Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and user confirmation

Bioinformatics and computer aided drug design rely on the curation of a large number of protocols for biological assays that measure the ability of potential drugs to achieve a therapeutic effect. These assay protocols are generally published by scientists in the form of plain text, which needs to be more precisely annotated in order to be useful to software methods. We have developed a pragmatic approach to describing assays according to the semantic definitions of the BioAssay Ontology (BAO) project, using a hybrid of machine learning based on natural language processing, and a simplified user interface designed to help scientists curate their data with minimum effort. We have carried out this work based on the premise that pure machine learning is insufficiently accurate, and that expecting scientists to find the time to annotate their protocols manually is unrealistic. By combining these approaches, we have created an effective prototype for which annotation of bioassay text within the domain of the training set can be accomplished very quickly. Well-trained annotations require single-click user approval, while annotations from outside the training set domain can be identified using the search feature of a well-designed user interface, and subsequently used to improve the underlying models. By drastically reducing the time required for scientists to annotate their assays, we can realistically advocate for semantic annotation to become a standard part of the publication process. Once even a small proportion of the public body of bioassay data is marked up, bioinformatics researchers can begin to construct sophisticated and useful searching and analysis algorithms that will provide a diverse and powerful set of tools for drug discovery researchers.

1 Title

2 Fast and accurate semantic annotation of bioassays exploiting a hybrid of machine learning and

3 user confirmation

4 Authors

5 Corresponding author:

6   Alex M. Clark

7   1633 Bayshore Hwy, Suite 342

8   Burlingame, CA, USA

9   alex@collaborativedrug.com

10   (650) 242-5259

11 Alex M. Clark,[a] Barry A. Bunin,[a] Nadia K. Litterman,[a] Stephan C. Schürer,[b] Ubbo Visser [b]

12 (a)   Collaborative Drug Discovery, Inc.

13   1633 Bayshore Hwy, Suite 342

14   Burlingame, CA, USA

15 (b)   Center for Computational Science

16   University of Miami

17   Miami, FL, USA

18 Abstract

19 Bioinformatics and computer aided drug design rely on the curation of a large number of

20 protocols for biological assays that measure the ability of potential drugs to achieve a therapeutic

21 effect. These assay protocols are generally published by scientists in the form of plain text, which

22 needs to be more precisely annotated in order to be useful to software methods. We have

23 developed a pragmatic approach to describing assays according to the semantic definitions of

24 the BioAssay Ontology (BAO) project, using a hybrid of machine learning based on natural

25 language processing, and a simplified user interface designed to help scientists curate their data

26 with minimum effort. We have carried out this work based on the premise that pure machine

27 learning is insufficiently accurate, and that expecting scientists to find the time to annotate their

28 protocols manually is unrealistic. By combining these approaches, we have created an effective

29 prototype for which annotation of bioassay text within the domain of the training set can be

30 accomplished very quickly. Annotations that are well-defined in the training data are generally

31 predicted correctly by the machine learning models, and require only single-click user approval,

32 while annotations from outside the training set domain can be identified using the search feature

33 of a well-designed user interface, and subsequently used to improve the underlying models. By

34 drastically reducing the time required for scientists to annotate their assays, we can realistically

35 advocate for semantic annotation to become a standard part of the publication process. Once

36 even a small proportion of the public body of bioassay data is marked up, bioinformatics

37 researchers can begin to construct sophisticated and useful searching and analysis algorithms

38 that will provide a diverse and powerful set of tools for drug discovery researchers.

Introduction

40 In recent decades scientific data has been almost entirely digitized: authors prepare their

41 manuscripts and presentations using a collection of text, graphics and data processing software.

42 Consumers of scientific data regularly download documents from publishers' websites, search

43 for content in databases, and share data with their colleagues electronically, often in an entirely

44 paperless fashion. Dozens of commercial and academic research groups are actively working on

45 ways to use software to analyze this rapidly expanding corpus of data to provide facile

46 information retrieval, and to build decision support systems to ensure that new research makes

47 the best possible use of all available prior art.

48 Despite the near complete migration from paper to computers (Khabsa & Giles 2014), the style

49 in which scientists express their results has barely changed since the dawn of scientific

50 publishing. Whenever possible, ideas and facts are expressed as terse paragraphs of English

51 text, kept as short as possible to minimize printing costs, and as stripped down diagrams that

52 often summarize vast numbers of individual data points in a form that can be visualized statically

53 by a scientific peer. These methods of communication have remained consistent because they

54 are effective for their primary purpose, but this presents a major hurdle to computer software that

55 is attempting to perform data mining operations on published results.

56 In the case of biological assays, experiments designed to measure the effects of introduced

57 substances for a model of a biological system or disease process, the protocols are typically

58 described in one or more textual paragraphs. Information about the target biology, the proteins or

59 cells, the measurement system, the preparation process, etc., are all described using information

60 rich jargon that allows other scientists to understand the conditions and the purpose. This

61 comprehension process is, however, expert-specific and quite time consuming. While one

62 scientist may read and understand dozens of published assay descriptions, this is not scalable

63 for large-scale analysis, e.g. clustering into groups after generating pairwise metrics, or

64 searching databases for related assays.

65 One of the most promising approaches to solving this problem is to express the assay design

66 experiments with terminology from a semantically rich ontology, which has the advantage of

67 being readily understood by software (Jonquet et al. 2010; Jonquet et al. 2009; Roeder et al.

68 2010). Efforts such as the BioAssay Ontology (BAO) project (Abeyruwan 2014; Vempati et al.

69 2012; Visser et al. 2011) were specifically designed to address this issue, and is part of a

70 pantheon of ontologies for expressing the chemistry and biology definitions and relationships

71 that are essential to drug discovery. Having all relevant scientific data expressed in semantic

72 form enables an inordinate number of options for building compelling decision support software,

73 but the biggest hurdle is the expression of the data. Expecting scientists to alter their

74 documentation habits to use computer-friendly ontologies rather than human-friendly natural

75 language is unrealistic, especially given that the benefits do not start to accrue until a critical

76 mass is achieved within the community. On the other hand, there has been a considerable

77 amount of research toward designing software to perform fully automated parsing of otherwise

78 intractable text and diagrams (Attwood et al. 2010), and add annotations in a machine-friendly

79   format. Many of these efforts have been found to be valuable for certain scenarios where the

80   high error rate is tolerable. For example, allowing a scientist to search the entire patent literature

81   for chemical reactions may be a very useful service even with a low signal to noise ratio,

82   because the effort required to manually filter out false positives is relatively low, and the portion

83   of false negatives may be no worse than more traditional methods (Hawizy et al. 2011; Jessop et

84   al. 2011a; Jessop et al. 2011b).

85   Nonetheless, such fully automated extraction procedures are likely to continue to have a very

86   high error rate for most scientific subject areas for the conceivably near future, and the poor

87   signal to noise ratio prevents most kinds of analysis from being effective. To address this urgent

88   issue, we have developed methods for combining automated extraction and manual curation in

89   order to optimize for both goals: minimal additional burden on practicing scientists, and minimal

90   transcription errors during the semantic markup process.

91   There are already examples of hybrid manual/automatic annotation technologies, for example

92   PubTator,(Wei et al. 2013) which is designed to help identify a variety of keywords in order to

93   classify papers within the PubMed collection. The web interface provides an initial attempt to

94   identify keywords that correspond to semantic content in the chemical or biological domain, and

95   allows the user to confirm them or add their own. On the other hand, some of the large scale

96   curation efforts, such as ChEMBL, provide funding for expert curators to manually annotate

97   bioassay data, but this is too labor intensive to execute in detail, and is currently limited to

98   identifying the target (Gaulton et al. 2012). Approaches such as Active Learning (AL) have also

99   been applied to classification of domain specific text documents (Cohn 1996; Dara 2014;

100 Tomanek 2007). Our objective in this work is to provide the necessary capabilities to annotate

101 bioassay protocols, in a significant level of detail, such that the semantic content is a relatively

102 complete description of the assay. In many ways our aims are similar to other natural language

103 text-based classification projects, but unusual in that we are ultimately seeking to use these

104 methods to express a very detailed description of a domain localized class of experiments.. We

105 can draw upon existing vocabularies, such as the BioAssay Ontology (BAO), and other

106 ontologies, which it in turn references, for the means to complete this description. To achieve the

107 objective of reducing the burden on the individual scientist to the bare minimum, we have made

108 use of natural language processing and machine learning techniques, and coupled the

109 algorithms to a prototype user interface with a workflow design that iterates back and forth

110 between automated inference and operator approval.

111 Rather than starting with the lofty objective of having an algorithm provide the right answers all of

112 the time, we merely require it to eliminate most of the wrong answers. To the extent that we are

113 able to achieve this comparatively realistic goal, this allows us to create a user-facing service for

114 which the scientist simply selects correct semantic markup options from a short list of options

115 proposed by the software. This is as opposed to the entirely manual curation approach, which

116 would require the operator to navigate through a densely packed hierarchy of descriptors. By

117 reducing the burden of markup to mere minutes by somebody who is not familiar with semantic

118 technology, and has had no special training for use of the software, it is quite reasonable to

119 expect scientists to use this software as part of their standard write-up and submission process.

120 As the number of correctly annotated bioassay protocols grows, it will improve the training set

121 and the machine learning algorithm will correspondingly improve in accuracy. Once the currently

122 high barrier to adoption has been overcome, and semantic markup of scientific knowledge such

123 as biological assay experiments is routine, assay protocols will be as readable to computer

124 software as they are to expert scientists. The informatics capabilities that this will unlock are

125 essentially limitless, but the first most clear example is the ability to search assays for specific

126 properties, e.g. target, assay type, cell line, experimental conditions, etc. Being able to

127 conveniently organize and aggregate assays by particular characteristics, cluster by similarity, or

128 assemble chemical structures and activity from multiple assays based on customizable criteria,

129 are all advantages that have a direct impact on drug discovery, which are currently held back by

130 the lack of semantic annotation. Once the corpus of marked up annotations becomes large, it

131 will also be possible to construct data mining algorithms to study large scale trends in bioassay

132 data, which will result in entirely new kinds of insight that are currently not possible.

Methods

Ontologies

The primary annotation reference for this project is the BioAssay Ontology (BAO), which is available from http://bioassayontology.org, and can be downloaded in raw RDF format. The BAO classes refer to a number of other ontologies, and of particular relevance are the Cell Line Ontology (CLO) (Sarntivijai 2011), Gene Ontology (GO) (Balakrishnan et al. 2013; Blake 2013), and NCBI Taxonomy (Federhen 2012), all of which are used for annotations within the training set. All of the source files for these ontologies were loaded into a SPARQL server (Apache Fuseki)(Website 2014a). SPARQL queries were used to organize the available values that correspond to each of the property groups.

Training data

In order to test the methodology of using text to create suggested annotations, we made use of a corpus of annotated bioassays that were that were provided by the BAO group (Schurer et al. 2011; Vempati et al. 2012) (See supplementary material). As part of the testing process for the BioAssay Ontology project, a simple annotation user interface was created in the form of an Excel spreadsheet template. Approximately 1000 assays were selected from PubChem, and each of these was painstakingly annotated, leading to an output document taking the form of: ‹assay ID› ‹property› ‹value›.

For each assay, 20-30 annotations were incorporated into the training set. The property values were individually mapped to the BAO space, e.g. 'has assay method' is mapped to the URI http://www.bioassayontology.org/bao#BAO_0000212, which is a part of the BAO ontology. Values that are string literals are not included in the training data. Those which map to a distinct URI are typically part of the BioAssay Ontology directly, or part of other ontologies that are referenced, such as the Cell Line Ontology (CLO), Gene Ontology (GO) and NCBI Taxonomy.

Once the annotations had been suitably collated for each distinct assay, the NCBI PubChem assays were obtained by a simple script making a call to the PUG RESTful API (Website 2014b). In each case, the *description* and *protocol* sections of the resulting content were merged into a

160 free text document. The manner in which these two fields are used by scientists submitting new

161 assay data varies considerably, but they are generally complete. For the collection of text

162 documents obtained, it was necessary to manually examine each entry, and remove non-

163 pertinent information, such as attribution, references and introductory text. The residual text for

164 each case was a description of the assay, including information about the target objective, the

165 experimental details, and the materials used. The volume of text varies from concisely worded

166 single paragraph summaries to verbosely detailed page length accounts of experimental

167 methodology. These reductively curated training documents can be found in the supplementary

168 information.

169 ### *Natural language processing*

170 There has been a considerable amount of effort in the fields of computer science and linguistics

171 to develop ways to classify written English documents in terms of classified tokens that can be

172 partially understood by computer software (Kang & Kayaalp 2013; Leaman et al. 2013; Liu et al.

173 2011; Santorini 1990). We made use of the OpenNLP project(Website 2014c), which provides

174 *part of speech* (POS) tagging capabilities, using the default dictionaries that have been trained

175 on general purpose English text. The POS tags represent each individual word as a token that is

176 further annotated by its type, e.g. the words "report" and "PubChem" were classified as an

177 ordinary noun and a proper noun, respectively:

178   (NN report)

179   (NNP PubChem)

180 Blocks of text are classified in an increasingly specific hierarchical form, e.g.

181   (NP (DT an) (JJ anti-cancer) (NN drug))

182   (VP (VBG developing) (NP (JJ potential) (JJ human) (NNS therapeutics)))

183   (NP (NP (NN incubation)) (PP (IN with) (NP (NN test) (NN compound))))

184   (NP (NP (DT the) (JJ metabolic) (NN activity)) (PP (IN of) (NP (DT a) (NN suspension)

185 (NN cell) (NN line))))

186   (VP (VB measure) (SBAR (WHADVP (WRB when))) (S (VP (VBG developing) (NP (JJ

187   potential) (JJ human) (NNS therapeutics))))))

188   (NP (JJ luciferase-based) (NN cell) (NN proliferation/viability) (NN assay) (NN endpoint))

189 An assay description of several paragraphs can generate many hundred distinct instances of

190 POS-tagged blocks. These marked up tokens contain a far larger amount of information about

191 the composition of the sentence than the words themselves. While building a model by

192 correlating words with annotations would be expected to achieve poor results, including markup

193 information about how the words are used in conjunction with each other might be able to

194 achieve far greater discrimination. For example, the POS-tagged block "(NP (DT an) (JJ anti-

195 cancer) (NN drug))" represents the words [an, anti, cancer, drug]. Each of these 4 words taken

196 out of context could be found in almost any assay description, but when they are associated

197 together in context, contribute an important statement about the corresponding biochemistry.

198 By collecting all sizes of POS-tagged blocks, up to a certain limit, it is possible to give many

199 different depths of linguistic structure the opportunity to distinguish themselves within a model. In

200 some cases a single word can have significant meaning on its own, especially proper nouns or

201 jargon (e.g. "luciferase"), and are likely to have a high correlation to certain kinds of annotations

202 (e.g. use of a luciferase-based assay). Other words are general to the English language, or

203 occur frequently in assay descriptions, such that they only have value in their proper context (e.g.

204 "interaction").

205 One of the useful properties of scientific writing is that authors have self-organized around a

206 narrow range of styles for presenting information such as assay descriptions. While the explicit

207 intent may not have been for the benefit of computerized natural language processing, the

208 motivation is the same: scientific authors also read many other published descriptions, and it is

209 in the best interests of the community to maintain a certain degree of consistency as well as

210 brevity. Because the literary style lacks prose and has a relatively little variation, there are certain

211 blocks of words, as identified by the POS-tagging, that are frequently correlated with particular

212 concepts, and hence the semantic annotations.

Machine learning models

213   **Machine learning models**

214   A collection of hundreds of assay descriptions will result in thousands of distinct POS-tagged

215   blocks after processing each of them with natural language analysis, and while certain blocks

216   are likely to be specifically correlated with certain annotations, there are many more with only

217   weak correlation or none at all. Matching thousands of potentially important tags with hundreds

218   or thousands of annotations requires the selection of an algorithm with favorable scaling

219   properties, and is well beyond the scope of manual curation.

220   In our initial explorations, we chose to apply a variation of Bayesian inference, which has been

221   used successfully in other aspects of computer aided drug discovery. The Laplacian-modified

222   naïve Bayesian variant is frequently used in conjunction with chemical structure based

223   fingerprints (Hassan et al. 2006; Mussa et al. 2013; Nidhi et al. 2006; Rogers et al. 2005), as it is

224   highly tolerant of large numbers of parameters. The score for each annotation is calculated as:

$$score = \sum_{n} \ln\left[\frac{A_n + 1}{T_n \cdot P + 1}\right]$$

225   where n is the tagged natural language block, A_n is the number of documents containing the

226   annotation and the tagged block, T_n is the total number of documents with the tagged block, and

227   P is the fraction of documents containing the annotation. The score is computed by adding up

228   the logarithms of these ratios, which circumvents issues with numeric precision, but produces a

229   score with arbitrary scale, rather than a probability.

230   When we considered each individual annotation as a separate observation, building a Bayesian

231   model using the presence or absence of each distinct POS-tagged block gave rise to a highly

232   favorable response for most annotations, as determined by the receiver-operator-characteristic

233   (ROC) curves. Selected examples of these models are shown in Figure 1: (a) shows annotations

234   with high training set coverage that perform well, due in part to having relatively unambiguous

235   word associations, while (b) shows well covered annotations that perform poorly, due to being

236   reliant on terms that can be used in a variety of contexts that do not necessarily imply the

237   presence of the annotation, and hence make it more difficult for the model to eliminate false

238 positives. Similarly, (c) shows the perfect recall for less well covered annotations, which are

239 easily identified due to very specific terms, while (d) shows a relatively poor response due to

240 small training set and terminology with variations in wording style.

241 One of the disadvantages of using this Laplacian corrected variant is that the computed value is

242 not a probability, but rather a score with arbitrary range and scale. This means that it is not

243 possible to compare the outcomes from two separate models, which is a problem, since the

244 objective of this technology is to *rank* the scores that are obtained from each separate model. In

245 order to achieve the ranking, the scores need to be directly comparable, and hence be suitable

246 for providing a list of suggestions for which annotations are most likely to be associated with the

247 text.

248 In order to make the scores from each of the models comparable, each model requires a

249 calibration function. This can be accomplished effectively by defining a simple linear correction

250 for each model, of the form $y = ax + b$, which is applied to each score prior to inter-model

251 comparison. Selecting appropriate values for a and b, for each model, can be achieved by

252 picking initial values that map each of the model outcomes to the range 0..1. By adjusting the

253 scale and offset of the linear calibration functions for each of these models, the overall ability of

254 the models to correctly rank the extant annotations with a higher score than those which are not

255 observed can be evaluated. It is straightforward to define a scoring term that measures the

256 ability of the calibrated models to distinguish between correct and incorrect annotations. This

257 score can be optimized by iteratively adjusting the calibration terms to get the best overall

258 separation in ranking.

259 Besides consistent use of linguistic descriptions of assays, one of the other observations about

260 the annotations defined for these assay protocols is that they are not in any way orthogonal: the

261 degree to which the annotations are correlated is very high. For example, if it is known that the

262 assay uses *luciferin* as a substrate, the probability that it also involves *luminescence* as a

263 detection method is far higher than it would be if the prior fact had not already been established.

264 Given that the calibrated Bayesian models have been established to perform very well at placing

265 the top few highest ranking annotations for the data used in this study, once these top scoring

266 annotations have been confirmed by the user, the amount of information that can be inferred

267 about the assay may be significantly greater, due to the high degree of correlation.

268 This second order correlation was implemented by building another set of Bayesian models with

269 each possible annotation considered separately as an observation. For each document, each

270 annotation's likely presence is modeled against the presence or absence of all the other

271 annotations recorded for the document, e.g. when building the correlation model for annotation

272 A, if document i contains annotations A, B and C, then it is considered to be "active", with priors

273 B and C; if document j contains annotations B, C and D, it is considered "inactive", with priors B,

274 C and D.

275 Thus, once one or more annotations have been assigned, the secondary Bayesian models are

276 consulted, and the score for each of the annotations is modified by applying the correlation

277 factor. Essentially this means that as the user approves annotations, the prediction scores of the

278 remaining annotations tends to improve, as the correlations are factored in.

279 Figure 2 provides an indication of how the ranking evolves during the model building steps, using

280 four example documents. For each of these diagrams, the left hand side shows two bands which

281 represent the *uncalibrated* predictions, which are linearly normalized so their values fall between

282 the minimum and maximum scores from the raw Bayesian prediction score. The annotations that

283 do not apply to the document are shown as red lines, while the annotations that are present are

284 shown in black. The height of each line is indicative of its score. As can be clearly seen, the

285 desired predictions score are significantly higher for those present than those which are absent,

286 but the extent to which the ranking separates the two groups varies, and is not initially a perfect

287 separation for any of these examples.

288 The main area of each diagram shows the progression of the relative predictions: at the

289 beginning of the sequence, the scores are ranked by the inter-model *calibration* functions, which

290 typically results in a significant improvement. For each of the subsequent steps, the highest

291 scoring correct annotation is added to the approved set, and the *correlation* model is updated

292 and applied. The ranking is redetermined, and the next highest scoring correct annotation is

293 selected. The diagram indicates the point at which each annotation is selected by plotting a

294   black circle, and changing the color of the line to green: since it has been confirmed, its ranking

295   order is no longer a concern, though its presence continues to affect the way the correlation

296   model is applied.

297   In the first example, shown in Figure 2 (a), application of these models in the given sequence

298   provides a perfect result: in each case the highest scoring annotation yet to be selected is at the

299   top of the list, with no false positives. In examples (b) and (c), the results are good but not

300   perfect: the red cross marks indicate when an incorrect annotation was presented as the best

301   next choice. Since this exercise is simulating curation by a human expert, the elimination of a

302   top-ranked incorrect proposal is equivalent to being recognized by the user as an incorrect

303   result, and explicitly excluded from further consideration. If the objective was to provide a pure

304   machine learning solution, each of these ranking mistakes would represent the accumulation of

305   bad data, rather than a small increase in the amount of effort required by the operator. In

306   example (d), the response of the model is relatively poor, with several false positives appearing

307   close to the top of the list, and the last few correct results being obscured by a large number of

308   incorrectly ranked proposals.

309   **Results**

310   We have designed the algorithm with the goal of ranking the available annotations such that

311   given a text description of an assay, the annotations that correctly apply to the document are

312   listed before any which do not. A perfect result is considered to be any case where all of the

313   correct proposals are ranked first. Because the objective of the machine learning is to assist and

314   accelerate the human-guided curation, a handful of mis-ordered annotations can be considered

315   as an inconvenience, rather than the means by which data becomes corrupted.

316   For evaluation purposes, we define a yardstick measure: the null hypothesis is that the

317   Bayesian-trained model using natural language processing performs no better than a trivial

318   method, such as ranking all proposed annotations by the frequency with which they occur in the

319   training set.

Cross validation

The 983 fully annotated assays, with corresponding text from PubChem, were split into training

and test sets using a simple partitioning algorithm. First of all, 208 documents were removed on

account of having the same list of property:value annotations. These documents may differ by

the free text annotations, but these are not included in the training set, and so duplicates need to

be pruned. Of the remaining documents, entries were selectively picked for the test set in order

to ensure that each annotation appears once in any one of the test set documents, but such that

the number of instances remaining in the training set was not reduced below 2. The training set

contained 698 documents, the test set 77.

The models were rebuilt using just the training set documents, and applied to the test set. For

evaluation purposes, we can consider the ranking of correct vs. incorrect answers to be

instructive for how well the model achieves its stated goal. Figure 3 shows several plots that

show the relative performance of the training and test sets.

The data for each plot is created according to the following procedure:

1. score each available annotation based on the model derived from the training set data;

2. pick the highest scoring annotation: if it is correct, add a positive mark, remove the
   annotation, and goto 1;

3. it is not correct, so add a negative mark, remove the annotation, and goto 2.

This essentially simulates an expert operator who only looks at the current top scoring

annotation, and either approves it, or excludes it from further consideration. The process stops

when all correct annotations have been approved.

Figure 3 illustrates this process graphically from several vantagepoints. In 3 (a), all of the test set

documents are considered: for each line, running from left to right, a correct top ranking

annotation is marked with a black square, while an incorrect top ranking annotation is marked

with a purple square. Once all of the correct annotations have been picked, the remaining space

is marked in grey. As can be seen, for the majority of cases the correct annotations are quickly

picked out. Nonetheless there are a number of test documents that contain a small number of

347   outliers, i.e. required annotations that are ranked very poorly, with many false positives getting a

348   higher score.

349   Figure 3 (b) shows the same datapoints, except that only the actual annotations are given a

350   color. The color is determined by a heatmap pattern, for which *green* indicates predictions that

351   were derived from a well-populated model with many examples, while *red* indicates those for

352   which very little training data was available. As can be seen, the outliers that rank very poorly

353   relative to the false positives are all colored red, which strongly suggests that poor performance

354   is due to sparsity of training data, rather than flaws with the method.

355   In Figure 3 (c), the method for scoring documents is set to the *frequency* of each annotation in

356   the overall training set, e.g. if an annotation occurs 100 times in 698 documents, its score is set

357   to 0.143. The same proposed ranking order is used for all documents, regardless of the text

358   description. This is used to test a reasonable null hypothesis, which is that picking the most

359   common annotations is an effective way to separate correct from incorrect. While it can be

360   clearly seen that the null hypothesis performs better than a random guess, at least for purposes

361   of identifying true positives, it is vastly inferior to the proposals generated by the trained

362   Bayesian-derived models, on account of the fact that every document has a very large number

363   of false positives that need to be eliminated before the annotation is complete.

364   Figure 3 (d) shows the same process as for (a), except that in this case the training data is used,

365   i.e. the models are used to predict the same documents from which they were trained. These

366   results are superior to applying the models to the test set, which is to be expected.

Operator workflow

368   The ultimate goal of combining machine learning with a user interface for bioassay annotation is

369   to have the models predict all the correct annotations with close to perfect accuracy, and have

370   the expert operator confirm these predictions. In practice this is realistic only when the document

371   being annotated consists of cases that are well covered in the training set. Due to the nature of

372   science, there will always be new methods being developed, which means that some of the

373   corresponding annotations may have insufficient examples to create a model. It is also possible

374 that the choice of phrasing for some of the assay details differs significantly from the language

375 used by the examples in the training set, which can reduce the efficacy of the models, until

376 additional data can be incorporated and used to re-train them.

377 For these reasons, the user interface needs to satisfy two main scenarios: 1) when the

378 predictions are correct, and 2) when the document is unable to accurately predict the

379 annotations. For the first scenario, confirming the correct annotations should be a matter of

380 quickly scanning the highest scoring proposals and confirming the assignments. In these cases,

381 the user interface must aspire to being unobtrusive. However, in the second scenario, when the

382 correct annotation does not appear at the top of the list, the interface needs to provide ways for

383 the user to hunt down the necessary annotations. We have conceived several options to help the

384 user deal with this case. In near-ideal cases, the user may find the correct annotation by simply

385 looking slightly further down the list of proposed annotations. Alternatively, the user may filter the

386 results by selecting a particular category of annotations, and browse through this refined subset

387 to find the correct annotation. Finally, if the user needs to include an annotation that is present in

388 the ontology, but has not been included in the list of proposals because there is not enough data

389 to build a model, the interface can provide assistance in searching through all of the unscored

390 options. Furthermore, there will also be occasions when the desired annotation does not exist in

391 the ontology, e.g. a previously unreported biological pathway, in which case it may be desirable

392 to allow the user to enter the information as plain text. While this has little immediate value for

393 semantic purposes, it could be considered as a placeholder for future additions to the underlying

394 ontology, which could be upgraded retroactively.

395 A mockup of the core elements of this interface is shown in Figure 4, which shows the same

396 layout principles for the proof of concept application that we created for testing the machine

397 learning methods and corresponding workflow. The box shown at the top left allows the user to

398 type in free text. This could be cut-and-pasted from another application, or it could be typed in

399 manually. The list immediately below shows a series of annotations, consisting of *property* and

400 *value*. These are ranked highest first. When the system is working perfectly, the user can click on

401 the *approve* button for the highest scoring annotation, shown at the top of the list. If the highest

402   scoring annotation is not correct, the user may look further down the list in order to find one that

403   is correct; or, they may *reject* an incorrect proposal. In either case, the proposals are

404   recomputed, and a new list of options is shown.

405   On the right hand side of the screen is shown all of the available properties, which are used to

406   organize the annotations: for each property or category, there can be zero-or-more assigned

407   annotations, of the property:value form. This simple hierarchical arrangement clearly shows the

408   annotations that have been assigned so far, and which properties have as yet no associations.

409   Making the property icons clickable is a way to allow filtering of the annotation list, i.e. only

410   showing the potential annotations that match the selected property. In this way, the operator can

411   carefully pick out assignments for each of the property groups, which is a workflow that becomes

412   important when working with documents that do not fall within the domain of pretrained data.

413   This process of picking out the correct assignment can either be done by scrolling through the

414   list of all possible annotations ranked according to the predictive score, or by partial text

415   searching.

416   *Domain example*

417   Figure 5 shows the annotation of an assay, which can be found in PubChem (Assay ID 761). The

418   annotation text has been composed by concatenating the assay description and protocol text

419   fields, and trimmed to remove superfluous content, which is shown in (a). This case is an

420   example where the performance of the machine learning models is strong, but still requires a

421   well-designed user interface for the portions that are less well covered.

422   Steps (b) through (y) show each of the assignment steps: in most of these examples, the 5

423   highest ranked annotations are shown. In most of the initial steps, the top ranked case is a

424   correctly predicted annotation. A green checkbox is used to indicate that the user confirms that

425   presence of the annotation, and in the following step, the list of proposals is updated to reflect

426   the modified scores, which take into account the correlation effects. In cases (l), (q), (r) and (t),

427   the top ranked prediction is incorrect, and a red cross mark indicates that the user explicitly

428   excludes the annotation from further consideration. In step (w) the desired annotation is further

429 down the list, and so the user scrolls the proposals in order to select the next correct one. In step

430 (x), the user needs to add the annotation *bioassay type : binding* assay, which has not been

431 ranked well in the overall scheme, and so the list of annotations is filtered by selecting the

432 *bioassay type* property, to only show these corresponding values. In step (y) the user is looking

433 to find the *GPCR signal pathway* annotation, which is not a part of the training set, due to

434 insufficient data to build a model. In order to locate this annotation, the user enters a search

435 string to narrow down the list and locate it.

436 In Figure 5 (z), the complete list of annotations, divided into property categories, is shown. This

437 list is updated dynamically as each of the annotations is added to the collection. The properties

438 for *cell line*, *assay kit* and *inducer* have no corresponding annotations, since these are not a part

439 of the assay.

440 ### *Semantic output*

441 The purpose of adding semantic annotations to bioassays is to enable a diverse range of queries

442 and automated analysis, and one of the most effective ways to enable this is load the annotation

443 markup into the same framework as the original BioAssay Ontology definition and all of its

444 related dependencies.

445 The output from an annotated description can easily be expressed in terms of RDF triples. The

446 properties and values are already mapped into the BAO space. A new URI needs to be defined

447 for each of the assays being annotated. For example, the annotation example used earlier,

448 converted into RDF "Turtle" format, is shown in Figure 6.

449 Once in this format, the assertions can be loaded into an existing SPARQL server. At this point

450 the content becomes accessible to the full suite of semantic technologies. Combining the generic

451 querying capabilities of the SPARQL syntax, with the semantic structure of the BioAssay

452 ontology, allows a variety of ad hoc questions to be answered.

453 For example, finding a list of annotated documents that make use of a specific assay kit:

```
454    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
455    PREFIX owl: <http://www.w3.org/2002/07/owl#>
456    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
457    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
458    PREFIX bao: <http://www.bioassayontology.org/bao#>
459    PREFIX cdd:
460    <http://www.collaborativedrug.com/bao/curation.owl#>

461    SELECT ?aid WHERE
462    {
463        ?assaykit rdfs:label "HTRF cAMP Detection Kit" .
464        ?has rdfs:label "has assay kit" .
465        ?document ?has ?assaykit .
466        ?document cdd:PubChemAID ?aid
467    }
```

468   This simple query extracts a list of assay identifiers for anything that makes use of a specific

469   manufacturer's cyclic AMP detector. Note that the property and value URIs are matched by cross

470   referencing the label. Based on the training data, this query returns AID numbers 933, 940,

471   1080, 1402, 1403, 1421, 1422 and 488980.

472   A slightly more advanced query can extract information other than just document identifiers:

```
473    SELECT ?instrument ?aid WHERE
474    {
475        ?document bao:BAO_0002855 bao:BAO_0000110 .
476        ?document bao:BAO_0000196 bao:BAO_0000091 .
477        ?document bao:BAO_0000207 bao:BAO_0000363 .
478        ?document bao:BAO_0002865 ?q .
479        ?q rdfs:label ?instrument .
480        ?document cdd:PubChemAID ?aid
481    }
482    ORDER BY ?instrument ?aid
```

483   In this case the restrictions are specified by directly referencing the BAO tags, which searches

484   for all protein-small molecule interaction assays, with inhibition as the mode of action, using

485   fluorescence intensity measurements. For each match, the detection instrument is looked up and

486   cross referenced by label:

EnVision Multilabel Reader	622
PHERAstar Plus	1986

ViewLux ultraHTS Microplate Imager	2323
ViewLux ultraHTS Microplate Imager	485281
ViewLux ultraHTS Microplate Imager	489008

487 The inheritance hierarchy of the BioAssay Ontology, and the ontologies it references, can also

488 be utilized in queries. The following query looks for assays that target GPCRs of mammals:

```
489    SELECT ?organism ?aid WHERE
490    {
491        ?mammal rdfs:label "mammalian" .
492        ?target rdfs:subClassOf* ?mammal .
493        ?target rdfs:label ?organism .
494        ?document bao:BAO_0002921 ?target .
495        ?q rdfs:label "G protein coupled receptor" .
496        ?document bao:BAO_0000211 ?q .
497        ?document cdd:PubChemAID ?aid
498    }
499    ORDER BY ?organism ?aid
```

500 The target variable is used to match any organism URI that is a subclass of mammals. The

501 result is a number of assays for humans, rats and mice.

502 Each of these examples shows how the semantic markup of the annotated assays can be put to

503 the test with very direct and specific adhoc questions. These queries can be composed on the fly

504 by software that provides a more intuitive user interface, or they can be used for developing new

505 kinds of analyses by experts. They can be applied to just the bioassay data in isolation, or they

506 can be spliced into the greater semantic web as a whole, and linked to all manner of other

507 information resources, e.g. screening runs measuring drug candidates, or medical

508 knowledgebases that go into more detail about the biological systems being assayed.

509 ### *Future work*

510 The hybrid interactive/machine learning approach to bioassay annotation is currently a proof of

511 concept product. The prototype user interface is presently being evaluated by scientists with an

512 interest in improving software annotation of biological data, and we are actively assessing the

513   results in order to improve the workflow. The long-term goal is to provide the user interface in the

514   form of a web application, which will be incorporated into larger products that provide data

515   capture functionality, such as the CDD Vault developed by Collaborative Drug Discovery Inc. or

516   potentially public databases such as PubChem. The semantic annotations will be recorded

517   alongside the text description, and immediately accessible, sharable, searchable and used by a

518   variety of features that can provide reasoning capabilities based on this data.

519   One of the obvious advantages of having user-approved annotations stored in a centralized

520   location is that the machine learning models can be retrained at periodic intervals, which will

521   ensure that the ease with which users can rapidly annotate their assays continues to improve as

522   more data is submitted. Also, as more data becomes available, the domain of the models will

523   continue to grow: annotations that were previously left out of the model building process due to

524   insufficient case studies will be added once they have been used.

525   Another potential advantage of centralization is to provide a pathway for new semantic

526   annotations, i.e. when the BioAssay Ontology and its dependencies do not provide an

527   appropriate term, users can resort to using a free text placeholder. Such annotations can be

528   examined on a regular basis, and either a manual or automated process can be devised to

529   collect together repeated use of related terms, and define a new annotation (e.g. for a new class

530   of biological target or a new measurement technique). This requires a single authority to decide

531   on a universal resource identifier (URI) for the new term, which could be done by the service

532   provider hosting the data, who may also take the opportunity to retroactively upgrade the existing

533   examples of free text labels to use the freshly minted semantic annotation. We have also

534   demonstrated creating a file containing RDF triples for the resulting annotations for a document,

535   and are looking into harmonizing the data format with the Assay Definition Standard format

536   (ADS/ADF)(de Souza et al. 2014; Website 2014d).

537   In addition to working with potential users of this software, we are also looking to incorporate

538   more public content, from large collection services such as PubChem (Wang et al. 2014; Zhang

539   et al. 2011), BARD (de Souza et al. 2014), ChEMBL(Bellis et al. 2011) and OpenPHACTS

540   (Williams et al. 2012). There are a number of research groups exploring ways to add semantic

541 markup to drug discovery data, including bioassays, and many of these annotations can be

542 mapped to the BAO annotations that we have chosen for this project. Even though we have

543 found in our internal evaluation efforts that annotation time can be plausibly reduced to a matter

544 of minutes, this is still a significant burden to impose on busy scientists, especially if participation

545 is voluntary. As we consider deployment of the service, it is important to ensure that the benefits

546 of assay annotation are realized as early as possible, rather than waiting for critical mass, which

547 might otherwise not be achieved. Allowing scientists to use their annotated assays to easily

548 search for similar assays within a database, or as a convenient way to label and categorize their

549 own collections of assays, are anticipated to be effective strategies to make the technology

550 useful during the early adoption phase.

551 Thinking broadly, one interesting possible use case for the annotation scheme is to run it in

552 reverse: to have the software use the annotations to assemble a paragraph of plain English text,

553 which is suitable for incorporation into a manuscript. In this case the workflow would likely be

554 quite different, e.g. the user types in a poorly formatted collection of terms involved in the assay

555 in order to help the inference engine rank the likely suggestions, selects the appropriate terms,

556 and then has the text produced by the software. Such a service could be of significant use to

557 scientists who are not experienced with writing assay procedures according to the standard style

558 guides.

559 As part of our ongoing work, we are evaluating our selection of annotations from the underlying

560 ontology. Our initial prototype is strongly influenced by the training data that we have available,

561 which is the result of hundreds of hours of work by qualified domain experts. We are actively

562 working with biologists and medicinal chemists to determine which properties are of primary

563 importance, and which are secondary, and to expand our collection of training data to reflect the

564 priorities of active drug discovery researchers.

565 Beyond the use of bioassays and BAO annotations for training data, the methodology developed

566 is broadly applicable and not specific to this domain. We anticipate that there are a number of

567 other distinct subject areas of scientific publications that would be amenable to this treatment,

568 e.g. experimental details of chemical reactions, computational chemistry protocols, and other

569 types of biological protocols beyond drug discovery, such as stem cell differentiation.

Conclusion

571 We have built a proof of concept framework that involves using machine learning based on plain

572 text assay descriptions and curated semantic markup, and matched this with a user interface

573 that is optimized for making machine-assisted annotation very rapid and convenient when

574 applied to text input that is well within the domain, and moderately efficient for annotating assays

575 that fall outside of the training set. By optimizing both the machine learning and user-centric

576 workflow at the same time, we avoid falling into the traps of both extremes, because both parts

577 complement each other. Annotation of plain text by purely automated methods has been limited

578 by the need to obtain an unrealistic level of accuracy, while purely manual annotation has to

579 overcome a very high motivational barrier, given that most scientists are too busy to take on

580 additional burdens, without an immediate benefit. By establishing that adding a very modest

581 amount of human effort to a well designed automated parser can achieve highly accurate

582 results, we believe that we can make a strong case for the use of this technology in the hands of

583 practicing scientists.

584 As the quantity of semantically rich annotated data increases, the opportunities for delivering

585 value to scientists increases in tandem. Making annotation easy is the first step, but it needs to

586 be followed by new capabilities. For example, the creators of assay screens should be able to

587 easily compare their experiments with others contained within the knowledgebase, and obtain a

588 list of experiments and published papers with common features. Researchers performing drug

589 discovery modeling studies should be able to gather together compounds that have been

590 screened under certain conditions, and use the annotations to make a judgment call as to

591 whether the measured activities can be incorporated into the same model. Additionally,

592 researchers can search for artifacts, such as compounds that are disproportionately active in

593 luminescent assays. New biological activities may also become mineable; for example, common

594 hits between cell-based assays and target based assays may reveal unknown molecular

595 mechanisms.

596 Beyond the specific domain of bioassay annotation, we believe that the hybrid approach to high

597 level markup is appropriate to many different areas of science, where use of English text jargon

598 or anachronistic diagrams is the norm for conveying concepts that are amenable to a highly

599 structured description. The understandable reluctance of scientists to redesign their

600 communication methods for the benefits of software, and the inability of software to provide

601 substantially useful results without such richly marked up data, is a proverbial chicken vs. egg

602 scenario that can be observed throughout the scientific disciplines. Combining machine learning

603 with modest demands on scientists' time, and rapid iteration of improved functionality, is a viable

604 strategy for advancing the goals of computer assisted decision support.

605 **Figure Captions**

606 **Figure 1**: Selected leave-one-out ROC plots for annotations, using Bayesian learning models

607 derived from marked-up natural language processing.

608 **Figure 2**: Representative examples of model building in action, showing relative ranking of

609 uncalibrated, calibrated, and stepwise application of the correlation models. The four examples

610 refer PubChem entries by assay ID: (a) 574, (b) 436, (c) 348 and (d) 346.

611 **Figure 3**: Effectiveness of ranking of activities: (a) hit/miss for test data; (b) heatmap for model

612 size; (c) null hypothesis; (d) hit/miss for training data.

613 **Figure 4**: A mockup of an interactive graphical user interface for annotating bioassays, with

614 guidance from pretrained models.

615 **Figure 5**: Stepwise annotation process for PubChem Assay ID 761,

616 http://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/aid/761/description/JSON

617   **Figure 6**: RDF Triples for the annotation of PubChem assay ID 761.

618   **References:**

619   Abeyruwan SV, UD; Küçük-McGinty, H; Visser, U; Koleti, A; Mir, A; Sakurai, K; Chung, C;
620   Bittker, JA; Clemons, PA; Brudz, S; Siripala, A; Morales, AJ; Romacker, M, Twomey, D;
621   Bureeva, S; Lemmon,V; Schürer, SC. 2014. Evolving BioAssay Ontology (BAO):
622   modularization, integration and applications. *J Biomed Semantics* 5(Suppl 1):S5.
623   Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, and Thorne D. 2010. Utopia
624   documents: linking scholarly literature with research data. *Bioinformatics* 26:i568-574.
625   Balakrishnan R, Harris MA, Huntley R, Van Auken K, and Cherry JM. 2013. A guide to best
626   practices for Gene Ontology (GO) manual annotation. *Database (Oxford)* 2013:bat054.
627   Bellis LJ, Akhtar R, Al-Lazikani B, Atkinson F, Bento AP, Chambers J, Davies M, Gaulton A,
628   Hersey A, Ikeda K, Kruger FA, Light Y, McGlinchey S, Santos R, Stauch B, and
629   Overington JP. 2011. Collation and data-mining of literature bioactivity data for drug
630   discovery. *Biochem Soc Trans* 39:1365-1370.
631   Blake JA. 2013. Ten quick tips for using the gene ontology. *PLoS Comput Biol* 9:e1003343.
632   Cohn DAG, Z.; Jordan, M.I. 1996. Active Learning with Statistical Models. *J Artifical
633   Intelligence Research* 4:129-145.
634   Dara AvG, J.; Liu, Q.; Judge, J.; Toral, A.; . 2014. Active Learning for Post-Editing Based
635   Incrementally Retrained MT. *Proceedings of the 14th Conference of the European
636   Chapter of the Association for Computational Linguistics*:185-189.
637   de Souza A, Bittker JA, Lahr DL, Brudz S, Chatwin S, Oprea TI, Waller A, Yang JJ, Southall N,
638   Guha R, Schurer SC, Vempati UD, Southern MR, Dawson ES, Clemons PA, and Chung
639   TD. 2014. An Overview of the Challenges in Designing, Integrating, and Delivering
640   BARD: A Public Chemical-Biology Resource and Query Portal for Multiple
641   Organizations, Locations, and Disciplines. *J Biomol Screen* 19:614-627.
642   Federhen S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res* 40:D136-143.
643   Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S,
644   Michalovich D, Al-Lazikani B, and Overington JP. 2012. ChEMBL: a large-scale
645   bioactivity database for drug discovery. *Nucleic Acids Res* 40:D1100-1107.
646   Hassan M, Brown RD, Varma-O'brien S, and Rogers D. 2006. Cheminformatics analysis and
647   learning in a data pipelining environment. *Mol Divers* 10:283-299.
648   Hawizy L, Jessop DM, Adams N, and Murray-Rust P. 2011. ChemicalTagger: A tool for semantic
649   text-mining in chemistry. *J Cheminform* 3:17.
650   Jessop DM, Adams SE, and Murray-Rust P. 2011a. Mining chemical information from open
651   patents. *J Cheminform* 3:40.
652   Jessop DM, Adams SE, Willighagen EL, Hawizy L, and Murray-Rust P. 2011b. OSCAR4: a
653   flexible architecture for chemical text-mining. *J Cheminform* 3:41.
654   Jonquet C, Musen MA, and Shah NH. 2010. Building a biomedical ontology recommender web
655   service. *J Biomed Semantics* 1 Suppl 1:S1.

Jonquet C, Shah NH, and Musen MA. 2009. The open biomedical annotator. *Summit on Translat Bioinforma* 2009:56-60.

Kang YS, and Kayaalp M. 2013. Extracting laboratory test information from biomedical text. *J Pathol Inform* 4:23.

Khabsa M, and Giles CL. 2014. The number of scholarly documents on the public web. *PLoS One* 9:e93949.

Leaman R, Islamaj Dogan R, and Lu Z. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 29:2909-2917.

Liu K, Hogan WR, and Crowley RS. 2011. Natural Language Processing methods and systems for biomedical ontology learning. *J Biomed Inform* 44:163-179.

Mussa HY, Mitchell JB, and Glen RC. 2013. Full "Laplacianised" posterior naive Bayesian algorithm. *J Cheminform* 5:37.

Nidhi, Glick M, Davies JW, and Jenkins JL. 2006. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 46:1124-1133.

Roeder C, Jonquet C, Shah NH, Baumgartner WA, Jr., Verspoor K, and Hunter L. 2010. A UIMA wrapper for the NCBO annotator. *Bioinformatics* 26:1800-1801.

Rogers D, Brown RD, and Hahn M. 2005. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen* 10:682-686.

Santorini B. 1990. Part-of-Speech Tagging Guidelines for the Penn Treebank Project *Available at* http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports

Sarntivijai SX, Z.; Meehan, T.F.; Diehl, A.D.;Vempati, U.; Schurer, S.C.; Pang, C.; Malone, J.; Parkinson, H.; Athey, B.D.; He, Y.; . 2011. Cell Line Ontology: Redesigning the Cell Line Knowledgebase to Aid Integrative Translational Informatics. ICBO: International Conference on Biomedical Ontology. Buffalo, NY, USA.

Schurer SC, Vempati U, Smith R, Southern M, and Lemmon V. 2011. BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *J Biomol Screen* 16:415-426.

Tomanek KW, J.; Hahn, U. 2007. An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and ComputationalNatural Language Learning*:486-495.

Vempati UD, Przydzial MJ, Chung C, Abeyruwan S, Mir A, Sakurai K, Visser U, Lemmon VP, and Schurer SC. 2012. Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay Ontology (BAO). *PLoS One* 7:e49198.

Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, and Schurer SC. 2011. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics* 12:257.

Wang Y, Suzek T, Zhang J, Wang J, He S, Cheng T, Shoemaker BA, Gindulyte A, and Bryant SH. 2014. PubChem BioAssay: 2014 update. *Nucleic Acids Res* 42:D1075-1082.

Website. 2014a. http://jena.apache.org (Accessed May 2014).

Website. 2014b. http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html (Accessed June 2014).

Website. 2014c. http://opennlp.apache.org (Accessed June 2014).

Website. 2014d. https://sites.google.com/site/assaydefinitionstandard (Accessed May 2014).

Wei CH, Kao HY, and Lu Z. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* 41:W518-522.

703 Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg
704   N, Ecker G, Goble C, and Mons B. 2012. Open PHACTS: semantic interoperability for
705   drug discovery. *Drug Discov Today* 17:1188-1198.
706 Zhang JD, Geer LY, Bolton EE, and Bryant SH. 2011. Automated annotation of chemical names
707   in the literature with tunable accuracy. *J Cheminform* 3:52.

Figure 1

Selected leave-one-out ROC plots for annotations, using Bayesian learning models derived from marked-up natural language processing.

Figure 1: Selected leave-one-out ROC plots for annotations, using Bayesian learning models derived from marked-up natural language processing.
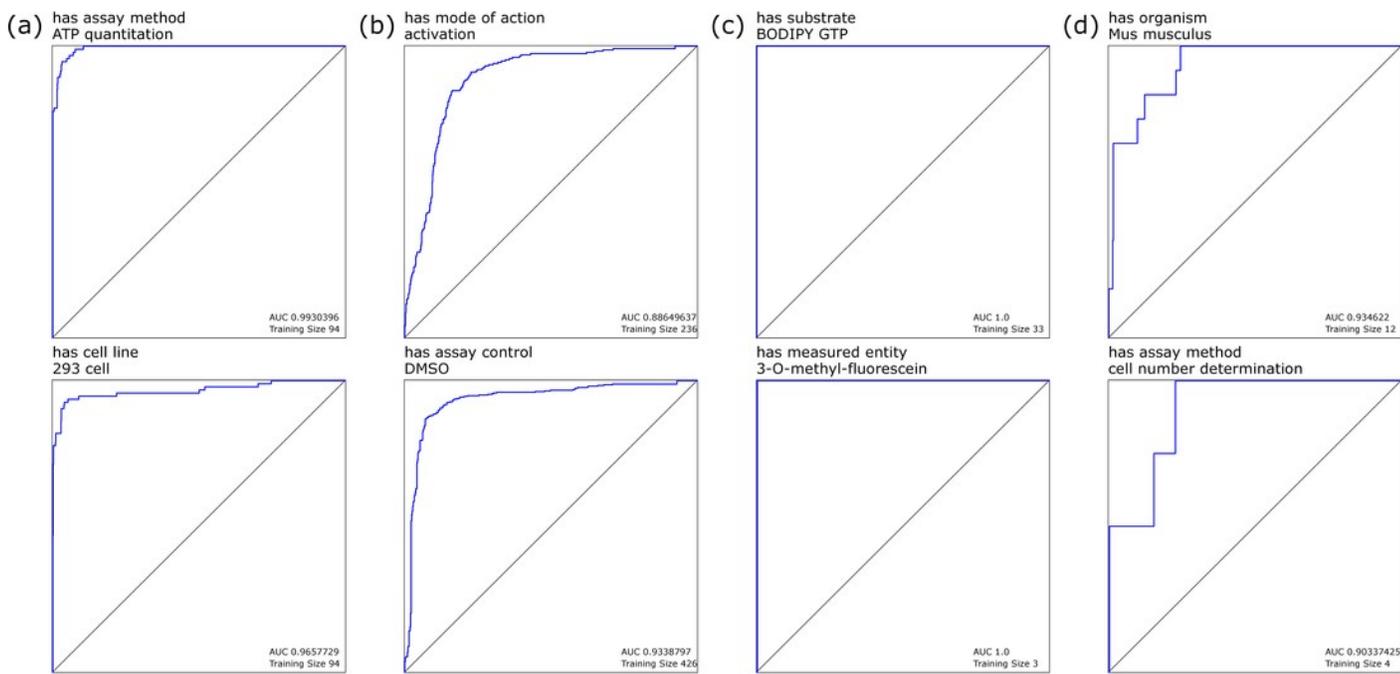
Figure 2

Representative examples of model building in action.

Figure 2: Representative examples of model building in action, showing relative ranking of uncalibrated, calibrated, and stepwise application of the correlation models. The four examples refer PubChem entries by assay ID: (a) 574, (b) 436, (c) 348 and (d) 346.
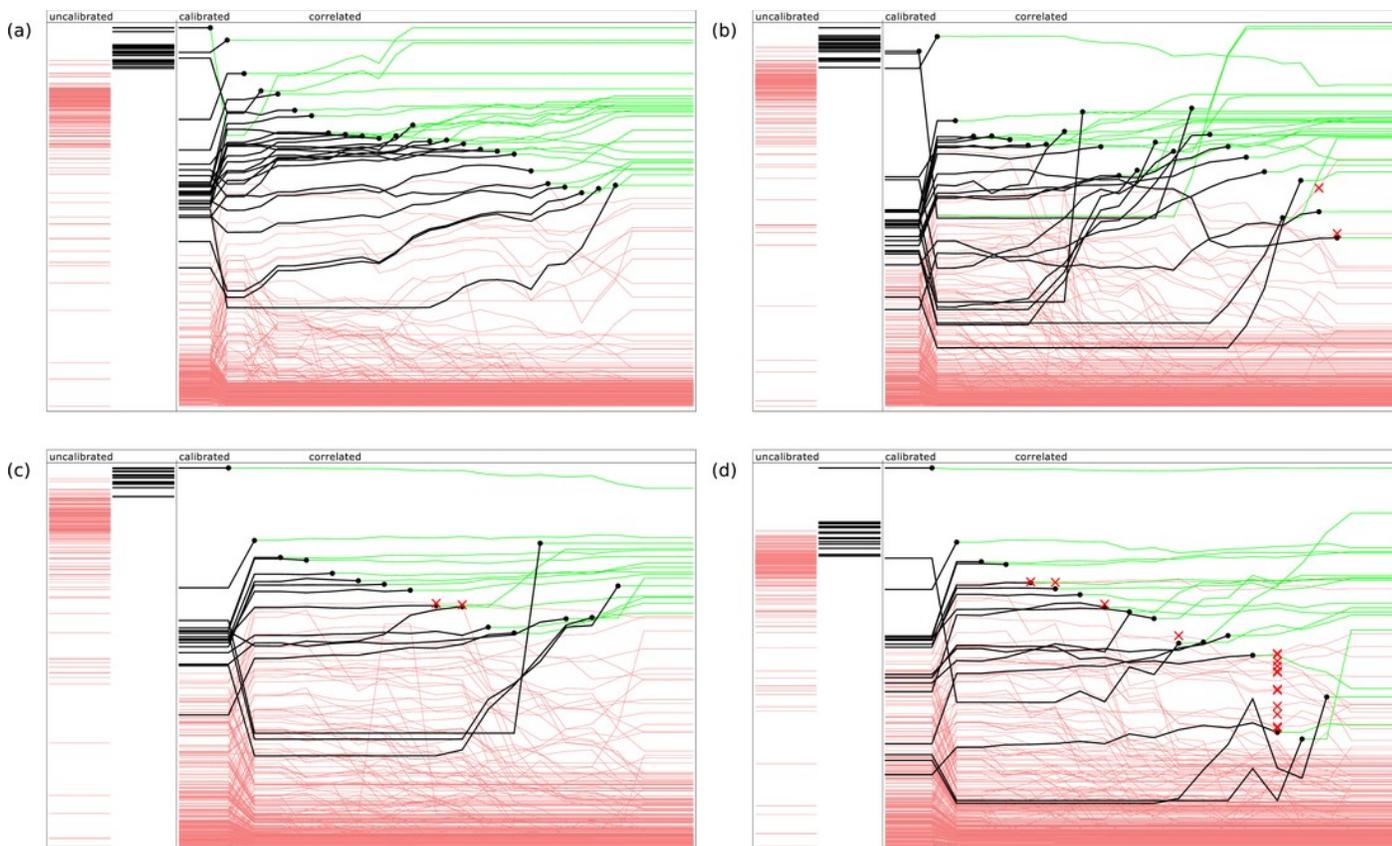
Figure 3

Effectiveness of ranking of activities

Figure 3: Effectiveness of ranking of activities: (a) hit/miss for test data; (b) heatmap for model size; (c) null hypothesis; (d) hit/miss for training data.
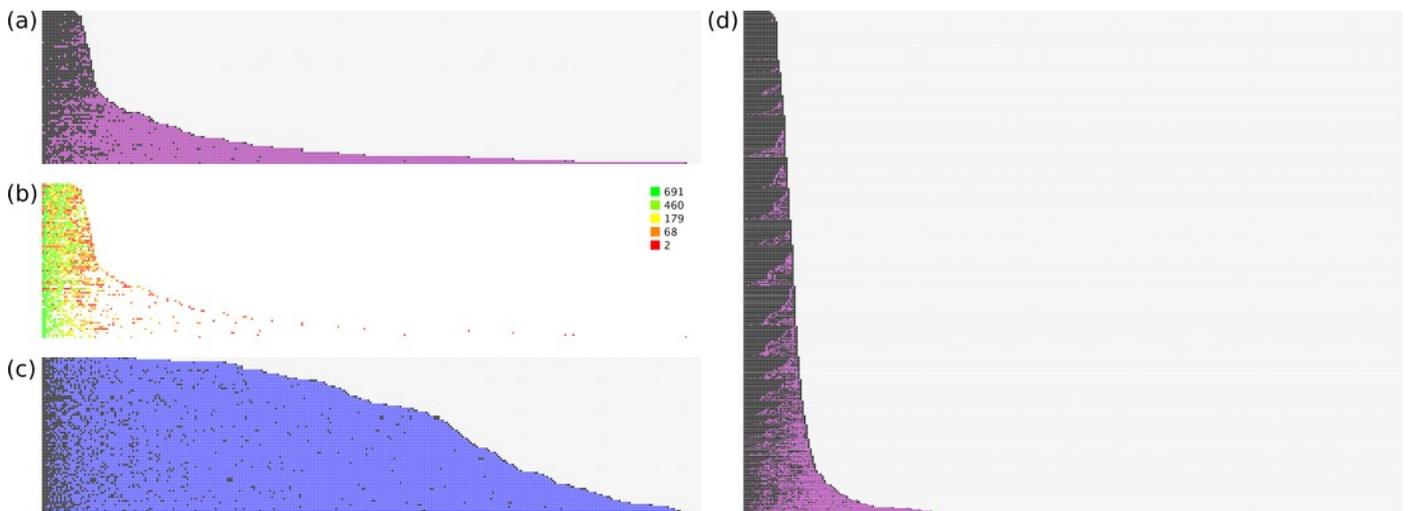
Figure 4

A mockup of an interactive graphical user interface for annotating bioassays, with guidance from pretrained models.

Figure 4: A mockup of an interactive graphical user interface for annotating bioassays, with guidance from pretrained models.

Figure 5

Stepwise annotation process for PubChem Assay ID 761,

http://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/aid/761/description/JSON

Figure 5: Stepwise annotation process for PubChem Assay ID 761,

http://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/aid/761/description/JSON
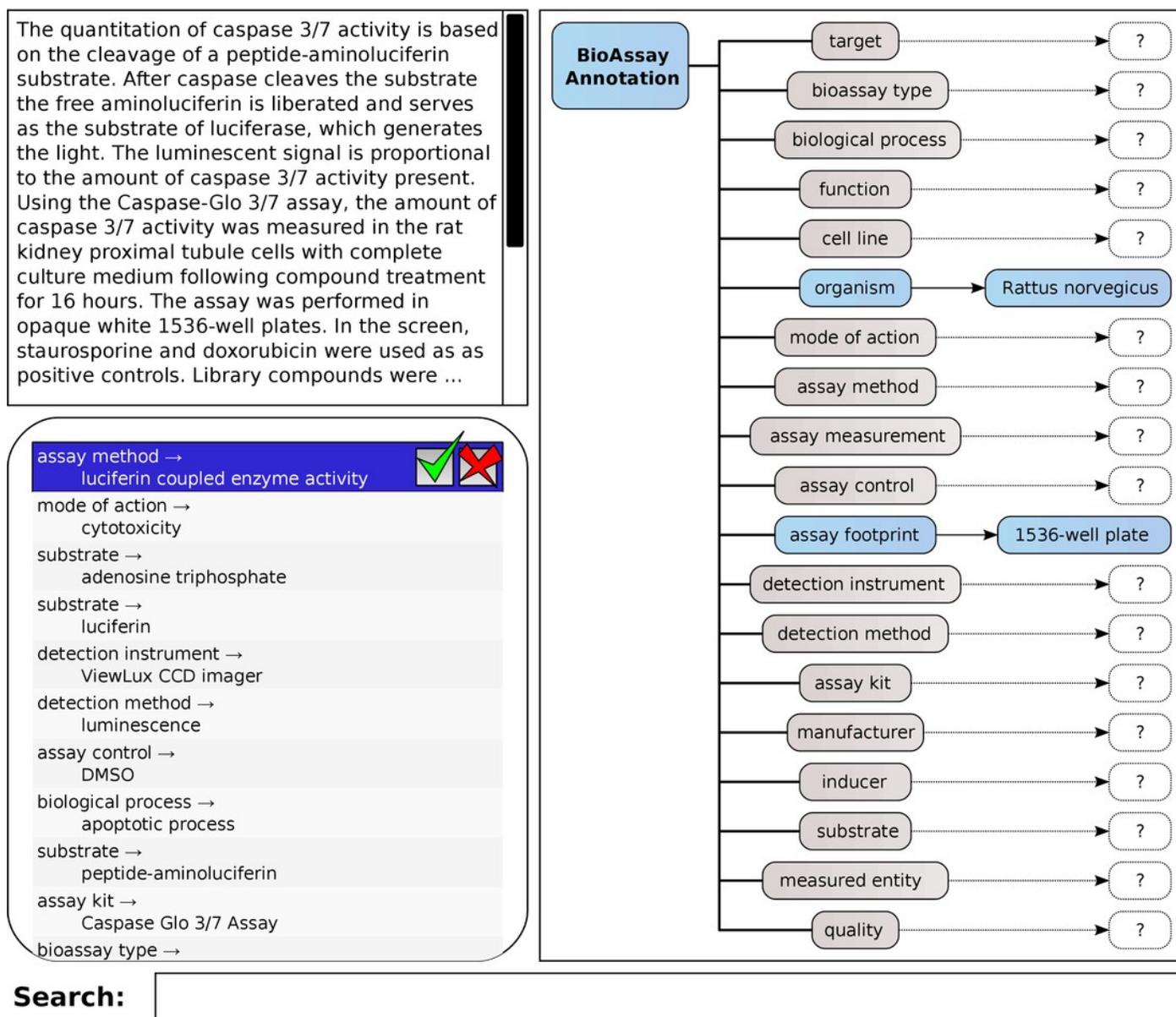
Figure 6

RDF Triples for the annotation of PubChem assay ID 761.

Figure 6: RDF Triples for the annotation of PubChem assay ID 761.

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix tag: <http://purl.org/ontology/tag/> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core> .
@prefix bao: <http://www.bioassayontology.org/bao#> .
@prefix : <http://www.collaborativedrug.com/bao/curation.owl#> .

#### base classes for assays

:AnnotatedAssay
    rdfs:label "BioAssay with Annotations" ;
    rdf:type owl:Class ;
    .

:PubChemAssay
    rdfs:label "Annotated Assay from PubChem" ;
    rdfs:subClassOf :AnnotatedAssay ;
    .

#### curation of assays from PubChem using BioAssay Ontology properties & values
#### provided by Collaborative Drug Discovery, Inc.

## all user-curated entries for PubChem assay #761
## see: http://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/aid/761/description/JSON

:AssayPubChem_761
    rdfs:label "Annotations for PubChem Assay ID 761" ;
    rdfs:subClassOf :PubChemAssay ;
    :PubChemAID "761" ;
    bao:BAO_0002855 bao:BAO_0000110 ; # "is bioassay type of" ->
                                      # "protein-small molecule interaction assay"
    bao:BAO_0002854 bao:BAO_0000041 ; # "has bioassay type" -> "binding type"
    bao:BAO_0002009 # "involves biological process"
        <http://purl.obolibrary.org/obo/GO_0007186> ; # -> "GPCR signaling pathway"
    bao:BAO_0002921 # "has organism"
        <http://purl.obolibrary.org/obo/NCBITaxon_9606> ; # -> "Homo sapiens"
    bao:BAO_0000196 bao:BAO_0000091 ; # "has mode of action" -> "inhibition"
    bao:BAO_0000196 bao:BAO_0000087 ; # "has mode of action" -> "activation"
    bao:BAO_0000205 bao:BAO_0000357 ; # "has assay format" -> "single protein format"
    bao:BAO_0000205 bao:BAO_0000217 ; # "has assay format" -> "biochemical format"
    bao:BAO_0000212 bao:BAO_0000123 ; # "has assay method" -> "binding assessment method"
    bao:BAO_0000409 bao:BAO_0000410 ; # "assay measurement type" -> "endpoint assay"
    bao:BAO_0000740 bao:BAO_0000079 ; # "has assay control" -> "negative control"
    bao:BAO_0000740 bao:BAO_0000080 ; # "has assay control" -> "positive control"
    bao:BAO_0002867 bao:BAO_0000515 ; # "has assay footprint" -> "384 well plate"
    bao:BAO_0002865 bao:BAO_0000943 ; # "uses detection instrument" ->
                                      # "HyperCyt High Throughput Flow Cytometry System"
    bao:BAO_0000207 bao:BAO_0000005 ; # "has detection method" -> "flow cytometry"
    bao:BAO_0000737 bao:BAO_0000946 ; # "has manufacturer" -> "IntelliCyt Corporation"
    bao:BAO_0002739 bao:BAO_0000931 ; # "has substrate" -> "BODIPY GTP"
    bao:BAO_0002000 bao:BAO_0000931 ; # "has measured entity" -> "BODIPY GTP"
    bao:BA0_0090012 bao:BAO_0000176 ; # "has participant" -> "small molecule"
    bao:BA0_0090012 bao:BAO_0000895 ; # "has participant" -> "DTT"
    bao:BA0_0090012 bao:BAO_0000693 ; # "has participant" -> "HEPES"
    bao:BA0_0090012 bao:BAO_0000978 ; # "has participant" -> "NP-40"
    bao:BA0_0090012 bao:BAO_0000368 ; # "has participant" -> "G protein"
    bao:BA0_0090012 bao:BAO_0000931 ; # "has participant" -> "BODIPY GTP"
    bao:BAO_0002662 bao:BAO_0002157 ; # "has quality" -> "native form"
```