Natural language processing methods in biological activity prediction

Szymon Nakoneczny[1] and Marek Śmieja[1]

Faculty of Mathematics and Computer Science, Jagiellonian University
Łojasiewicza 6, 30-348 Kraków, Poland
{szymon.nakoneczny, marek.smieja}@ii.uj.edu.pl

Abstract. Virtual screening is a process in which databases of chemical compounds are searched in order to find structures characterized with a high biological activity, possible drug candidates [12]. Our goal is to combine the natural language processing methods with SMILES, a text representation of compounds, in the classification of active compounds. Since SMILES encodes a graph structure of a compound into a sequence of symbols, it is not reasonable to build a language model directly from SMILES. In this paper we propose various strategies to adjust the underlying representation to create the best possible language model of compounds. The introduced modifications are verified in an extensive experimental study. The results show that the proposed approach outperforms classical fingerprint representations and does not require any specialized chemical knowledge.

Keywords: n-gram language model, bag-of-n-grams, support vector machine, virtual screening, SMILES

1 Introduction

Only a small percentage of a huge number of organic compounds can serve as drugs. As it is not possible to synthesize all of them in a laboratory, the first move is to find biological active compounds by means of a computer analysis. Before machine learning, a common approach was to simulate a compound to target protein docking and estimate a ligand activity [13]. Currently, one of the most popular methods is to represent chemical compounds as a vector called fingerprint and analyze them with machine learning approaches. Each position in this representation refers to some substructure of a compound's graph and often in order to use the most meaningful substructures, a huge number of them is being handcrafted by chemists [11].

Our goal is to apply natural language processing methods with a text representation called SMILES. In this representation, a compound graph is flattened into a sequence of symbols which makes the structure information much harder to understand. Fortunately, working with a representation which computational needs matches the use of fingerprints while still having all of the information encoded inside of the representation opens new possibilities for applying machine learning methods. The effort is taken to use and improve n-gram language model and bag-of-words representation known from NLP. In scope of improvements, SMILES modifications with an aim to increase its informativity will be tested. The best modifications were chosen with experimental analysis which also proves that our solution can achieve higher scores than standard fingerprint representations without a use of any specialized chemical knowledge.

The paper is organized as follows. Next section gives a brief review of related SMILES based approaches. Section 3 provides a more detailed description of SMILES representation and recalls n-gram language model. The proposed modifications, which allow to combine SMILES with NLP, are placed at the end of this section. Experimental results are included in Section 4. Finally, the conclusion is given.

2 Related work

The SMILES-based approaches to chemical compounds analysis are rather rare. A problem of bioactivity prediction with text representation was tackled by Apilak Worachartcheewan et al. [16]. The idea of this approach is to define a set of features based on a presence of some abstract subsequences in SMILES, which should be relevant for activity analysis, and then optimizing the model with Monte Carlo approach. It is worth to mention that this approach is very similar to building a structural fingerprint.

David Vidal et al. [14] proposed to build a general vector representation based on all subsequences of a given length found in SMILES. This representation can be then used in any task concerning chemical compounds analysis.

In comparison to those ideas, our solution is mostly motivated by natural language processing achievements. Due to the n-gram language model characteristics, subsequences used are varying in length and their set is not limited by any domain knowledge. It is a model objective to learn the value of subsequences given the problem of bioactivity prediction. Thereby, our approach does not require much of organic chemistry knowledge.

3 Natural language processing with SMILES

In this section, we provide more detailed description of SMILES representation and recall the n-gram language model. We also propose here various strategies how to combine these two tools to create an efficient representation for applying machine learning methods.

3.1 SMILES representations

SMILES (Simplified Molecular Input Line Entry System) is a simple, easy to understand language in which molecules and chemical reactions can be written with

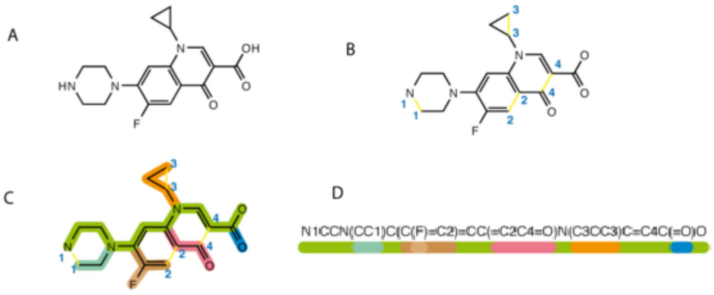


Fig. 1: Example of SMILES representation.

ASCII characters representing atoms and bonds [3]. Its most important advantage is an ability to uniquely identify chemical compound, which is something that simple molecular formula does not provide, while taking approximately 50%-70% less space than corresponding matrix representation. SMILES is created by traversing a compound graph and writing down atoms and bonds along the way.

Basic rules which govern this process are (see Figure 1 for example):

- Atoms are represented by their atomic symbol. If an atom does not belong to subset of organic elements or it is an isotope or its charge has to be specified, it is putted with all this information into square brackets. Besides the isotope case, hydrogen atoms are always implied by compound structure in which case they are being omitted.
- Single, double and triple bonds are represented by symbols '-', '=' and '#' respectively. Th default bond is a single one and '-' symbol can be omitted.
- Branches are putted inside brackets.
- Cyclic structures are represented by removing one of the bonds in compound graph in order to create a tree. Removed bond is then marked with one and same digit following an atom which opens and closes this bond. Digits can be reused after a bond closing and in rare cases when number higher than 9 is used, it has to be preceded with '%' symbol.
- Not connected structures are written individually and separated with a '.' symbol.

Extensions to SMILES covers aromaticity and unique representation. In case of aromaticity, the idea is to encode this information with small atomic symbols, thanks to which it can be easily detected without any algorithm and SMILES is being even further simplified. The example is aromatic ring C1=COC=C1 which will be now written as c1cocc1. The rules described above, starting with an order in which a graph is being traversed, are not deterministic. The idea with canonical SMILES is to create additional set of rules and default behaviors in order to create only one unique SMILES for each of the compounds.

3.2 N-gram language model

One of the first problems tackled by natural language processing methods was a missing last word in sentence [7]. N-gram language model, which builds a probabilistic language model based on n previous words, turns out to be a perfect solution to this problem [1]. Having a sentence of words $w_1, \ldots, w_N$, the probability of word w_i occurring after the last n words is given by[1]:

$$P(w_i|w_{i-n+1} \ldots w_{i-1}) = \frac{|w_{i-n+1} \ldots w_i|}{|w_{i-n+1} \ldots w_{i-1}|}$$

Therefore, the probability of a sequence can be calculated as:

$$P(w_1, \ldots, w_N) = \prod_{i=1}^{N} P(w_i|w_{i-n+1} \ldots w_{i-1})$$

Because longer sequences will automatically get lower probability, the perplexity is used to calculate how well a given sentence fits to a created model.

Definition 1. *Perplexity of a $w = w_1, \ldots, w_N$ sequence of elements, is defined as*

$$PPL(w) = P(w_1, \ldots, w_N)^{-\frac{1}{N}}$$

This measure is not dependent on a sequence length and is minimized by the best fitting sequence.

A problem with an n-gram model arises when a sequence contains such n-grams which were not present in a training set, therefore their probability equals 0. Smoothing techniques were introduced to deal with this and other n-gram model problems. One of such methods is Jelinek-Mercer interpolation [17], which idea is to use information of smaller contexts to interpolate the probability of a longer one. It is given by:

$$\begin{aligned}&\hat{P}(w_i|w_{i-n+1} \ldots w_{i-1}) = \\ &\lambda_{w_{i-n+1}^{i-1}} \hat{P}(w_i|w_{i-n+1} \ldots w_{i-1}) + (1 - \lambda_{w_{i-n+1}^{i-1}})\hat{P}(w_i|w_{i-n+2} \ldots w_{i-1})\end{aligned}$$

where $\lambda \in [0, 1]$ parameters can be grouped and then fitted with an expectation-maximization algorithm.

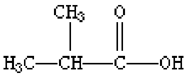
Applications of n-gram model go beyond sentence modeling with words. One can use the perplexity to construct a decision function which allows to classify a given sentence to one of underlying classes. For a simplicity, let us consider a binary classification problem and assume that separate n-gram models were constructed for two groups of texts. Then the probability of assigning a new text x to class c_i, for $i = 1, 2$ equals:

$$P(c_i|x) = \frac{PPL_{c_i}(x)}{PPL_{c_1}(x) + PPL_{c_2}(x)}. \tag{1}$$

N-gram model provides good results in many areas and its biggest advantages are simplicity and scalability. With increase of data, the space requirement grows slower as more n-grams are being repeated.

[1] we replace w_{-i}, for $i = 0, 1, \ldots$ by an empty symbol $< s >$

Table 1: Adding a context to SMILES 3-gram model

compound	SMILES	SMILES with context added
	CC(C)C(=O)O	CC(C)eCCC(=O)eCCO

3.3 Combining SMILES with NLP

The first step of constructing n-gram model relies on the tokenization of text into base symbols (words). This is motivated with balancing information carried by different tokens. It resembles a situation from natural language processing in which word lengths are completely ignored by an n-gram model.

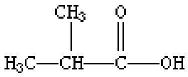
Two approaches are proposed to tokenize SMILES:

- **Baseline.** The simplest way to apply n-gram model to SMILES is to split SMILES into single characters and build a model upon such elements.
- **Tokenization.** Let us observe that it might not be reasonable to split an atom symbol 'Br' int o 'B' and 'r'. Following that idea, elements longer than one symbol, like 'Br', are gathered into single tokens. Also, numbers which are present in square brackets or after '%' symbol are matched together to distinct them from single digits which stand for circular structures.

SMILES provides a text representation of chemical compound structure. However, to build a reasonable N-gram model from SMILES one has to keep in mind that this representation encodes a graph structure into a sequence of symbols. To deal with that problem, the following SMILES modifications are proposed:

- **Simplification.** Idea proposed in [16] and [14] is to simplify circular structure information by replacing all the digits with '0' and to simplify aromaticity information by replacing all the letters with their capitals.
- **Context.** While traversing a branch in graph, ending it and going back to where it starts makes the n-gram model to still read the elements on the end on that branch which can actually be placed far away from a current position. The idea is to duplicate the last $(n-1)$ elements which were read before the branch started and are connected to the current path. Also proceed them with $(n-2)$ symbols 'e' to fully disconnect the structures which are not connected in compound graph. Table 1 shows an example of such SMILES modification.
- **Short paths.** In order not to multiply the information contained in the branching (as in the above context addition), the idea is to only add $(n-2)$ symbols 'e' which results in disconnecting some parts of a graph and working on shorter paths. Similarly to an NLP situation in which individual parts of a sentence are separated with a coma symbol, in case of SMILES such separation is always done with going back to the beginning of a branch represented by symbol ')'.

Table 2: An example of new SMILES creation

compound	SMILES	new reversed SMILES
	CC(C)C(=O)O	CC(C(O)=O)C

– **New SMILES.** Because there is more than one way of traversing a graph, to allow n-gram model to use more of the paths available one can create new SMILES by changing the order of graph traversing. Both sequences can be then separated with a special symbol in order to create one representation. Because the number of all possible SIMILES is extremely high and it makes our model more complex, there is a need to solve an information-complexity trade-off by maximizing the information gain when adding new SMILES. Starting with creation of one additional representation, it was done by reversing the order in which branches are read so that the first branch is read as the last one. Table 2 shows an example. As you can see, non of the paths used before is being repeated in the new SMILES which maximizes the information gain.

4 Experiments

In this section we present a complete experimental analysis, which verifies methods described in previous section. First, we describe the preparation of data. Next, we test how the modifications introduced in Section 3.3 affect the results obtained by applying a perplexity classifier (1). Finally, we compare different classifiers on a vector representation created from the n-gram model.

4.1 Data preparation

Chemical compounds data were downloaded from the ChEMBL database [4]. The target proteins are called receptors and each of those defines different activation value. In order to reliably test our approaches, a set of 6 serotonin receptors was chosen: 5-HT1a, 5-HT6, 5-HT7, 5-HT2a, 5-HT2b and 5-HT2c. They all define separated datasets in which the same compounds may or may not occur. Because activity test on human and rat provides similar results, both of those sets were downloaded. Activity function is measured by an inhibition constant K_i [15]. Compounds with an inhibition constant less than or equal to 100 nM were considered active; ligands with K_i higher than 1000 nM were used as inactive. The range (100, 1000) is removed from data as not clear enough.[2]

[2] From a machine learning point of view, removing those compounds creates an artificial absence of data. However, to make the problem easier and be consistent with the methodology used by chemists, the given approach is used.

Table 3: Number and ratio of different compounds for receptors

	5-HT1a	5-HT6	5-HT7	5-HT2a	5-HT2b	5-HT2c
actives	4376	1597	888	2041	410	1289
inactives	1057	379	365	995	335	1023
ZINC	39384	14373	7992	18369	3640	11601
act. / inact.	4.14	4.21	2.43	2.05	0.32	1.26
act. / ZINC	0.11	0.11	0.11	0.11	0.11	0.11

Unfortunately, bioactivity data can be very noised as varying laboratory conditions can highly change test results. In case of duplicated records, in order to reduce an influence of outliers, a median value is calculated.

The last problem with ChEMBL database is an unbalanced amount of active and inactive elements. Because the majority of molecules in the real world are the inactive ones, their discovery is nothing special and they are not published to databases. It results in a completely opposite compounds ratio. To deal with that problem, ZINC compounds are being used [6]. Those are artificially generated compounds with a high probability of being inactive. Because those compounds are significantly different than the ones obtained from ChEMBL, mixing ZINC with ChEMBL compounds would result in a yet another problem different than the real one. The most common strategy is to create two unbalanced datasets, one containing all the compounds from ChEMBL and another with active molecules from ChEMBL and inactive ones in the form of ZINC. An advantage of the first dataset is a high similarity between active and inactive compounds which makes the problem much harder, while second dataset describes better the real ratio between compounds. Having those datasets and 6 different receptors, it gives us total of 12 problems to solve. The number of compounds and their ratio in all of the datasets is summarized in the Table 3.

The models are tested with 10-fold cross-validation and the error is measured with the Matthews correlation coefficient [9] defined as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where:

TP - number of examples correctly classified as positive

TN - number of examples correctly classified as negative

FP - number of examples incorrectly classified as positive

FN - number of examples incorrectly classified as negative

MCC results vary in range [-1, 1]. The reason for choosing it is the fact that it is one of the most reliable errors in case of unbalanced problems.

Table 4: Mean MCC score of 6-gram model in active and ZINC compounds classification

	5-HT1a	5-HT6	5-HT7	5-HT2a	5-HT2b	5-HT2c
baseline	0.944	0.969	0.943	0.930	0.897	0.928
tokenization	0.944	0.969	0.944	0.931	0.894	0.929
context	0.926	0.963	0.929	0.914	0.879	0.903
short paths	0.926	0.958	0.925	0.909	0.877	0.900
simplification	0.926	0.959	0.930	0.916	0.885	0.915
new SMILES	**0.960**	**0.974**	**0.952**	**0.952**	**0.912**	**0.939**

Table 5: Mean MCC score of 6-gram model in active and inactive compounds classification

	5-HT1a	5-HT6	5-HT7	5-HT2a	5-HT2b	5-HT2c
baseline	0.611	0.737	0.671	0.657	0.562	0.674
tokenization	0.614	0.733	0.668	0.656	**0.564**	0.676
context	0.591	0.731	0.633	0.641	0.553	0.655
short paths	0.601	**0.759**	0.650	0.649	0.542	0.665
simplification	0.593	0.720	**0.672**	0.637	0.552	0.653
new SMILES	**0.621**	0.749	0.666	**0.657**	0.559	**0.697**

4.2 N-gram language model

In this section we verify the usefulness of SMILES modifications proposed in Section 3.3. The goal of experiments will be to make the n-gram model extract from SMILES as much information relevant to bioactivity as possible.

We use a perplexity classifier (1) applied on these models. To refer the particular types of modifications, we use the names given in Section 3.3. In experiments, the KenLM implementation of n-gram model is used [5]. Both 3 and 6-grams models were tested, however 6-gram model performed much better in all of the experiments and will be used to present the results.

It is clear from Tables 4 and 5 that the tokenization made the results slightly better than the baseline. The experiments reveal the interesting thing that addition of the context does not help the bioactivity prediction. It may be due to a fact that during a ligand to target protein connection, branches are much more important than chemical compounds center. In above approach however, it is centers of a compound that are being multiplied and result in adding noise to the representation. We also observed a slight deterioration if the classification results when the short paths or the simplification of SMILES was considered. On the other hand, the creation of new SMILES worked well in case of active and ZINC compounds. With active and inactive compounds, the results are not so clear but in general it is the best approach developed.

Table 6: Representations for 5-HT1a receptor

				nonzeros	
		dataset size	length	mean	density
active + inactive	3-gram	5433	1975	99,71	0,050
	6-gram	5433	43562	308,51	0,007
	KRFP	5433	4860	64,08	0,013
active + ZINC	3-gram	43760	2599	103,72	0,040
	6-gram	43760	103933	318,19	0,003
	KRFP	43760	4860	62,38	0,013

4.3 Bag-of-n-grams

In order to enable natural language processing with standard methods of machine learning, one has to create a vector representation. The simplest representation of this kind is called bag-of-words and its idea is to create a vector of length N where every position shows a number of occurrence of a word from a dictionary. Dictionary can be extracted from a training set. This representation can be extended to the case of n-grams (bag-of-n-grams), where every position of a vector corresponds to occurrence of a given n-gram. This representation will be used in the present experiment.

Modifications introduced previously should also have a positive effect on a bag-of-n-grams representation build upon modified and tokenized SMILES. A popular Klekota-Roth fingerprint (KRFP) will be used to compare the results [8]. It is based on counting substructures which were designed to increase the bioactivity information. Table 6 shows statistics of created representations. KRFP has a constant length of 4860 features while between 3 and 6-gram representations one can observe a huge difference due to an exponential grow of n-grams. However, our representation scales well with the amount of data and for active and ZINC compounds, which set is about 8 times bigger than active and inactive compounds, the representation is only about twice longer.

The classification of vector representations was then solved with Support Vector Machine (SVM) [2] which is a reliable model of the machine learning. When using SVM, one has to choose a kernel and the most important parameter of SVM which is the regularization value C. The kernel chosen is radial basis function, while regularization was fitted with cross-validation method after performing features standardization by removing the mean and scaling to unit variance. The parameter space searched was 10^x for $x \in \{-3, -2, \cdots, 4\}$ and for all the data sets and representations, the best C value equals 10^3. Because the classification problem comes from mapping the activation function into a set of labels {-1, 1}, one can also treat this problem as a regression. Both approaches were tested with a use of SVM implementation from scikit-learn library [10]. Answers $\hat{y}_i$ of regression models are mapped into the set of labels with a function:

$$f(\hat{y}) = \begin{cases} 0 \text{ if } \hat{y} \leq 0.5 \\ 1 \text{ if } \hat{y} > 0.5 \end{cases}$$

Table 7: Mean MCC score of vector classification for active and ZINC compounds.

		5-HT1a	5-HT6	5-HT7	5-HT2a	5-HT2b	5-HT2c
	6-gram model	0.960	0.974	0.952	0.952	0.912	0.943
6-gram	SVM cls.	0.983	0.987	**0.977**	**0.975**	0.944	**0.970**
	SVM rgs.	**0.984**	**0.988**	**0.977**	**0.975**	**0.947**	**0.970**
KRFP	SVM cls.	0.971	0.975	0.956	0.953	0.892	0.943
	SVM rgs.	0.969	0.946	0.955	0.913	0.887	0.918

Table 8: Mean MCC score of vector classification for active and inactive compounds.

		5-HT1a	5-HT6	5-HT7	5-HT2a	5-HT2b	5-HT2c
	6-gram model	0.621	0.749	0.666	0.657	0.559	0.697
6-gram	SVM cls.	0.689	**0.780**	0.701	0.711	**0.589**	**0.703**
	SVM rgs.	**0.695**	0.774	0.686	0.695	0.576	0.697
KRFP	SVM cls.	0.655	0.726	0.733	0.722	0.529	0.677
	SVM rgs.	0.618	0.741	**0.744**	**0.727**	0.529	0.678

Tables 7 and 8 show results of classification for sets containing inactive and ZINC compounds respectively. The 'cls.' and 'rgs.' annotations stand for classification and regression approaches. First of all, the bag-of-n-grams classification gives much better results than n-gram model. SVM regression approach works better than the classification. In general, it achieves better results with 6-gram representation than with KRFP fingerprint (two exceptions are 5-HT7 and 5-HT2a receptors in the case of actives-inactives separation).

To have more detailed analysis of the classification, the results for receptors 5-HT1a and 5-HT6 (actives and inactives) are shown in Figure 2. For all of the classification models, 6-gram representation works better than KRFP and allows the SVM regression to achieve a high mean MCC scores equal to 0.7 and 0.78.

5 Conclusion

In this paper, the application of NLP methods to bioactivity prediction was presented. The best modifications were chosen during experimental analysis which shows that virtual screening with a use of textual representation was successful.

In case of the n-gram model, a general tokenization approach and new SMILES generation algorithm were designed. This lead to a creation of bag-of-n-grams representation which is a big success of this work. It allows SVM to achieve higher scores than when using KRFP . Considering the fact that in comparison to KRFP in order to create a bag-of-n-grams representation no chemical domain knowledge is required, it proves high capabilities of this representation.

Open door are also left for further study on SMILES. One of the biggest improvements was new SMILES generation. It would be definitely worth to find

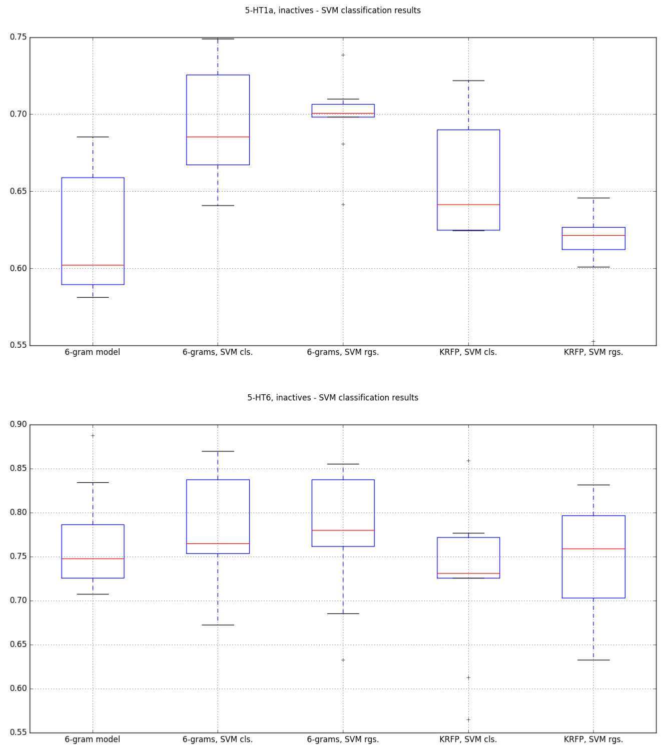


Fig. 2: Classification results of active and inactive compounds for 5-HT1a and 5-HT6, respectively

an optimum between results improvement and time complexity due to a new SMILES generation. In case of bag-of-n-grams representation, it is a common approach in NLP to reduce its dimensionality by removing from a dictionary words which occur in too many or too few documents, as they may be not relevant to a problem. Similar techniques based on n-grams frequency could be applied here which would result in reducing the dimensionality of the representation. Lastly, satisfactory results of bag-of-n-grams representation and n-gram model which works directly on SMILES representation lead to a conclusion that learning compounds representation directly from SMILES with deep learning methods could be a promising approach to undertake.

Acknowledgement

This research was partially supported by the National Science Centre (Poland) grant no. 2014/13/N/ST6/01832 and grant no. 2014/13/B/ST6/01792.

References

1. Brown, P.F.: Class-based n-gram models of natural language. Computational Linguistics 18, 467–479 (1992)
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273297 (1995)
3. Daylight Chemical Information Systems: Daylight (2008), `http://www.daylight.com`
4. Gaulton, A.: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research 40, D1100–D1107 (2016)
5. Heafield, K.: KenLM: faster and smaller language model queries. WMT '11 Proceedings of the Sixth Workshop on Statistical Machine Translation pp. 187–197 (2011)
6. Irwin, J.J., Shoichet, B.K.: ZINC - A free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. 45, 177–182 (2005)
7. Jurafsky, D., Martin, J.H.: Speech and Language Processing. Prentice Hall, Upper Saddle River, New Jersey 07458, 2nd edition edn. (2008)
8. Klekota, J., Roth, F.P.: Chemical substructures that enrich for biological activity. Bioinformatics 24, 2518–2525 (2008)
9. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405, 442–451 (1975)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
11. Raevsky, O.A.: Molecular structure descriptors in the computer-aided design of biologically active compounds. Russ. Chem. Rev. 68, 505–524 (1999)
12. Shoichet, B.K.: Virtual screening of chemical libraries. Nature 432, 862865 (2004)
13. Sousa, S.F., Fernandes, P.A., Ramos, M.J.: Proteinligand docking: Current status and future challenges. Proteins: Structure, Function and Bioinformatics 65, 15–26 (2006)
14. Vidal, D., Thormann, M., Pons, M.: LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J. Chem. Inf. Model. 45, 386393 (2005)
15. Warszycki, D., Mordalski, S., Kristiansen, K., Kafel, R.and Sylte, I.C.Z., Bojarski, A.J.: A linear combination of pharmacophore hypotheses as a new tool in search of new active compounds–an application for 5-HT1A receptor ligands. PloS ONE 8(12), e84510 (2013)
16. Worachartcheewan, A., Mandi, P., Prachayasittikul, V., Toropova, A., Toropov, A., Nantasenamat, C.: Large-scale QSAR study of aromatase inhibitors using SMILES-based descriptors. Chemometrics and Intelligent Laboratory Systems 138, 120–126 (2014)
17. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 334–342. ACM (2001)