector Machine (SVM) and Gradient

264   Boosted Machine (GBM) chosen as high-performing predictors, and a Generalised Linear Model

265   (GLM) regarded as a simple classifier, but with which a baseline could be established that would

266   allow comparison of the performance of the SVM and GBM machines. A detailed description of

267   the annotation models is given in Methods section, but briefly the annotation models used the

268   following sequence information from the training set as classifier inputs: either the frequency of

269   amino acids (TBSim) or combinations of two amino-acids (BIF), the presence of absence or

270   'tox-bits' (SToxA), HMM scores for 'tox-bits' (SToxB), a selection of BLAST output co-

271   variants (TBEa) or a variation on TBSim and TBEa (TBEb).

272   The training set was constructed by merging 75% of arbitrarily selected sequences from

273   each of the following datasets: 1/ A 'Positive' dataset that contained all 8,093 protein sequences

274   deposited in the UniProtKB/SwissProt-ToxProt database of animal venom toxins (Jungo et al.

275   2012). 2/ An 'Easy' dataset composed of 47,144 random, non-duplicate sequences from

276   UniProtKB/SwissProt database (Bateman et al. 2015). 3/ A 'Moderate' dataset comprised of

277   8,034 unique sequences from the manually curated UniProtKB/SwissProt database considered to

278   be non-toxic but with homology to toxin proteins in UniProtKB/SwissProt-ToxProt. 4/ A 'Hard'

279   dataset that included 7,403 non-duplicate sequences extracted from the computer annotated

280   TrEMBL (Bairoch & Apweiler 2000) database, also considered to be non-toxic but with

281   homology to animal venom toxins in UniProtKB/SwissProt-ToxProt.

282   All training was performed using 10 internal bootstrap cross-validations on the training

283   set, and learning curves showing the accuracy of predictions versus the number of sequences in

284   the training sets were constructed, thereby allowing a comparative evaluation of training

285   efficiency (Fig. S1). The trained classifiers were then tested for prediction accuracy using the

286   remaining 25% of sequences not included in the training set. Performance values were calculated

287   to give an overall comparative classification of protein sequences as toxins or non-toxins (Table

288   1). By comparing learning curves (Fig. S1) and accuracy of predictions (Table 1), 9 of the

289   annotation model and classifier combinations were chosen to construct the 'ToxClassifier'

290   ensemble. The trained classifiers were: BIF_SVM, BIF_GBM, STB_SVM, STB_GBM,

291   TBS_SVM, TBS_SVM, TBEa_SVM, TBEb_SVM and TBEb_GBM. These classifiers all gave

292   excellent accuracy scores to predict toxins from the Positive dataset (range 0.82 to 0.96) and

293   non-toxin proteins from the Easy, Moderate and Hard datasets (range 0.92 to 1.00). No GLM

294   classifiers were included in the ensemble because prediction accuracies were considerably lower

295   when compared with SVM and GBM machines. Classifiers using NTB and OF annotation

296   models were also abandoned in favour of better performing STB and BIF models. Furthermore,

297   the TBEa_GBM model consistently underperformed compared to the TBE_SVM model and was

298   excluded, giving an odd number of classifiers in the ensemble, thereby avoiding a 'tied vote'

299   scenario when the outputs were interpreted collectively.

300   The prediction accuracy of the trained machine learning classifiers was next compared to

301   more conventional annotation methods based on sequence alignment, to determine if machine

302   learning predictions were superior or inferior to well established and accepted bioinformatics

303   tools. A detailed description of the annotation models based on these bioinformatics tools is

304   given in Methods. Briefly, simple predictions were made by taking the best-hit from BLAST

305   comparisons between a query sequence and the UniProtKB/SwissProt-ToxProt database

306   (naiveBLAST method), or the best-hit following a HMMER hmmsearch comparison between a

307   query sequence and either existing HMM models for toxin protein families in the Pfam database

308   (hmmerVenom model), or our own 'tox-bits' HMM models (hmmerToxbits classifier). More

309   sophisticated annotation models also used BLAST or HMMER searches, but the best-hit was

310   extracted following simultaneous comparisons between the query sequence and multiple

311   datasets. These sophisticated annotation models are also described in Methods, but briefly these

312   models were constructed from sequence information extracted from either

313   UniProtKB/SwissProt-ToxProt sequences supplemented with additional toxin-like sequences

314   from the UniProtKB/SwissProt and TrEMBL databases, or UniProtKB/SwissProt-ToxProt

315   sequences supplemented with non-toxin sequences from the 'Moderate' and 'Hard' datasets used

316   to train the machine classifiers. Training and test sets were analogous in design and execution to

317   the machine classifier learning, with 75% of sequence information used to construct the BLAST

318   and HMMER databases and the remaining 25% of data used to evaluate performance. Prediction

319   accuracy measures for each query sequence using each of the bioinformatics models were

320   repeated 10 times to give a final balanced accuracy value. Accuracy measure calculations are

321   described in Methods. A range of sequence-alignment scoring was also tested to select the lowest

322   BLAST and HMMER cut-off scores that gave the most precise toxin annotation. This value was

323   1.0e-20 for both BLAST and HMMER searches (Fig. S2).

324   Machine learning classifiers were also evaluated against currently available published

325   tools for toxin prediction and annotation; Animal toxin prediction server ClanTox (Kaplan et al.

326   2007) was benchmarked using *Positive, East, Moderate* and *Hard* datasets and summary of these

327   datasets. As ToxinPred (Gupta et al. 2013) tools predict only small peptide toxins, it was tested

328   using a subset of *Positive dataset* with sequences no longer than 30 amino acids (868 sequences)

329   and separate negative dataset composed of 5,673 random short proteins from UniProtKB

330   database. SpiderP (Wong et al. 2013) was not benchmarked as the server no longer seems

331   publically available. Finally, PredCSF (Fan et al. 2016) is a conotoxin-specific tool and was

332   deemed not comparable to general annotation tools; it also only allows single sequence input,

333   making it unsuitable for large scale testing. Final performance measures compared between the

334   different tools are listed in Table 2.

335   Testing of sequence-alignment based annotation models (Table 2) demonstrated that the

336   simplistic methods (naiveBLAST, hmmerToxBits and hmmerVenom) gave high prediction

337   accuracies for sequences in the Easy dataset (ranging from 0.95 for hmmerVenom to 0.99 for

338   naiveBLAST), but underperformed in annotation of the physiological toxin-like sequences in the

339   Moderate and Hard datasets (accuracies ranging from 0.74 to 0.83 for Moderate and 0.07 to 0.23

340   for Hard dataset (the poor performance here, also evinced by the low F1 and MCC scores). More

341   sophisticated BLAST-based methods (oneBLAST and triBLAST) gave very high prediction

342   accuracy scores (0.93 − 0.999) for sequences in the Easy and Moderate datasets, but somewhat

343   lower performance on sequences in the Positive and Hard datasets (0.86 − 0.90). Pfam-based

344   twinHMMER gave the highest accuracy prediction for non-toxin sequences, but underperformed

345   compared to the other annotation models against sequences in the positive toxin dataset

346   (accuracy 0.56). The 'tox-bits' based variant accurately predicted sequences in the Easy and

347   Moderate datasets (accuracy 0.85 − 0.999), but suffered from a high false positive rate when

348   sequences in the Hard dataset were analysed (accuracy 0.44). When compared to machine

349   learning-based methods, even the most accurate of the sequence alignment-based models

350   (oneBLAST and triBLAST) were surpassed by the majority of the machine learning based

351   classifiers, especially by TBEa and TBEb models (SVM and GBM variants), which gave the

352   highest accuracy of prediction for sequences in all test datasets. All prediction methods showed

353 higher performance for negative prediction (predicting non-toxin as non-toxin) compared to

354 positive prediction (correctly predicting toxin as toxic), with Specificity (Spec) and Negative

355 Prediction Value (NPV) significantly higher than Sensitivity (Sens) and Positive Prediction

356 Value (PPV).

357   Each of the 9 machine learning classifiers used in the 'ToxClassifier' ensemble gives a

358 simple bit (1 or 0) value as output to predict whether the likely biological activity of the input

359 sequence is as a toxin (1), or has a non-toxic (0) physiological role and scores are summed into

360 final prediction score ranging from 0 to 9. Evaluation of this final prediction score was

361 performed on test sets obtained from randomly sampling 75% of sequences in the published

362 annotated genomes from a selection of venomous animals (king cobra *Ophiophagus hannah*, the

363 duck-billed platypus *Ornithorhynchus anatinus*, the snakelocks sea anemone *Anemonia viridis*,

364 the starlet sea anemone *Nematostella vectensis*; and all proteins deposited in the

365 UniProtKB/SwissProt and TrEMBL databases attributed to snakes, spiders, wasps and *Conus*

366 snails) and other animals considered to be non-venomous (human *Homo sapiens*, the house

367 mouse *Mus musculus* and the Burmese python *Python bivittatus*). Calibration was performed by

368 assessing the performance measures of the Toxclassifier ensemble relative to prediction score;

369 calibration curves for summary of all animal genome data are presented in Figure 1. When the

370 average correct annotation of all input sequences for all genomes was calculated, a combined

371 score from 5 out of the 9 classifiers giving correct classification provided a good balance

372 between the detection of toxins and the filtering of non-toxins. Hence, a calibration for the

373 ToxClassifier ensemble was possible where an input sequence giving a combined score of > 6

374 would be considered a likely toxin, a combined score of < 3 would be regarded as non-toxic,

375 while an input sequence presenting with a score 4 or 5 would suggest a potential toxin, but would

376   require manual evaluation using additional tools, for example, InterProScan (Zdobnov &

377   Apweiler 2001).

378   Performance of calibrated 'ToxClassifier' meta-classifier was evaluated on a test set

379   comprising 25% of the animal genome data not used for calibration; these results were compared

380   to naiveBLAST and OneBLAST conventional methods and to ClanTox server for animal toxin

381   prediction (Kaplan et al. 2007). Performance measurements are reported in Table 3 and

382   comparison of F1-scores and MCC values across all datasets is presented in Figures 2, S3/A and

383   S3/B. Finally, the 'ToxClassifier' was assessed in a blinded experiment that used as input a set of

384   protein sequences derived from the venom gland transcriptome of the Amazonian rain forest pit

385   viper *Bothrops atrox* (Supplementary Data 1). The sequences had been annotated using standard

386   methods and manually inspected, with the biological activities of some also being authenticated

387   experimentally. The results of the 'ToxClassifier' predictions matched with the expert annotation

388   (Table 4).

389   **Discussion**

390   The continued decline in proteomics sequencing costs over recent years has led to an

391   explosion in venomics data characterising the toxic peptide and protein components in many

392   venomous animals (Kaas & Craik 2015). However, there is currently no widely accepted and

393   standard method for functional annotation of toxins from these data sources, leading to

394   inconsistent estimates for the number of toxins in the venom of the same animal. For example,

395   the venom of the duck-billed platypus *Ornithorhynchus anatinus* has only 6 toxins listed

396   following manual annotation in the latest release of the UniProtKB/SwissProt-ToxProt database

397   (11th May 2016), yet 107 putative toxins were identified by a simple pair-wise BLASTp search

398 using venom gland transcriptome sequences as input to search the UniProtKB/SwissProt ToxProt

399 database (Jungo et al. 2012). In addition to separate homology searching methods to interpret the

400 same data, many venomics projects now also include different manual filtering steps as part of

401 the annotation process (Rachamim et al. 2014; Gacesa et al. 2015), exacerbating the problem of

402 results verification.

403   In this study, a selection of machine learning-based classifiers implementing a range of

404 BLAST and HMMER-based annotation models were trained on datasets of known toxins,

405 protein sequences assumed to be non-toxic but with homology to known toxins, and predicted

406 proteins encoded in the genome, transcriptome or proteome of a range of venomous and non-

407 venomous animals. A comparison between the results presented in Tables 1-3 demonstrated that

408 the majority of the machine learning methods consistently out-performed standard bioinformatics

409 approaches of functional annotation. Interestingly, all tested methods demonstrated higher

410 performance for negative prediction (classification of non-toxic sequences) compared to positive

411 classification (prediction of toxic sequences as toxins). These results demonstrate that

412 differentiating between physiological toxin-like proteins and actual toxins is more difficult then

413 prediction of random proteins as non-toxic, which is to be expected considering the similarity

414 and common origin of many toxins and toxin-like sequences (Fry 2005; Chang & Duda 2012;

415 Hargreaves et al. 2014). As such, it is important to consider balanced performance measurements

416 when assessing toxin classifiers, with scores such as F1-score and MCC value (Matthews 1975;

417 Powers 2011) providing more appropriate measurements of performance than simple accuracy.

418 Another issue of toxin classification lies in unbalanced datasets, because most venomous animal

419 genomes encode less than 100 toxins and 20,000 to 30,000 physiological non-toxic proteins; as a

420 result, even a high performing method can generate a high number of false positive predictions.

421   For example, 99.5% correct prediction of non-toxins results in ~100 false positive toxins for an

422   average animal proteome, which is in fact more than the actual number of true toxins. In order to

423   minimize both of these problems, the ToxClassifier training scheme was conservative, using only

424   well-annotated toxins from UniProtKB/ToxProt database as positives, and while this might lead

425   to a somewhat lower positive prediction rate (due to missing likely toxins which are not

426   annotated as such), it does serve to minimise the false positive rate.

427   Notably, predictions were less accurate on some genome datasets, especially *Conus* snail

428   proteins, with low performance metrics observed for all tested annotation methods. This

429   discrepancy was likely caused by the assumption that sequences deposited in the

430   UniProtKB/SwissProt-ToxProt sequence are *bona fide* toxins, while sequences in the

431   UniProtKB/SwissProt and TrEMBL databases without 'toxin' or 'venom' keywords are not

432   toxins. Given that toxin activity is attributed to most sequences without biological validation, it is

433   likely that the training datasets almost certainly excluded a number of toxin sequences and

434   included some yet unknown toxins as non-toxic. Another limitation of the 'ToxClassifier' lies in

435   the inherent bias of the training sets; an underrepresentation in sequences from certain animal

436   lineages, particularly the basal Metazoa, e.g. Cnidaria, could lead to incorrect assignment and

437   suspicious quality of existing annotation of conotoxins is a reason to treat prediction on this

438   protein class with caution. To elevate these problems, 'ToxClassifier' has been designed to also

439   report sequences suspected to have closest homology to underrepresented taxa as 'suspicious

440   toxin' and recommends manual annotation with other tools, such as InterProScan (Zdobnov &

441   Apweiler 2001).

442   Use of machine learning for toxin prediction has been attempted before and a range of

443   such tools exists; however, most of the available tools are heavily specialised for toxins of

444   specific   animal   origins.   For   example,   SpiderP   (Wong   et   al.   2013)

445   (http://www.arachnoserver.org/spiderP.html) is a predictor for spider toxins while ToxinPred

446   (Gupta et al. 2013) (http://crdd.osdd.net/raghava/toxinpred/) predicts only small peptide toxins;

447   while ClanTox (Kaplan et al. 2007) (http://www.clantox.cs.huji.ac.il/tech.php) was trained only

448   on   an   ion-channel   toxin   dataset   and   PredCSF   (Fan   et   al.   2016)

449   (http://www.csbio.sjtu.edu.cn/bioinf/PredCSF/) is conotoxin specific. In addition, the reported

450   training set sizes are low (for example ClanTox was trained on ~600 ion channel toxins; the

451   ToxinPred toxin positive training set is 1805 sequences, while as of 11th May 2016, the

452   UniProtKB/SwissProt-ToxProt database contained ~6500 sequences). None of the currently

453   available machine learning methods also gives a comparison with other currently used accepted

454   bioinformatics   annotation   methods.   When   compared   to   ToxClassifier   and   conventional

455   annotation tools (Tables 2 and 3), ClanTox and ToxinPred tools were found to perform similar to

456   BLAST based methods, while ToxClassifier demonstrated higher performance across all metrics,

457   which is likely a result of comparatively larger training sets and combination of different internal

458   classifiers.

459   In addition to high performance, the user interface of the 'ToxClassifier' web service

460   reports the best-scoring hit annotation either to UniProtKB/SwissProt-ToxProt (allowing

461   placement of the toxin into the most appropriate toxin protein family), or to the best hit in

462   UniProtKB/SwissProt (giving the closest homology to a non-toxin protein). In summary, this

463   study has established baseline prediction accuracies for a selection of toxin annotation methods

464   and integrates these methods into an easy-to-use, high-precision, machine learning-based

465   classification system named 'ToxClassifier'.This tool offers a reliable and reproducible

466   framework for toxin annotation to enable standardised toxin prediction in venomics projects and

467   to allow for semi-automatic annotation or re-annotation of existing datasets.

468   **References**

469   Altschul SF., Madden TL., Schäffer AA., Zhang J., Zhang Z., Miller W., Lipman DJ. 1997.

470   Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

471   *Nucleic Acids Research* 25:3389–3402. DOI: 10.1093/nar/25.17.3389.

472   Bairoch A., Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement

473   TrEMBL in 2000. *Nucleic Acids Research* 28:45–48. DOI: 10.1093/nar/28.1.45.

474   Bateman A., Martin MJ., O'Donovan C., Magrane M., Apweiler R., Alpi E., Antunes R.,

475   Arganiska J., Bely B., Bingley M., Bonilla C., Britto R., Bursteinas B., Chavali G., Cibrian-

476   Uhalte E., Da Silva A., De Giorgi M., Dogan T., Fazzini F., Gane P., Castro LG., Garmiri

477   P., Hatton-Ellis E., Hieta R., Huntley R., Legge D., Liu W., Luo J., Macdougall A., Mutowo

478   P., Nightingale A., Orchard S., Pichler K., Poggioli D., Pundir S., Pureza L., Qi G.,

479   Rosanoff S., Saidi R., Sawford T., Shypitsyna A., Turner E., Volynkin V., Wardell T.,

480   Watkins X., Zellner H., Cowley A., Figueira L., Li W., McWilliam H., Lopez R., Xenarios

481   I., Bougueleret L., Bridge A., Poux S., Redaschi N., Aimo L., Argoud-Puy G., Auchincloss

482   A., Axelsen K., Bansal P., Baratin D., Blatter MC., Boeckmann B., Bolleman J., Boutet E.,

483   Breuza L., Casal-Casas C., De Castro E., Coudert E., Cuche B., Doche M., Dornevil D.,

484   Duvaud S., Estreicher A., Famiglietti L., Feuermann M., Gasteiger E., Gehant S., Gerritsen

485   V., Gos A., Gruaz-Gumowski N., Hinz U., Hulo C., Jungo F., Keller G., Lara V., Lemercier

486   P., Lieberherr D., Lombardot T., Martin X., Masson P., Morgat A., Neto T., Nouspikel N.,

487   Paesano S., Pedruzzi I., Pilbout S., Pozzato M., Pruess M., Rivoire C., Roechert B.,

488   Schneider M., Sigrist C., Sonesson K., Staehli S., Stutz A., Sundaram S., Tognolli M.,

489   Verbregue L., Veuthey AL., Wu CH., Arighi CN., Arminski L., Chen C., Chen Y.,

490   Garavelli JS., Huang H., Laiho K., McGarvey P., Natale DA., Suzek BE., Vinayaka CR.,

491   Wang Q., Wang Y., Yeh LS., Yerramalla MS., Zhang J. 2015. UniProt: A hub for protein

492   information. *Nucleic Acids Research* 43:D204–D212. DOI: 10.1093/nar/gku989.

493   Bromberg Y., Yachdav G., Ofran Y., Schneider R., Rost B. 2009. New in protein structure and

494   function annotation: hotspots, single nucleotide polymorphisms and the "Deep Web".

495   *Current Opinion in Drug Discovery & Development* 12:408–419.

496   Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009.

497   BLAST+: architecture and applications. *BMC Bioinformatics* 10:1–9. DOI: 10.1186/1471-

498   2105-10-421.

499   Chang D., Duda TF. 2012. Extensive and continuous duplication facilitates rapid evolution and

500   diversification of gene families. *Molecular Biology and Evolution* 29:2019–2029. DOI:

501   10.1093/molbev/mss068.

502   Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Computational Biology* 7:e1002195.

503   DOI: 10.1371/journal.pcbi.1002195.

504   Fan YX., Song J., Kong X., Shen H Bin. 2016. PredCSF: An integrated feature-based approach

505   for predicting conotoxin superfamily. *Protein & Peptide Letters* 18:261–267. DOI:

506   10.2174/092986611794578341.

507   Fry B. 2005. From genome to "venome": molecular origin and evolution of the snake venom

508   proteome inferred from phylogenetic analysis of toxin sequences and related body proteins.

509   *Genome Research* 15:403–420. DOI: 10.1101/gr.3228405.

510   Gacesa R., Chung R., Dunn SR., Weston AJ., Jaimes-Becerra A., Marques AC., Morandini AC.,

511   Hranueli D., Starcevic A., Ward M., Long PF. 2015. Gene duplications are extensive and

512   contribute significantly to the toxic proteome of nematocysts isolated from *Acropora*

513   *digitifera* (Cnidaria: Anthozoa: Scleractinia). *BMC Genomics* 16:774. DOI:

514   10.1186/s12864-015-1976-4.

515   Gupta S., Kapoor P., Chaudhary K., Gautam A., Kumar R., Raghava GPS. 2013. *In silico*

516   approach for predicting toxicity of peptides and proteins. *PLoS ONE* 8, e73957 DOI:

517   10.1371/journal.pone.0073957.

518   Hargreaves AD., Swain MT., Hegarty MJ., Logan DW., Mulley JF. 2014. Restriction and

519   recruitment - gene duplication and the origin and evolution of snake venom toxins. *Genome*

520   *Biology and Evolution* 6: 2088-2095. DOI: 10.1093/gbe/evu166.

521   Jungo F., Bougueleret L., Xenarios I., Poux S. 2012. The UniProtKB/Swiss-Prot Tox-Prot

522   program: A central hub of integrated venom protein data. *Toxicon* 60:551–557. DOI:

523   10.1016/j.toxicon.2012.03.010.

524   Junqueira-de-Azevedo ILM., Bastos CMV., Ho PL., Luna MS., Yamanouye N., Casewell NR.

525   2015. Venom-related transcripts from *Bothrops jararaca* tissues provide novel molecular

526   insights into the production and evolution of snake venom. *Molecular Biology and*

527   *Evolution* 32:754–66. DOI: 10.1093/molbev/msu337.

528   Kaas Q., Craik DJ. 2015. Bioinformatics-aided venomics. *Toxins* 7:2159–2187. DOI:

529   10.3390/toxins7062159.

530   Kaplan N., Morpurgo N., Linial M. 2007. Novel families of toxin-like peptides in insects and

531   mammals: A computational approach. *Journal of Molecular Biology* 369:553–566. DOI:

532   10.1016/j.jmb.2007.02.106.

533   Krogh A., Brown M., Mian IS., Sjölander K., Haussler D. 1994. Hidden markov models in

534   computational biology, applications to protein modeling. *Journal of Molecular Biology*

535   235:1501–1531. DOI: 10.1006/jmbi.1994.1104.

536   Li R., Yu H., Xing R., Liu S., Qing Y., Li K., Li B., Meng X., Cui J., Li P. 2012. Application of

537   nanoLC-MS/MS to the shotgun proteomic analysis of the nematocyst proteins from jellyfish

538   *Stomolophus meleagris*. *Journal of Chromatography B: Analytical Technologies in the*

539   *Biomedical and Life Sciences* 899:86–95. DOI: 10.1016/j.jchromb.2012.05.006.

540   Li R., Yu H., Xue W., Yue Y., Liu S., Xing R., Li P. 2014. Jellyfish venomics and venom gland

541   transcriptomics analysis of *Stomolophus meleagris* to reveal the toxins associated with

542   sting. *Journal of Proteomics* 106:17–29. DOI: 10.1016/j.jprot.2014.04.011.

543   Liu G., Zhou Y., Liu D., Wang Q., Ruan Z., He Q., Zhang L. 2015. Global transcriptome

544   analysis of the tentacle of the Jellyfish *Cyanea capillata* using deep sequencing and

545   expressed sequence tags: Insight into the toxin-and degenerative disease-related transcripts.

546   *PLoS ONE* 10:1–22. DOI: 10.1371/journal.pone.0142680.

547   Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage

548   lysozyme. *BBA - Protein Structure* 405:442–451. DOI: 10.1016/0005-2795(75)90109-9.

549   Neumann RS., Kumar S., Shalchian-Tabrizi K. 2014. BLAST output visualization in the new

550   sequencing era. *Briefings in Bioinformatics* 15:484–503. DOI: 10.1093/bib/bbt009.

551   Powers D. 2011. Evaluation: From precision, recall and f-measure to roc., informedness,

552   markedness & correlation. *Journal of Machine Learning Technologies* 2:37 − 63.

553   Rachamim T., Morgenstern D., Aharonovich D., Brekhman V., Lotan T., Sher D. 2014. The

554   dynamically evolving nematocyst content of an anthozoan, a scyphozoan, and a hydrozoan.

555   *Molecular Biology and Evolution* 32:740–753. DOI: 10.1093/molbev/msu335.

556   Reyes-Velasco J., Card DC., Andrew AL., Shaney KJ., Adams RH., Schield DR., Casewell NR.,

557   Mackessy SP., Castoe TA. 2015. Expression of venom gene homologs in diverse python

558   tissues suggests a new model for the evolution of snake venom. *Molecular Biology and*

559   *Evolution* 32:173–183. DOI: 10.1093/molbev/msu294.

560   Schwartz EF., Diego-Garcia E., Rodríguez de la Vega RC., Possani LD., Whittington CM.,

561   Papenfuss AT., Locke DP., Mardis ER., Wilson RK., Abubucker S., Mitreva M., Wong

562   ESW., Hsu AL., Kuchel PW., Belov K., Warren WC., Sher D., Zlotkin E., Knebel A., Bsor

563   T., Nesher N., Tal T., Morgenstern D., Cohen E., Fishman Y., Zlotkin E., Schwartz EF.,

564   Diego-Garcia E., Rodríguez de la Vega RC., Possani LD., Liu G., Zhou Y., Liu D., Wang

565   Q., Ruan Z., He Q., Zhang L. 2015. Transcriptome analysis of the venom gland of the

566   Mexican scorpion *Hadrurus gertschi* (Arachnida: Scorpiones). *Toxicon* 11:865–879. DOI:

567   10.1371/journal.pone.0142680.

568   Sher D., Zlotkin E. 2009. A hydra with many heads: Protein and polypeptide toxins from hydra

569   and their biological roles. *Toxicon* 54:1148–1161. DOI: 10.1016/j.toxicon.2009.02.036.

570   Starcevic A., Moura-Da-Silva AM., Cullum J., Hranueli D., Long PF. 2015. Combinations of

571   long peptide sequence blocks can be used to describe toxin diversification in venomous

572   animals. *Toxicon* 95:84–92. DOI: 10.1016/j.toxicon.2015.01.005.

573   Starcevic A., Long PF. 2013. Diversification of animal venom peptides-were jellyfish amongst

574   the first combinatorial chemists? *ChemBioChem* 14:1407–1409. DOI:

575   10.1002/cbic.201300305.

576   Vonk FJ., Casewell NR., Henkel C V., Heimberg AM., Jansen HJ., McCleary RJR., Kerkkamp

577   HME., Vos R a., Guerreiro I., Calvete JJ., Wüster W., Woods AE., Logan JM., Harrison R

578   a., Castoe T a., de Koning a PJ., Pollock DD., Yandell M., Calderon D., Renjifo C., Currier

579   RB., Salgado D., Pla D., Sanz L., Hyder AS., Ribeiro JMC., Arntzen JW., van den Thillart

580   GEEJM., Boetzer M., Pirovano W., Dirks RP., Spaink HP., Duboule D., McGlinn E., Kini

581   RM., Richardson MK. 2013. The king cobra genome reveals dynamic gene evolution and

582   adaptation in the snake venom system. *Proceedings of the National Academy of Sciences of*

583   *the United States of America* 110:20651–20656. DOI: 10.1073/pnas.1314702110.

584   Weston AJ., Dunlap WC., Shick JM., Klueter A., Iglic K., Vukelic A., Starcevic A., Ward M.,

585   Wells ML., Trick CG., Long PF. 2012. A profile of an endosymbiont-enriched fraction of

586   the coral *Stylophora pistillata* reveals proteins relevant to microbial-host Interactions.

587   *Molecular & Cellular Proteomics* 11:M111.015487–M111.015487. DOI:

588   10.1074/mcp.M111.015487.

589   Weston AJ., Chung R., Dunlap WC., Morandini AC., Marques AC., Moura-da-Silva AM., Ward

590   M., Padilla G., da Silva LF., Andreakis N., Long PF. 2013. Proteomic characterisation of

591   toxins isolated from nematocysts of the South Atlantic jellyfish *Olindias sambaquiensis*.

592   *Toxicon* 71:11–17. DOI: 10.1016/j.toxicon.2013.05.002.

593   Wheeler DL., Church DM., Federhen S., Lash AE., Madden TL., Pontius JU., Schuler GD.,

594   Schrimi LM., Sequerira E., Tatusova TA., Wagner L. 2003. Database resources of the

595   National Center for Biotechnology. *Nucleic Acids Research* 31:28–33. DOI:

596   10.1093/nar/gkg033.

597   Whittington CM., Papenfuss AT., Locke DP., Mardis ER., Wilson RK., Abubucker S., Mitreva

598   M., Wong ESW., Hsu AL., Kuchel PW., Belov K., Warren WC. 2010. Novel venom gene

599   discovery in the platypus. *Genome Biology* 11:R95. DOI: 10.1186/gb-2010-11-9-r95.

600   Wong ESW., Hardy MC., Wood D., Bailey T., King GF. 2013. SVM-based prediction of

601   propeptide cleavage sites in spider toxins identifies toxin innovation in an *Australian*

602   *tarantula*. *PloS One* 8:e66279. DOI: 10.1371/journal.pone.0066279.

603   Zdobnov EM., Apweiler R. 2001. InterProScan-an integration platform for the signature-

604   recognition methods in InterPro. *Bioinformatics (Oxford, England)* 17:847–848. DOI:

605   10.1093/bioinformatics/17.9.847.

606

607   **Acknowledgments**

608   The authors are grateful to Prof. Dr. Ana M. Moura-da-Silva who provided the unpublished

609   anonymised snake venom transcriptome sequences. We also thank Prof. J. Malcolm Shick and

610   Prof Dr John Cullum for critically reviewing the manuscript.

611

612   **Authors' Contributions**

613   PFL conceived the study. RG carried out the bioinformatics analysis. All authors interpreted the

614   data, wrote the manuscript and have approved the final version of the manuscript.

615

616   **Competing financial interests**

617   The authors declare that they have no competing interests.

618

619   **Figure 1 legend:** Figure displays calibration curves used to select final prediction scores for

620   ToxClassifier ensemble. Each of performance measures, described in detail in Methods, is shown

621   for ToxClassifier prediction with toxin prediction cut-off values 1 to 9. Dotted green line shows

622   trends of prediction measurement with increase in ToxClassifier cut-off value and final cut-off

623   value implemented in calibrated ToxClassifier is highlighted with blue line.

624

625   **Figure 2 legend:** Figure shows comparison between selected toxin prediction tools for all animal

626   test datasets. Calibrated ToxClassifier, with positive prediction cutoff 5, is shown by green bar,

627   ClanTox prediction server is displayed in orange, oneBLAST prediction performance in red and

628   naiveBLAST in blue. F1-value and Matthews' correlation coefficient are displayed for each of

629   test sets and for summary of all data with exclusion of conus snail proteins.

630

631   **Table 1 legend:** Table displays prediction accuracy on positive and negative datasets, as well as

632   range of measurements calculated for all test data, and described in detail in Methods.

633   Annotation models used as classifier inputs either: the frequency of amino acids (TBSim) or

634   combinations of two amino-acids (BIF); the presence of absence or 'Tox-Bits' (SToxA); HMM

635   scores for 'ToxBits' (SToxB); a selection of BLAST output co-variants (TBEa); a variation on

636   TBSim and TBEa (TBEb). Classifier Learning Machines used were: Gradient Boosted (GBM),

637   Support Vector (SVM) and Generalised Linear Model (GLM). The datasets were a 'Positive'

638   control containing only validated animal toxins, an 'Easy' dataset composed of non-toxin

639   sequences, a 'Moderate' dataset comprising curated non-toxin sequences but with homology to

640   'Positive' sequences, and a 'Hard' dataset that included all sequences from the 'Moderate'

641   dataset, together with un-curated sequences also with homology to 'Positive' sequences.

642

643   **Table 2 legend:** Table compares performance for selected annotation models and published

644   toxin prediction tools. Prediction accuracy is listed for positive and negative datasets and

645   measurements are also shown for summary of all test data. ToxinPred tool, marked with star and

646   displayed in italic, was tested with short protein sequences only. Annotation models were

647   constructed from sequence information extracted from either: BLAST (naiveBLAST), 'tox-bits'

648   HMM   (hmmerToxBits)   or   Pfam   HMM   (hmmerVenom)   comparisons   with   the

649   UniProtKB/SwissProt-ToxProt   database;   or   BLAST   (oneBLAST),   'tox-bits'   HMM

650   (twinHmmerPfam)   or   Pfam   HMM   (twinHmmerPfam)   comparisons   with   the

651   UniProtKB/SwissProt-ToxProt sequences supplemented with additional toxin-like sequences

652   from the UniProtKB/SwissProt and TrEMBL databases; or BLAST (triBLAST) comparisons

653   with the UniProtKB/SwissProt-ToxProt sequences supplemented with non-toxin sequences from

654   the 'Moderate' and 'Hard' datasets used to train the machine classifiers. The datasets were a

655   'Positive' control containing only validated animal toxins, an 'Easy' dataset composed of non-

656   toxin sequences, a 'Moderate' dataset comprising curated non-toxin sequences but with

657   homology to 'Positive' sequences, and a 'Hard' dataset that included all sequences from the

658   'Moderate' dataset, together with un-curated sequences also with homology to 'Positive'

659   sequences.

660

661   **Table 3 legend:** Table lists comparison of classification performance for calibrated

662   ToxClassifier to BLAST based annotation models and ClanTox toxin prediction server. All tests

663   were conducted test dataset not used in calibration of ToxClassifier.

664

665   **Table 4 legend:** Table compares results of expert manual annotation to ToxClassifier annotation

666   for set of novel proteins from venom gland transcriptome of *Bothrops atrox* snake.

Figure 1

Calibration curves used to select final prediction scores for ToxClassifier ensemble

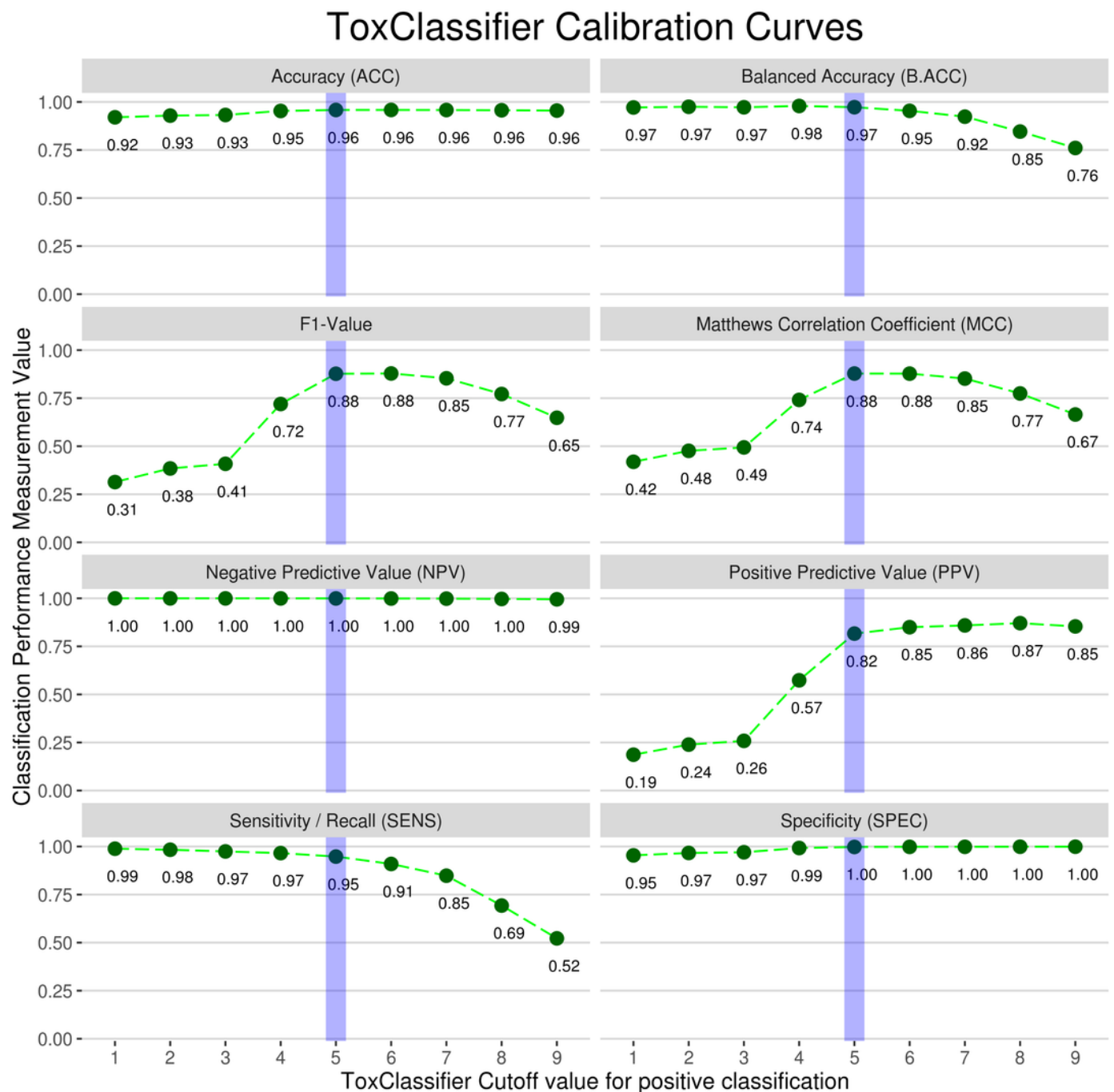
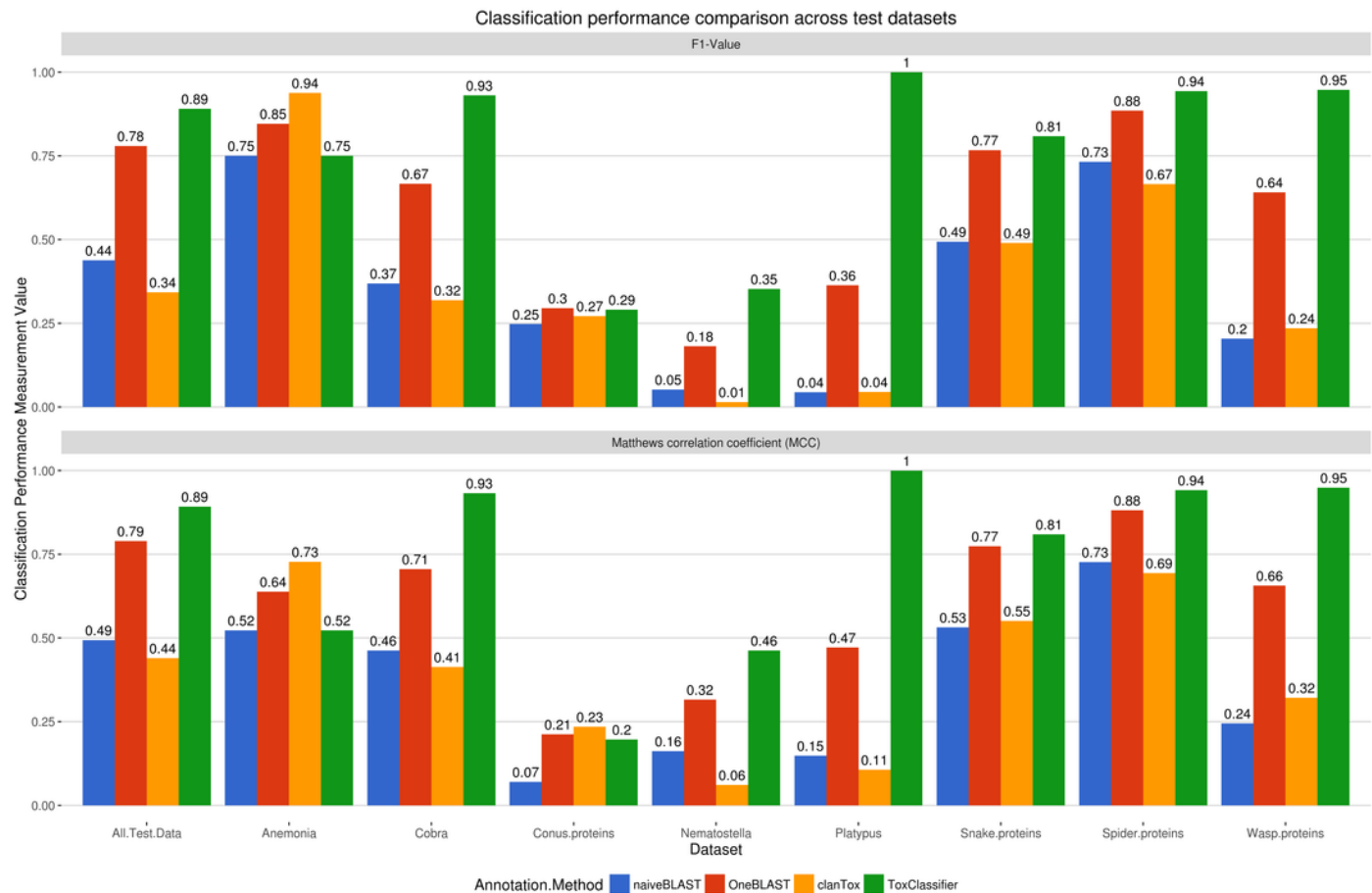



Figure 2

Comparison between selected toxin prediction tools for all animal test datasets.

Prediction accuracy on positive and negative datasets, as well as range of measurements calculated for all test data

1   Table 1: Comparative accuracy values for the classification of protein sequences as toxins or non-toxins using 6 different annotation
2   models trained on 3 separate classifier learning machines.
3

		Classification scores for:				Test Set Summary					
Annotation Model	Classifier	Accuracy (Positive toxin dataset)	Accuracy (Easy non-toxin dataset)	Accuracy (Moderate non-toxin dataset)	Accuracy (Hard non-toxin dataset)	PPV	NPV	Sens.	Spec.	F1-value	MCC
TBSim	GBM	0.80	0.99	0.98	0.92	0.80	0.98	0.84	0.97	0.82	0.80
	SVM	0.80	1.00	0.98	0.94	0.80	0.99	0.91	0.97	0.85	0.84
	GLM	0.55	0.99	0.96	0.84	0.55	0.97	0.69	0.94	0.61	0.57
BIF	GBM	0.83	1.00	0.98	0.94	0.83	0.99	0.92	0.98	0.87	0.86
	SVM	0.89	1.00	0.98	0.96	0.89	0.99	0.94	0.99	0.91	0.90
	GLM	0.71	0.99	0.99	0.91	0.71	0.98	0.82	0.96	0.76	0.74
SToxA	GVM	0.64	1.00	0.98	0.94	0.64	0.99	0.90	0.96	0.75	0.73
	SVM	0.84	1.00	0.96	0.91	0.84	0.98	0.87	0.98	0.86	0.84
SToxB	GBM	0.75	1.00	1.00	0.93	0.75	0.99	0.92	0.97	0.83	0.84
	SVM	0.85	1.00	0.99	0.92	0.85	0.99	0.91	0.98	0.88	0.81
	GLM	0.03	1.00	0.99	0.99	0.03	1.00	0.61	0.89	0.06	0.12
TBEa	GBM	0.88	1.00	1.00	0.99	0.88	1.00	0.99	0.98	0.93	0.93
	SVM	0.93	1.00	1.00	0.97	0.93	1.00	0.97	0.99	0.95	0.94
	GLM	0.96	1.00	1.00	0.94	0.96	0.99	0.95	0.99	0.95	0.95
TBEb	GBM	0.82	1.00	1.00	1.00	0.82	1.00	1.00	0.98	0.90	0.90
	SVM	0.96	1.00	1.00	0.97	0.96	1.00	0.97	0.99	0.97	0.96
	GLM	0.93	1.00	1.00	0.99	0.93	1.00	0.99	0.99	0.96	0.95

Performance for selected annotation models and published toxin prediction tools.

1 Table 2: Comparative balanced accuracy values for classification of protein sequences as toxins or non-toxins using different
2 annotation methodologies based on BLAST or HMMER alignments.

		Classification scores for:				Test Set Summary statistics					
Annotation Model	Tool	Accuracy (Positive toxin dataset)	Accuracy (Easy non-toxin dataset)	Accuracy (Moderate non-toxin dataset)	Accuracy (Hard non-toxin dataset)	PPV	NPV	Sens.	Spec.	F1-value	MCC
naiveBLAST	BLAST	0.90	0.99	0.83	0.07	0.90	0.86	0.46	0.99	0.60	0.58
oneBLAST	BLAST	0.86	1.00	0.98	0.90	0.86	0.99	0.89	0.98	0.87	0.86
triBLAST	BLAST	0.87	0.99	0.93	0.87	0.87	0.97	0.78	0.98	0.82	0.80
hmmerToxBits	HMMER	0.91	0.99	0.80	0.19	0.91	0.87	0.48	0.99	0.63	0.60
hmmerVenom	HMMER	0.65	0.95	0.74	0.23	0.65	0.84	0.34	0.95	0.45	0.38
twinHmmerPfam	HMMER	0.56	0.99	0.98	0.91	0.56	0.98	0.78	0.95	0.65	0.62
twinHmmerToxBits	HMMER	0.85	1.00	0.93	0.44	0.85	0.92	0.59	0.98	0.70	0.67
ClanTox server	ML	0.66	0.99	0.93	0.73	0.66	0.95	0.65	0.96	0.66	0.61
*ToxinPred server**	*ML*	*0.55*	*0.98*	*N/A*	*N/A*	*0.55*	*0.98*	*0.82*	*0.93*	*0.66*	*0.63*

Comparison of classification performance for calibrated ToxClassifier to BLAST based annotation models and ClanTox toxin prediction server

1   Table 3: ToxClassifier calibrated meta-classifier, Test set results compared to oneBLAST, naiveBLAST and ClanTox annotation
2   methods

Classification performance of ToxClassifier meta-classifier, compared to BLAST based methods and ClanTox server								
Annotation Model	**Tool**	**Accuracy**	**Positive Prediction Value**	**Negative Prediction Value**	**Sensitivity**	**Specificity**	**F1-value**	**MCC value**
naiveBLAST	BLAST	0.976	0.294	0.998	0.857	0.977	0.438	0.494
oneBLAST	BLAST	0.994	0.660	1.000	0.951	0.995	0.779	0.790
ClanTox server	ML-based	0.947	0.209	1.000	0.954	0.976	0.342	0.440
ToxClassifier	ML-based	0.997	0.825	1.000	0.967	0.998	0.890	0.892

Results of expert manual annotation to ToxClassifier annotation for set of novel proteins from venom gland transcriptome of *Bothrops atrox* snake.

1 Table 4: Blinded 'ToxClassifier' prediction of proteins as toxins or non-toxins from the standard
2 annotation of the venom gland transcriptome of *Bothrops atrox*.

	Sequence annotation	
Test Sample	**'ToxClassifier'**	**Standard annotation**
Test 01	Potential toxin	BPP precursor - Toxin
Test 02	Potential toxin	Crisp - Toxin
Test 03	Toxin	CTL - Toxin
Test 04	Toxin	CTL - Toxin
Test 05	Potential toxin	Hyaluronidase - Toxin
Test 06	Potential toxin	Kunitz –like sequence - Probably a toxin
Test 07	Toxin	LAAO - Toxin
Test 08	Not a toxin	Nucleotidase – Probably not a toxin
Test 09	Not a toxin	Phosphodiesterase – Probably not a toxin
Test 10	Toxin	PLA2 - Toxin
Test 11	Toxin	SVMP class PI - Toxin
Test 12	Toxin	SVMP class PII - Toxin
Test 13	Toxin	SVMP class PIII - Toxin
Test 14	Toxin	SVSP - Toxin
Test 15	Potential toxin	VEGF - Toxin