Классификация протоколов тестирования биологической активности для моделирования взаимосвязей «структура-активность»

Ольга Тарасова

Институт Биомедицинской химии

Москва

XXIV Российский национальный конгресс
«Человек и лекарство»

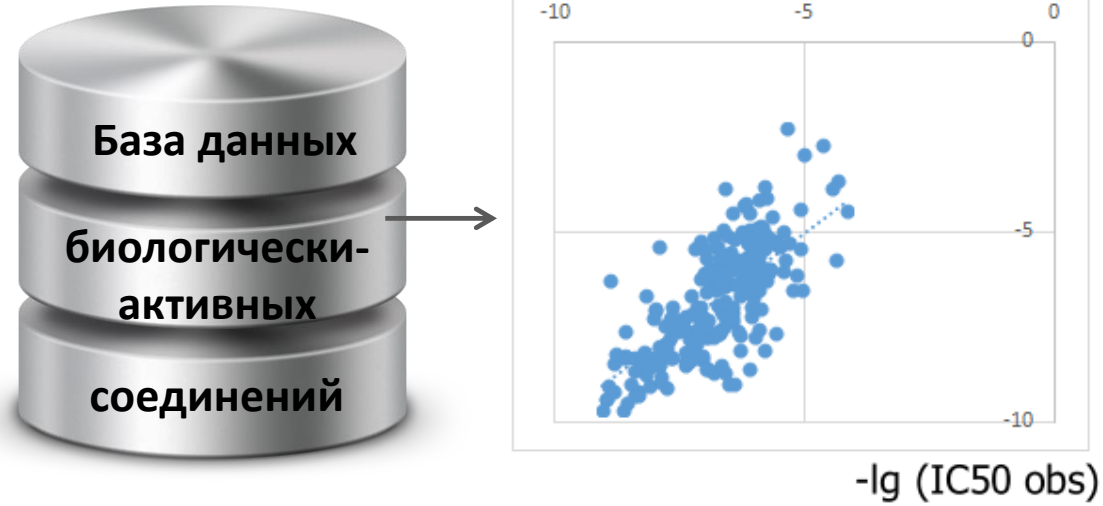
Моделирование взаимосвязей «структура-активность»

Идеальные данные

"…high-quality data are those that have been obtained from the same experimental protocol, ideally in the same laboratory by the same workers <…>, for a standardized and pertinent endpoint."

R. Benigni. "Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens CRC Press, 2003, 304 P.

Реальность



D. Fourches et al., *J Chem Inf Model,* 2010, 50(7): 1189–1204.

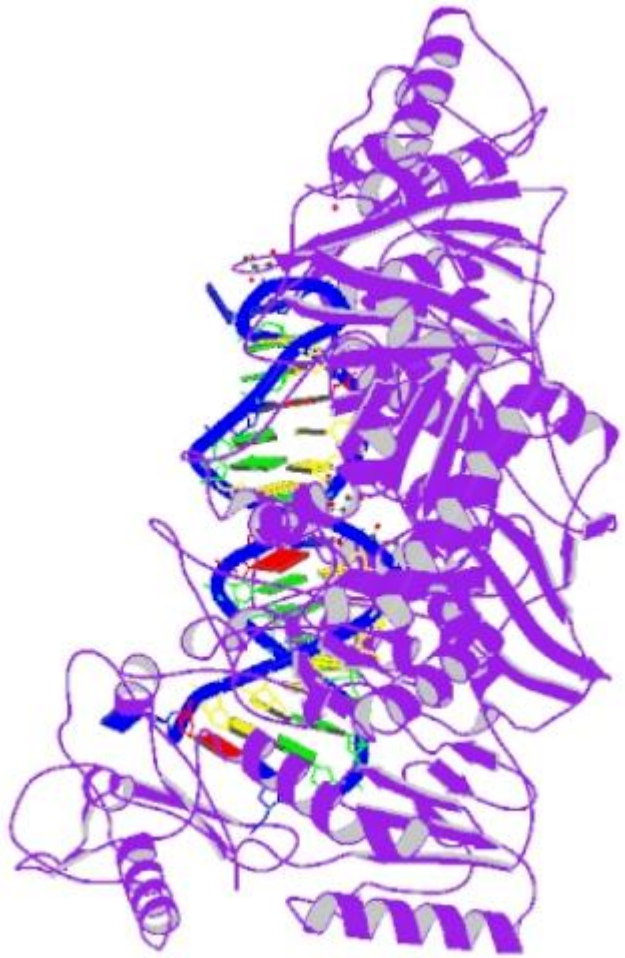
D. Fourches et al., *J Chem Inf Model,* 2016, 56(7): 1243-52.

A.V. Zakharov, *ACS meeting*, Fall 2015

U. Visser et al., *BMC Bioinformatics*, 24;12:257,2011

Цель работы: разработка и тестирование алгоритма классификации протоколов тестирования органических соединений на основе анализа текстов научных публикаций.

Обратная транскриптаза (ОТ) ВИЧ-1 – объект исследования



Основная мишень механизма действия антиретровирусных соединений (в том числе – составляющих ВААРТ);

Потребность в разработке новых антиретровирусных веществ*

Большое число данных о биологической активности, полученных в результате множества биологических тестов;

Ранее исследованы возможности агрегирования данных о биологической активности из различных баз данных (включая ChEMBL, Integrity)**

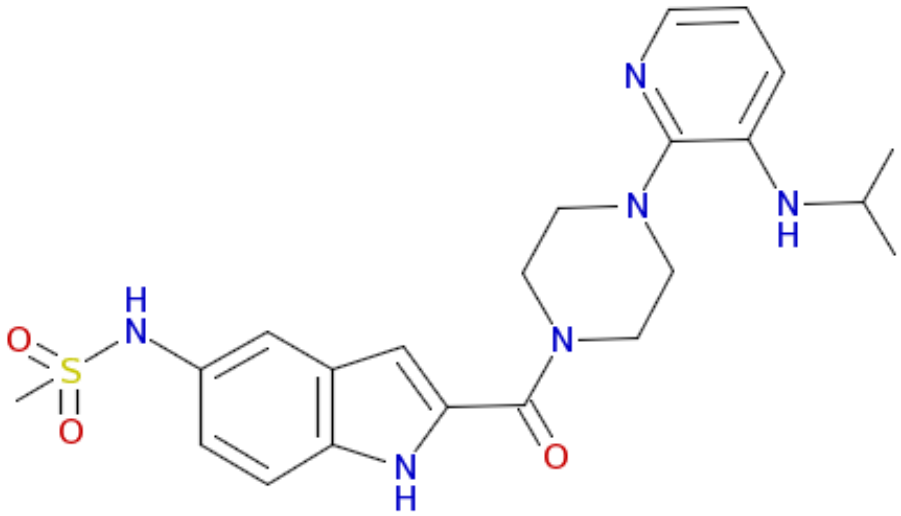
K. Das et al., *Acs Chem.Biol.* , 2016, 11 (8): 2158–2164; PDB: **5I3U**

* *W.L. Jorgensen. Bioorg Med Chem.,* 2016 Jul 21. pii: S0968-0896(16)30546-6. doi: 10.1016/j.bmc.2016.07.039.

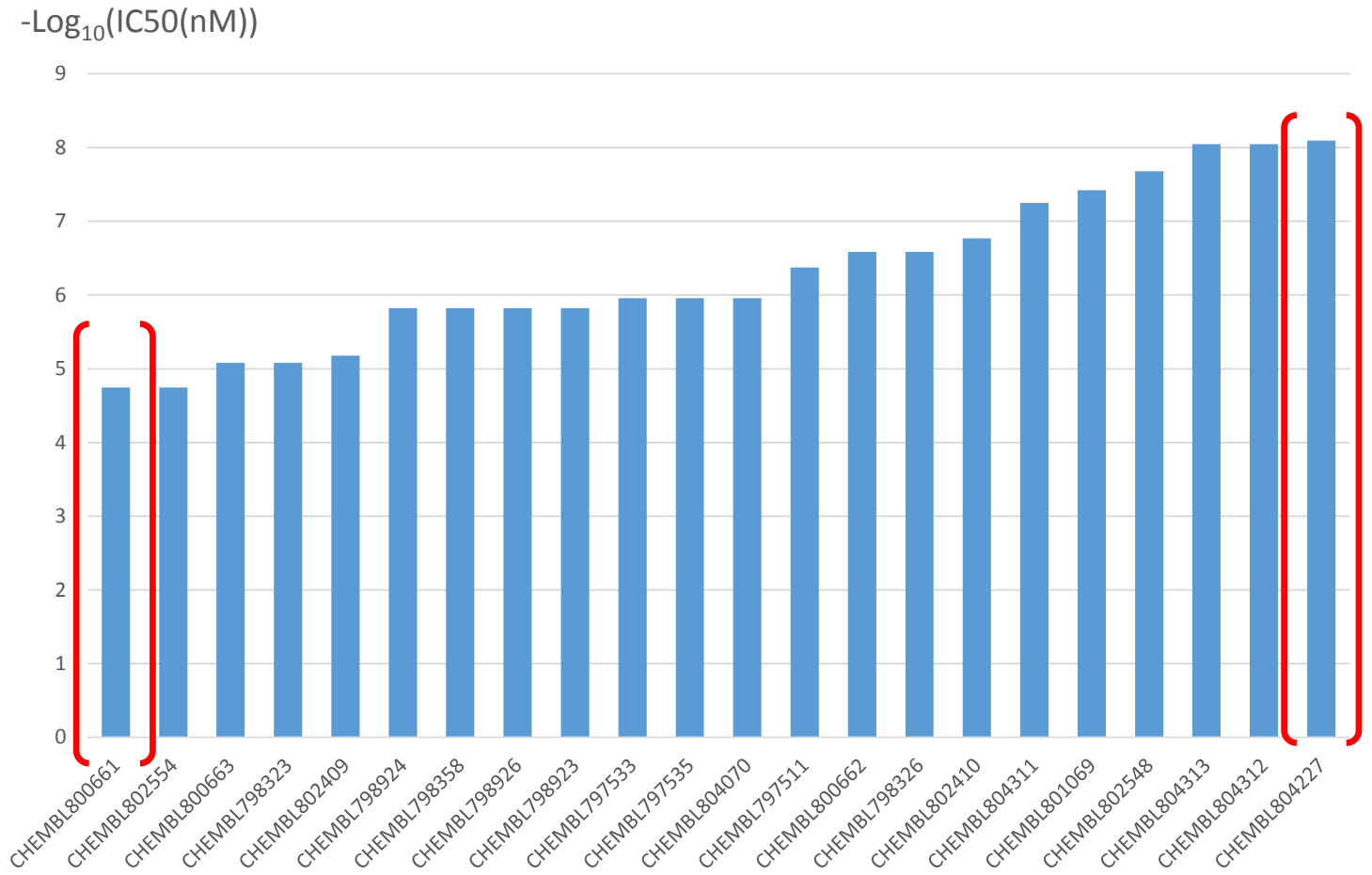
**Tarasova, O. et al., *J Chem Inf Model*, 2015 Jul 27;55(7):1388-1399.

Распределение значений $Log_{10}(IC_{50})$ в зависимости от деталей эксперимента (ChEMBL v.21)

Delavirdine



CHEMBL593



IC_{50}(max) = 18 µM

IC_{50}(min) = 8nM;

Автоматизированный анализ текста в биомедицине

Алгоритмы анализа текста

- Определение онтологии генов [1-5]
- Определение функции белков [3-5]
- Идентификация белок-белковых взаимодействий [7]
- Определение взаимосвязей между генами, белками и заболеваниями [6-8]

[1] J. Finkel et al., *BMC Bioinformatics 2005*, 6 Suppl 1:S5.
[2] J. Hakenberg et al., *BMC Bioinformatics 2005*, 6 Suppl 1:S9.
[3] S. Kinoshita et al., *BMC Bioinformatics 2005*, 6 Suppl 1:S4.
[4] R. McDonald, F. Pereira, *BMC Bioinformatics 2005*, 6 Suppl 1:S6.
[5] T. Mitsumori et al., *BMC Bioinformatics 200*5, 6 Suppl 1:S8.
[6] G. Lee et al., *PLoS One. 2016* Jul 15;11(7):e0159088.
[7] J. Evans, A. Rzhetsky. J Biol Chem. 2011,286(27):23659-66.
[8] L. Yao et al., *Trends Biotechnol.* 2010 Apr; 28(4): 161–170.
[9] P. Thomas, Bioinformatics. 2015 Apr 15;31(8):1258-66[10]

Whether text mining can help us to identify similar assay protocols from the scientific publications ???

План исследования

Автоматизированный отбор текстов научных публикаций

- Полные тексты публикаций
- Резюме публикаций

1

Автоматическое извлечение фрагментов текста с описанием эксперимента

- С применением полных текстов публикаций

2

Автоматическое сравнение фрагментов текста

- С применением фрагментов текста публикаций

3

Выбор релевантных публикаций

- Описание новых молекул, активных против обратной транскриптазы ВИЧ-1
- Детальные описания биологических экспериментов

ScienceDirect;
Elseivier;
PubMed

⇨ (“HIV-1” **and** “REVERSE TRANSCRIPTASE” **and** “INHIBITORS”)

Bioorganic & Medicinal Chemistry 17 (2009) 3980–3986

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Design, microwave-assisted synthesis and HIV-RT inhibitory activity of 2-(2,6-dihalophenyl)-3-(4,6-dimethyl-5-(un)substituted-pyrimidin-2-yl)-thiazolidin-4-ones

Hua Chen *, Jie Bai, Lingling Jiao, Zaihong Guo, Qingmei Yin, Xiaoliu Li *

Key Laboratory of Chemical Biology of Hebei Province, Key Laboratory of Medicinal Chemistry and Molecular Diagnosis, Ministry of Education, College of Chemistry and Environmental Science, Hebei University, Baoding 071002, China

ARTICLE INFO

Article history:
Received 21 January 2009
Revised 4 April 2009
Accepted 6 April 2009
Available online 18 April 2009

Keywords:
Thiazolidin-4-one
Microwave-assisted
Synthesis
Anti-HIV-RT activity

ABSTRACT

A series of novel thiazolidin-4-ones bearing a hydrophobic substituent at 5-position on the 4,6-dimethylpyrimidine ring at N-3 (**5c–i** and **6c–i**) were designed on the prediction of QSAR studies, synthesized in good yields of 60.1–85.3% by microwave-assisted one-pot protocol with the combination of using dicyclohexylcarbonimide (DCC) as the promotor, and evaluated as HIV-1 reverse transcriptases inhibitors. The results of in vitro HIV-1 RT kit assay showed that some of the new compounds, such as **5c**, **6c**, **5d**, **6d**, **5g**, **5h** and **6i**, could effectively inhibit RT activity. Among them, compounds **5c** and **6c** where ethyl group existed at 5-position on N-3 pyrimidine ring were the best ones with the IC_{50} value of 0.26 μM and 0.23 μM, respectively. Structure-activity relationship analysis of these analogues suggested that the overall hydrophobicity and steric factor were important to the anti-HIV RT activity. The mechanism of the intramolecular cycloamidation promoted by DCC was also investigated with the key uncyclized intermediate **13**.

Пример публикации

Нерелевантные публикации

- Содержат описание соединений которые исследованы НЕ на активность против обратной транскриптазы ВИЧ-1

ИЛИ

- Отсутствует детальное описание биологических экспериментов

Пример публикации

Non-nucleoside reverse transcriptase inhibitors emerging as an attractive alternative to protease inhibitors in HIV combination therapy regimens

Antiretroviral therapy aims to suppress, and maintain the suppression of, HIV replication in infected individuals. Triple combinations of antiretroviral agents are accepted as being the minimum 'standard-of-care' regimens to achieve this goal. Recommended initial starting regimens usually consist of two nucleoside reverse transcriptase inhibitors (NRTIs) plus either a protease inhibitor (PI), a non-nucleoside reverse transcriptase inhibitor (NNRTI) or possibly a third NRTI.

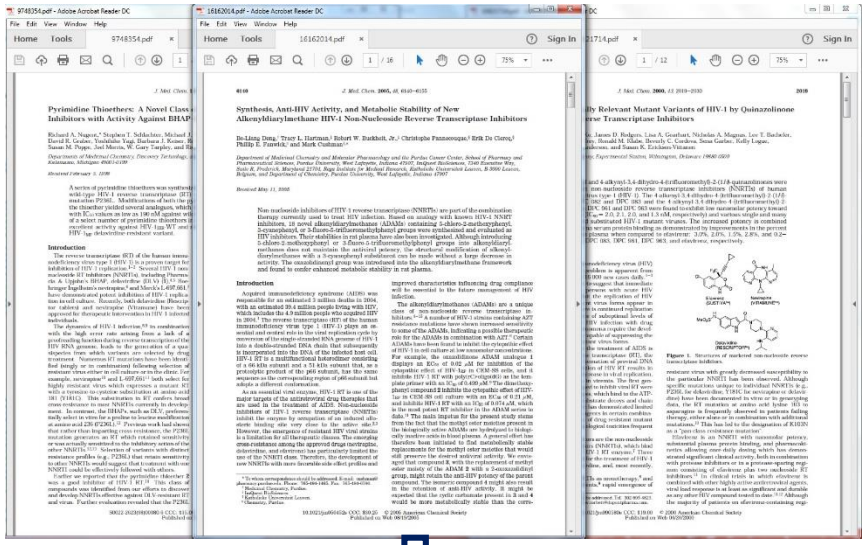
Three NNRTIs (efavirenz, nevirapine and delavirdine) are currently available for inclusion in NNRTI-containing regimens but delavirdine has not yet received approval in the European Union.

PI-combination regimens not ideal

Although protease inhibitors (in combination with two NRTIs) have demonstrated improved survival and reduced disease progression compared with single or double agent therapy, they are not ideal.[1] In particular, the recently recognised metabolic abnormalities (e.g. lipodystrophy syndrome, raised cholesterol, etc.) appear to be most common with protease inhibitor therapy [see article entitled: 'The aetiology of antiretroviral-associated lipodystrophy remains elusive posing problems for its prevention and treatment' *Drugs & Therapy Perspectives 2001 Jul 2;*

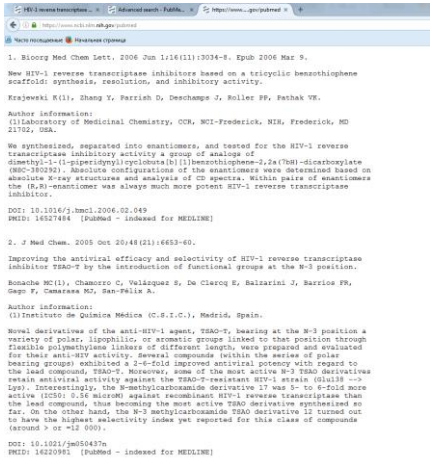
Vol. 17, No. 19; September 24, 2001

1. Автоматизированный выбор релевантных публикаций



Создание классификатора

Полнотекстовые публикации



Резюме

TET PDFLib v.5.0 (PDFlib GmbH)

Преобразование PDF в текст

32 релевантные + 31 нерелевантная статьи

Автоматизированная обработка текста

LingPipe v. 4.1.2*

Классификатор

Обучающая выборка (66.7%)
21 релевантная + 21 нерелевантная статьи

Тестовая выборка(33.3%)
11 релевантных+ 10 нерелевантных

Результаты

*B. Carpenter. *Proceedings of the 2nd BioCreative workshop*. Valencia, Spain, 2007.

1. Отбор релевантных публикаций. Анализ результатов классификации

	Полнотекстовые статьи	Резюме
TP:	10	9
TN:	9	9
FP:	1	1
FN:	1	2
Sens:	90.0	81.7
Spec:	91.0	90.0
BA:	90.5	85.5

Классификация на основе резюме публикаций может быть применена для автоматизированного выбора релевантных публикаций

Применение классификатора для автоматизированного отбора

Анализ названий релевантных публикаций, формирование запроса

⇩

NCBI PubMed (запрос):
“HIV-1 reverse transcriptase inhibitors and design and (evaluation OR activity)”

⇩

Более 240 резюме публикаций

⇩

Автоматический отбор релевантных публикаций

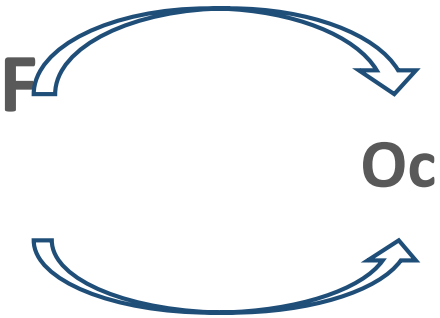
⇩

Загрузка публикаций для обработки

⇩

72 текста публикаций в формате PDF

32 релевантных публикации
из обучающей выборки



Основная выборка

Полный текст публикаций для анализа

Резюме:

Описание биологических экспериментов недостаточно детализировано или отсутствует вообще

CHEMMEDCHEM

DOI: 10.1002/cmdc.200500020

Design, Synthesis, Biological Evaluation, and Molecular Modeling Studies of TIBO-Like Cyclic Sulfones as Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors

Roberto Di Santo,*[a] Roberta Costi,[a] Marino Artico,[a] Rino Ragno,[a] Antonio Lavecchia,[b] Ettore Novellino,[b] Enrico Gavuzzo,[c] Francesco La Torre,[d] Roberto Cirilli,[d] Reynel Cancio,[e] and Giovanni Maga[e]

*TIBO- and TBO-like sulfone derivatives **1** and **2** were designed, synthesized, and tested for their ability to block the replication cycle of HIV-1 in infected cells. The anti-HIV-1 activities of sulfones **3**, which were intermediates in the syntheses of **1** and **2**, were also evaluated. Surprisingly, the sulfone analogues of TIBO R82913 (compounds **1**) were inactive, whereas interesting results were obtained for truncated derivatives **2**. Compound **2w** was the most potent among this series in cell-based assays (EC_{50} = 0.07 μM, CC_{50} > 200 μM, SI > 2857). It was twofold less potent than R82913, but more selective. An X-ray crystallographic analysis was carried out to establish the absolute configuration of **2w** and its enantiomer **2x**, which were obtained by semipreparative HPLC of **2v**, one of the most potent racemates. Compounds **1–3** were proven to target HIV-1 RT. In fact, representative derivatives inhibited recombinant HIV-1 RT in vitro at concentrations similar to those active in cell-based assays. 3D QSAR studies and docking simulations were developed on TIBO- and TBO-like sulfone derivatives to rationalize their anti-HIV-1 potencies and to predict the activity of novel untested sulfone derivatives. Predictive 3D QSAR models were obtained with a receptor-based alignment by docking of TIBO- and TBO-like derivatives into the NNBS of RT.*

R. Di Santo et al., *ChemMedChem, 2006*

Полный текст публикаций для анализа

Описание биологических экспериментов приведено и максимально детализировано

TET PDFLib v.5.0 (PDFlib GmbH)

Преобразование PDF в текстовый формат

Cells and viruses: MT-4 cells were grown at 37 °C in a CO_2 atmosphere (5%) in RPMI 1640 medium, supplemented with fetal calf serum (FCS, 10%), penicillin G (100 IU mL^{-1}), and streptomycin (100 μg mL^{-1}). Cell cultures were verified periodically for the absence of mycoplasma contamination with a MycoTect Kit (Gibco). Human immunodeficiency virus type 1 (HIV-1, IIIB strain) was obtained from supernatants of persistently infected cells. HIV-1 stock solutions had titers of 1.5×10^7 mL^{-1} for a 50% cell culture infectious dose ($CCID_{50}$).

HIV titration: Titration of HIV was performed in C8166 cells by the standard limiting dilution method (dilution 1:2, four replica wells per dilution) in 96-well plates. The infectious virus titer was determined by light-microscope scoring of syncytia after four days of incubation. Virus titers were expressed as $CCID_{50}$ per mL.

Anti-HIV assays: The activity of the compounds against multiplication of wild-type HIV-1 in acutely infected cells was based on the inhibition of virus-induced cytopathicity in MT-4 cells. Briefly, culture medium (50 μL) containing 1×10^4 cells was added to the wells of flat-bottom microtiter trays containing 50 μL of culture medium with or without various concentrations of test compounds. Then, HIV suspensions (20 μL, each containing the appropriate amount (by $CCID_{50}$) to cause complete cytopathicity at day 4) were added. After incubation at 37 °C, cell viability was determined by the 3-(4,5-dimethylthiazol-1-yl)-2,5-diphenyltetrazolium bromide (MTT) method.[51] The cytotoxicity of each test compound was evaluated in parallel with its antiviral activity and was based on the viability of mock-infected cells, as monitored by the MTT method.

RT assays: Assays were performed as previously described.[52] Briefly, purified rRT (20–50 nM) was assayed for its RNA-dependent DNA polymerase activity in a volume of 25 μL containing Tris-HCl (50 mM, pH 7.8), KCl (80 mM), $MgCl_2$ (6 mM), DTT (1 mM), BSA (0.1 mg mL^{-1}), template–primer duplex: [poly(rA)–oligo$(dT)_{12-18}$] (0.3 μM, determined by 3′-OH end concentration), and [^{3}H]dTTP (10 μM, 1 Ci mmole^{-1}). After a incubation at 37 °C for 30 min, samples were spotted onto glass fiber filters (Whatman GF/C) and the acid-insoluble radioactivity was determined.

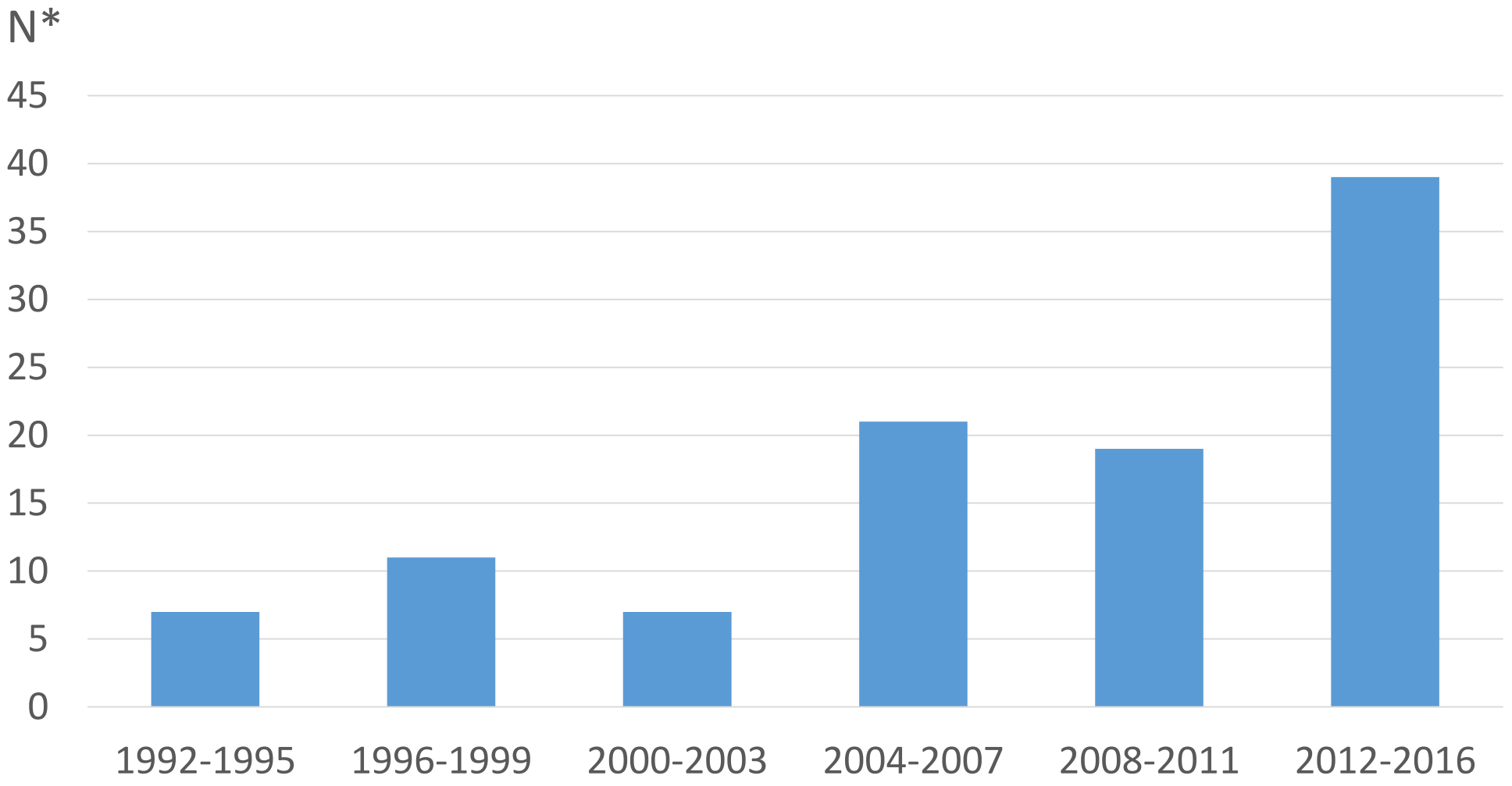
and electrostatic interactions. An initial random velocity was applied to all atoms corresponding to 300 K. Three subsequent MD runs were then performed. The first was carried out for 10 ps with a time-step of 1.5 fs at a constant temperature of 300 K for equilibration purposes. The next MD was carried out for 20 ps, during which time the system was coupled to a thermal bath (150 K) with a time constant of 5 ps. The time constant represents approximately the half-life for equilibration with the bath; consequently, the second MD command caused the molecule to slowly cool to ≈150 K. The third and last MD cooled the molecule to 50 K over 20 ps. A final energy minimization was then carried out for 250 iterations with a conjugate gradient. Similarly to the previous minimization, during all the MD procedures described above, only the residues of a core 3 Å from the ligands were allowed to relax. Energy minimizations and MD were performed in vacuo in all cases. Again, after the MD procedure was applied, all ligands were extracted while maintaining their absolute Cartesian coordinate positions. This afforded the second alignment rule (Dyn alignment).

The binding mode of TIBO- and TBO-like derivatives **1–3** was analyzed with a docking procedure by using the program AutoDock.[44] For the docking, a grid with spacing of 0.375 Å and with point dimensions of 60×80×60 was used. The grid was centered on the mass center of the experimentally bound R82913 coordinates. The Lamarckian genetic algorithm with local search (GA-LS) method was adopted using the default settings. Amber united-atom charges and solvation parameters were assigned to the protein with the program ADT (AutoDock Tools). AutoDock generated 200 possible binding conformations grouped in clusters, setting the RMSD tolerance to 1 Å. For docking assessment, the same docking protocol was used on the reference drug R82913. AutoDock successfully reproduced the bound conformation with a RMSD value of only 0.63.

Once docked based on the settings described above, derivatives **1–3** were receptor-based aligned by using the AutoDock program. To this, each molecule of the training set was docked into the NNBS. The starting conformations used for docking were those ex-

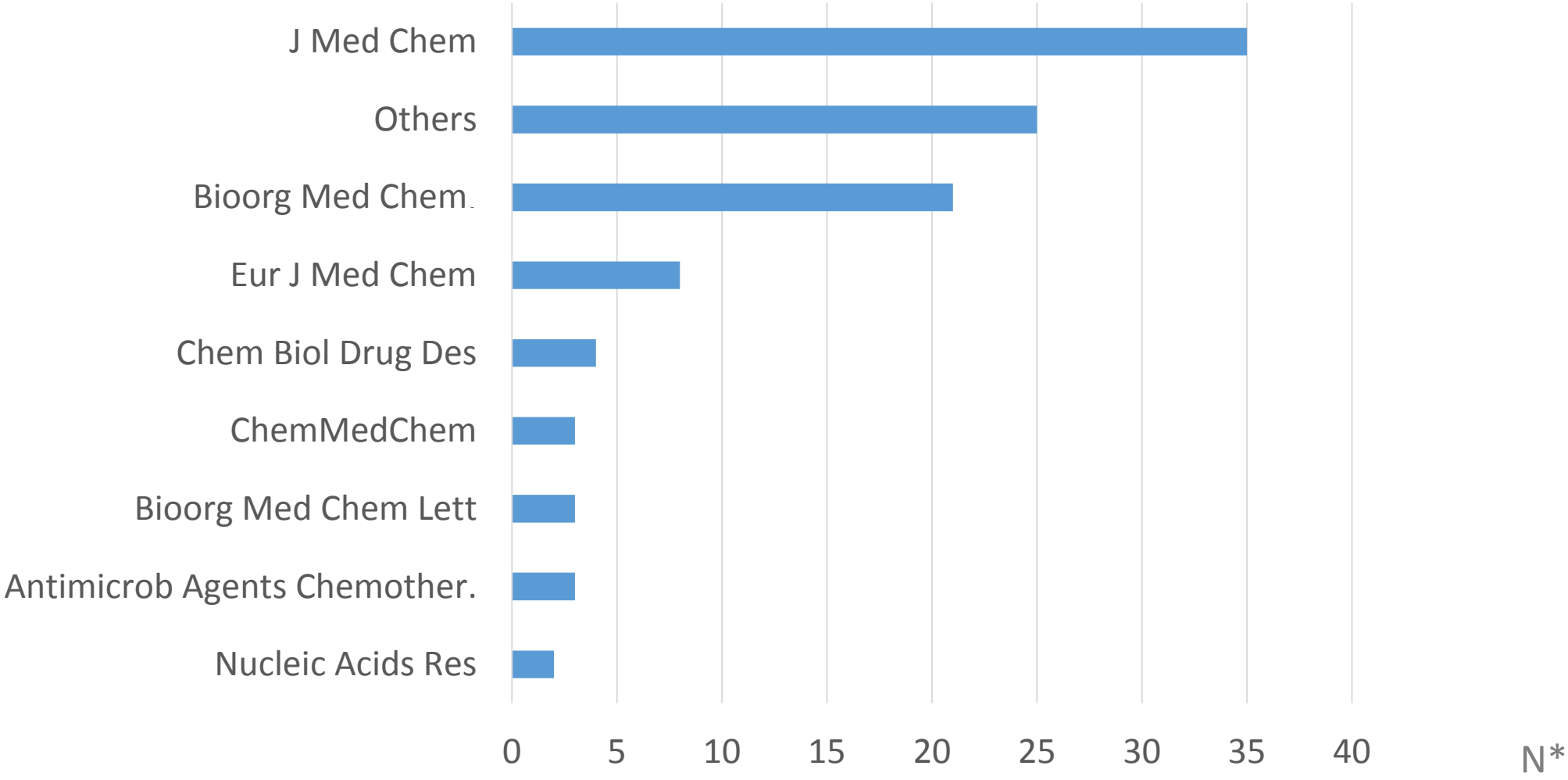
R. Di Santo et al., *ChemMedChem, 2006*

Основная выборка публикаций. Распределение по годам выхода



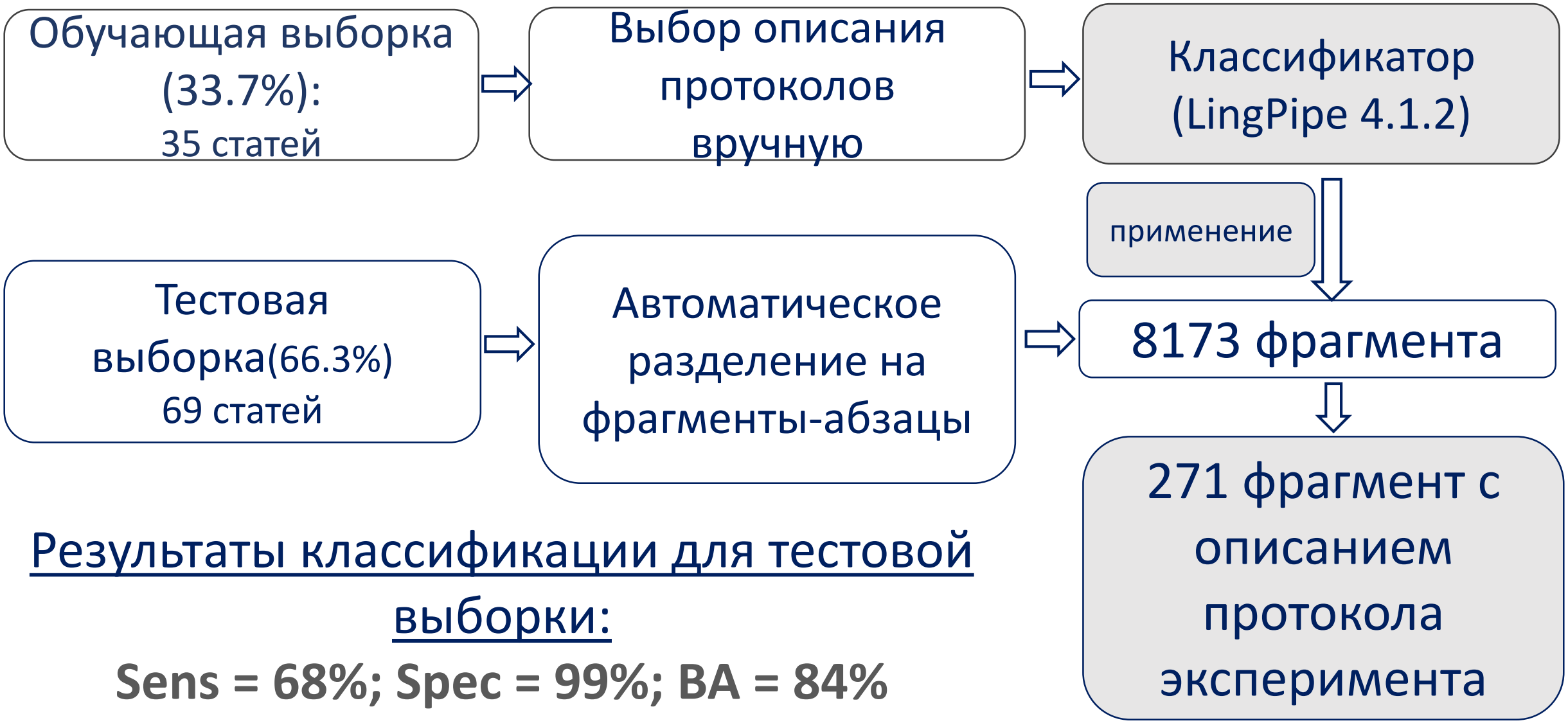
* Количество публикаций в указанном диапазоне

Основная выборка. Журналы



*Количество статей, опубликованных в данном журнале

2. Извлечение протоколов тестирования



Результаты классификации для тестовой выборки:

Sens = 68%; Spec = 99%; BA = 84%

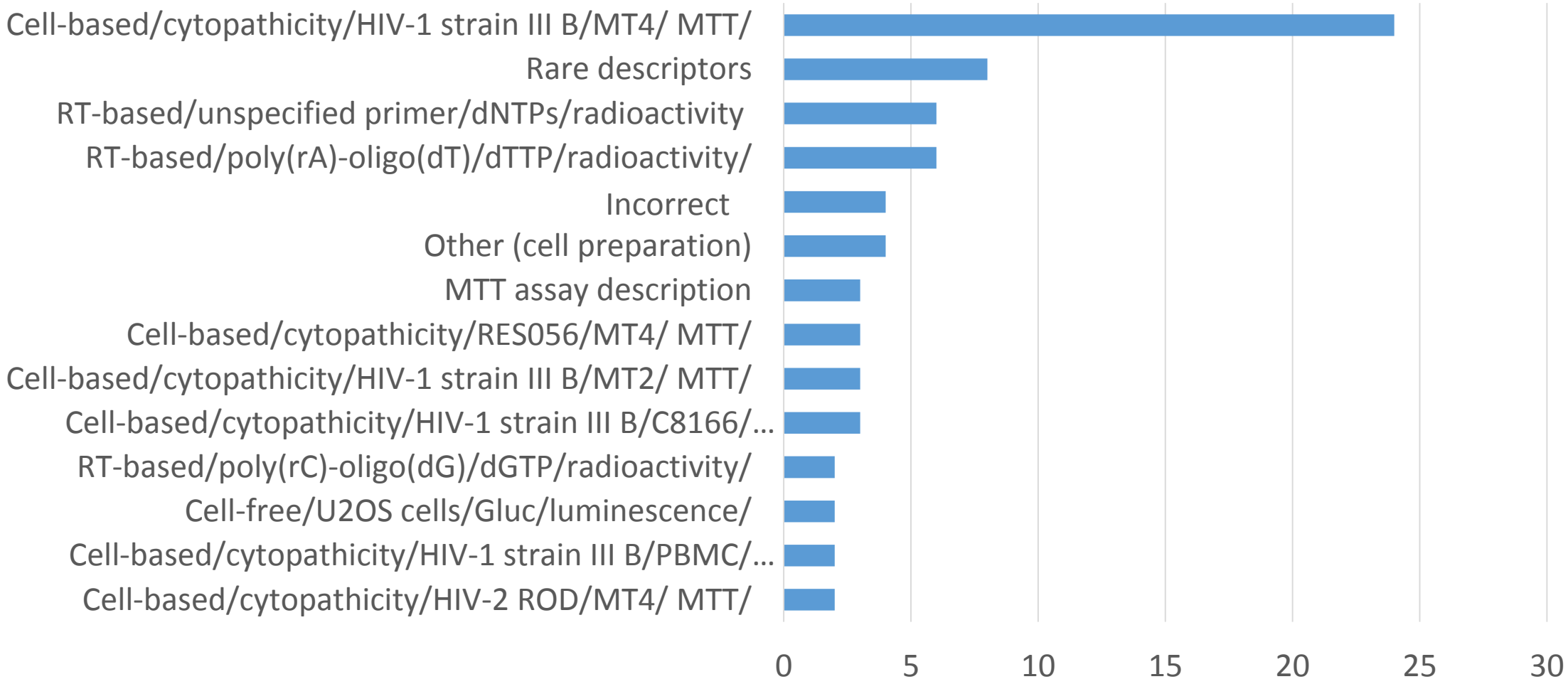
3. Сравнение протоколов экспериментального тестирования

3. Применение методов обучения без учителя

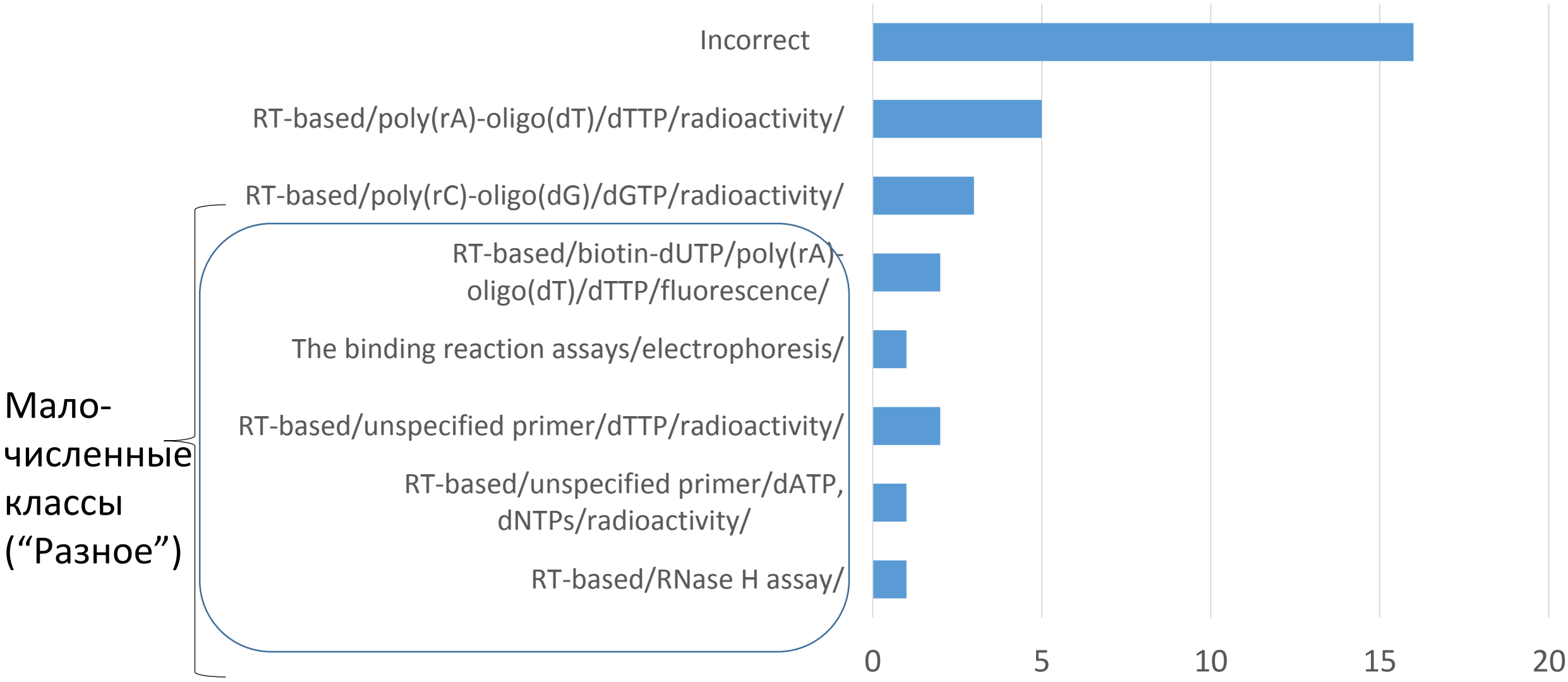
Weka 3.8.0. Кластеризация EM-алгоритм

Четыре кластера

Кластер 1.

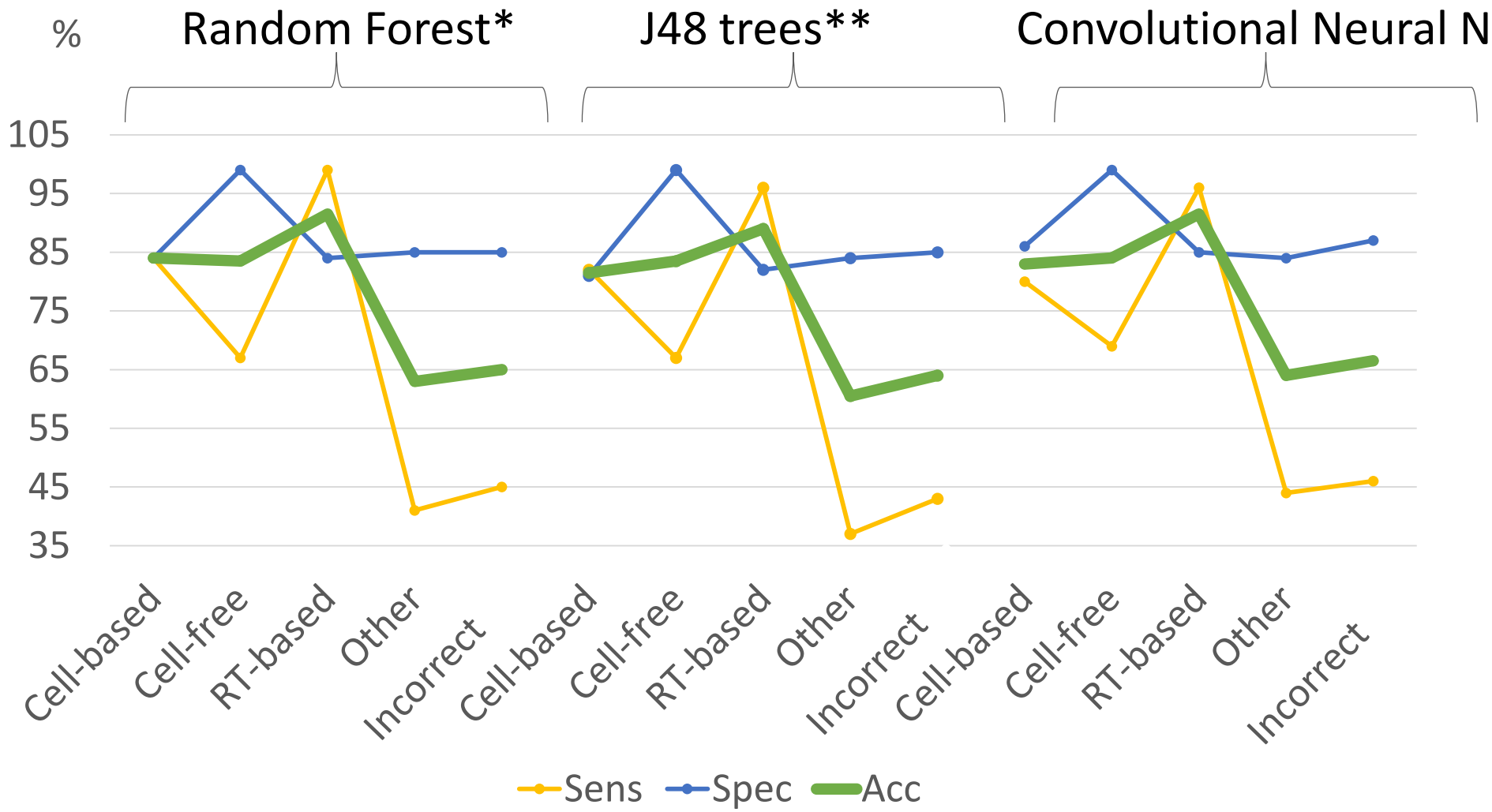


Кластер 2.


Incorrect
RT-based/poly(rA)-oligo(dT)/dTTP/radioactivity/
RT-based/poly(rC)-oligo(dG)/dGTP/radioactivity/
RT-based/biotin-dUTP/poly(rA)-oligo(dT)/dTTP/fluorescence/
The binding reaction assays/electrophoresis/
RT-based/unspecified primer/dTTP/radioactivity/
RT-based/unspecified primer/dATP, dNTPs/radioactivity/
RT-based/RNase H assay/
0
5
10
15
20
Мало-численные классы (“Разное”)

4. Применение методов машинного обучения

Weka 3.8.0. Разделение на группы “cell-based”/ “RT-based”/ “cell-free”, 20fold CV



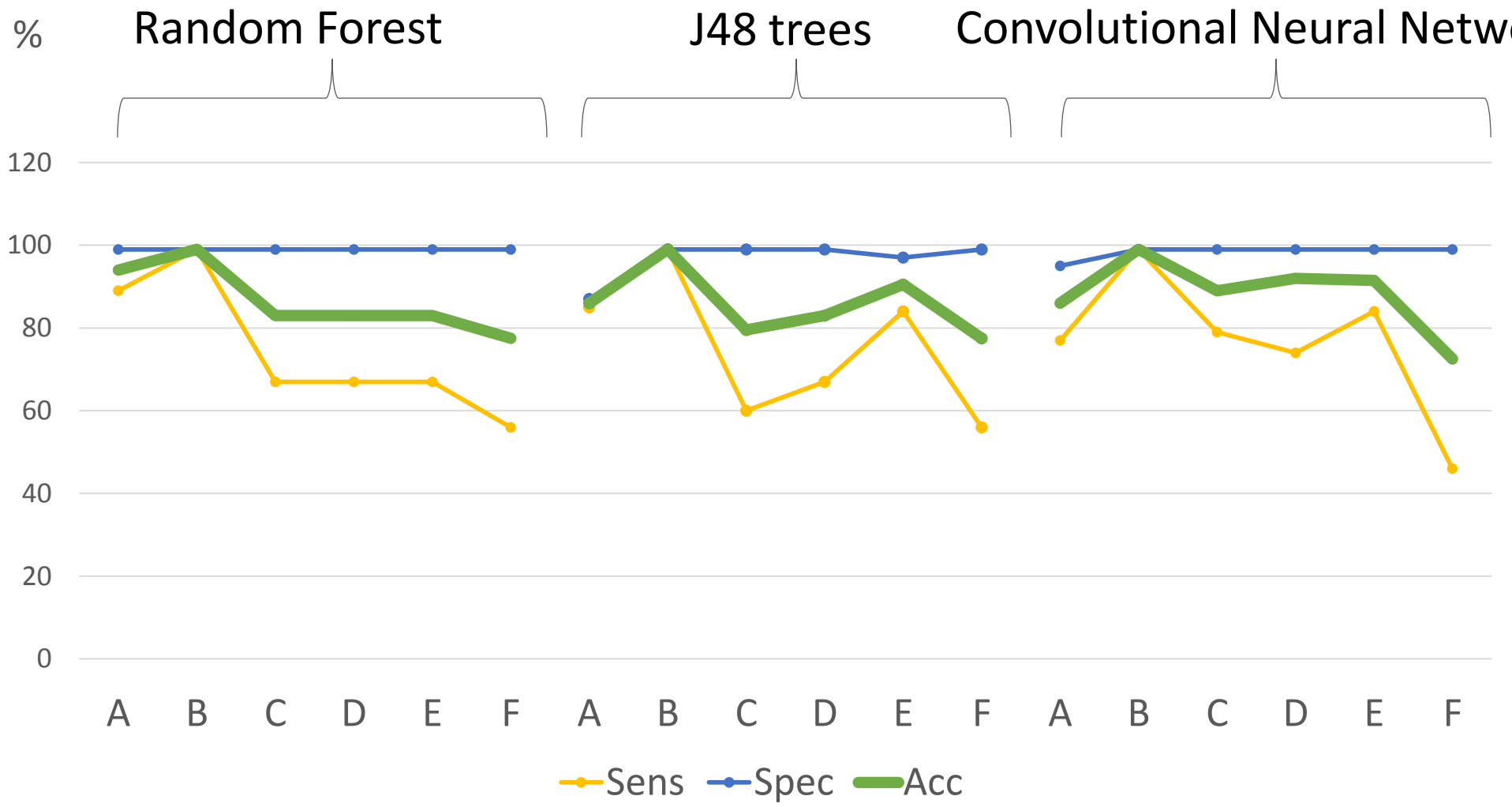
* L. Breiman. Random Forests. *Machine Learning*. **45**:5-32, 2001.

** R. Quinlan. In: *Programs for Machine Learning*, San Mateo, CA, 1993

*** https://github.com/amten/NeuralNetwork

4. Применение методов машинного обучения

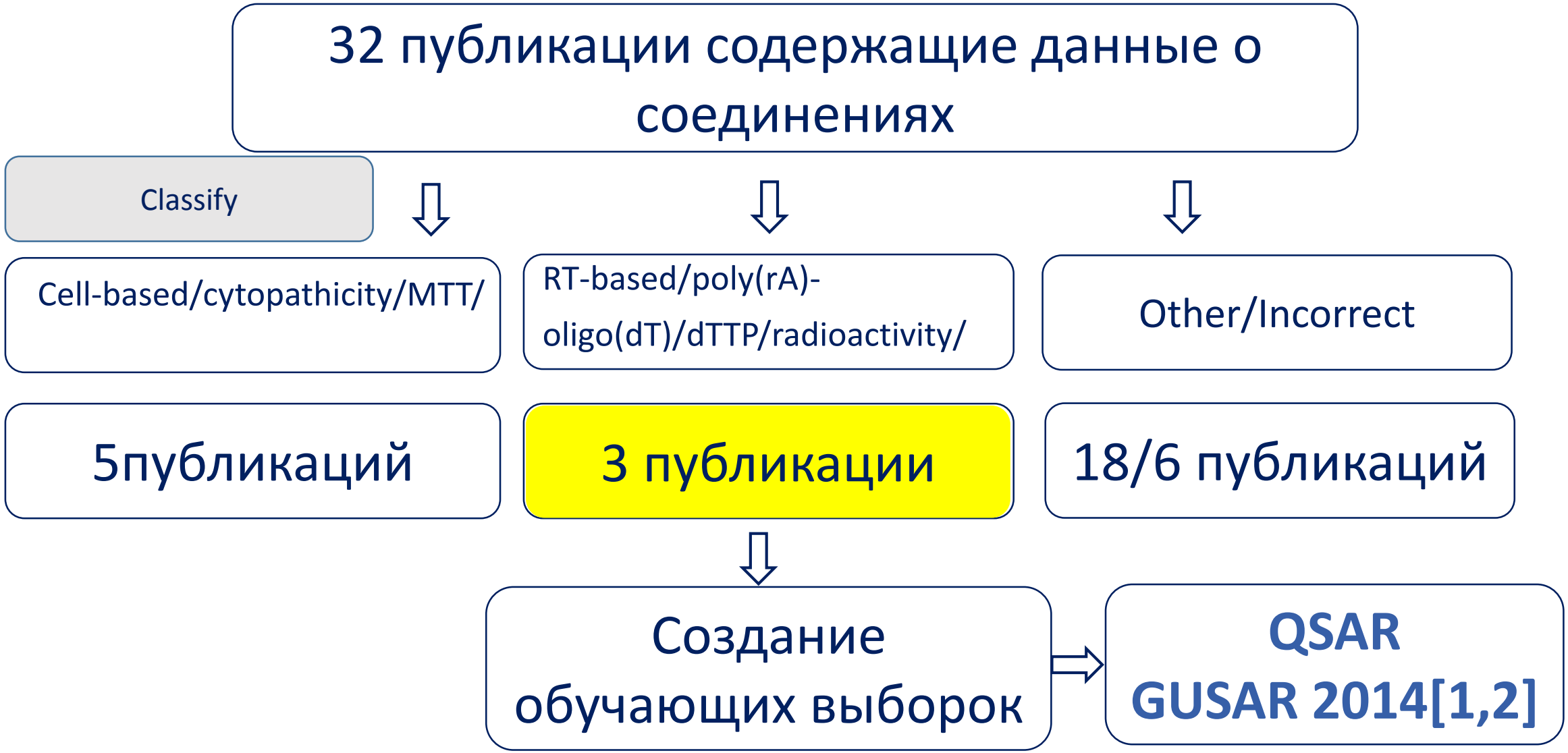
Weka 3.8.0. Разделение на группы с учетом детального описания, 20fCV



A	Cell-based/cytopathicity/MTT/	D	RT-based/unspecified primer/dNTPs/radioactivity/
B	RT-based/poly(rA)-oligo(dT)/fluorescence/	E	Other
C	RT-based/poly(rA)-oligo(dT)/dTTP/radioactivity/	F	Incorrect

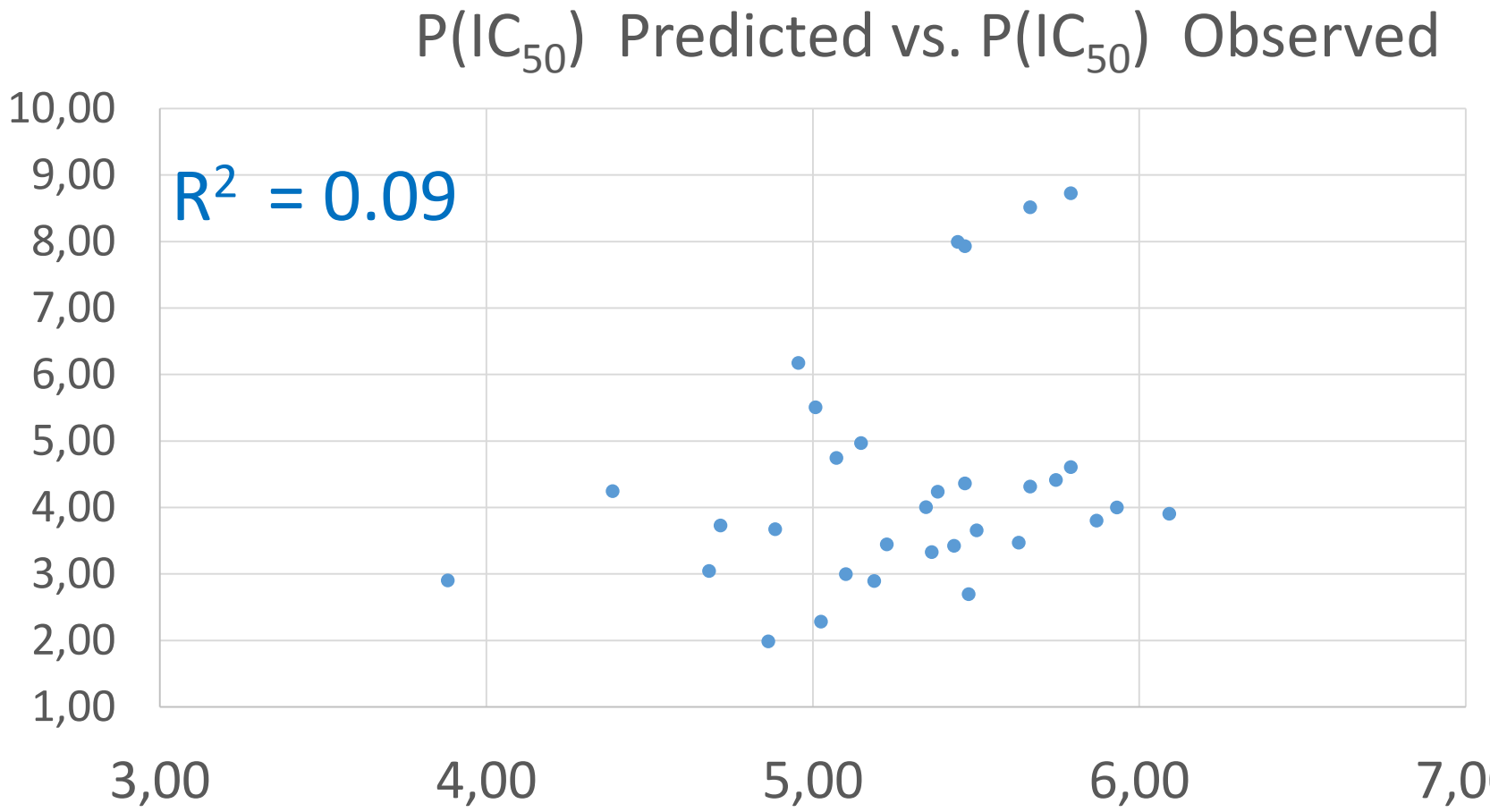
Как применять разработанный алгоритм на практике?

Классификация соединений из публикаций



[1] Lagunin A. et al., *Molecular Informatics*, 2011, 30(2-3), 241–250.
[2] A.V. Zakharov et al., *J Chem Inf Model* 2014, 54(3):705-12

Результаты прогноза для обучающей выборки ChEMBL



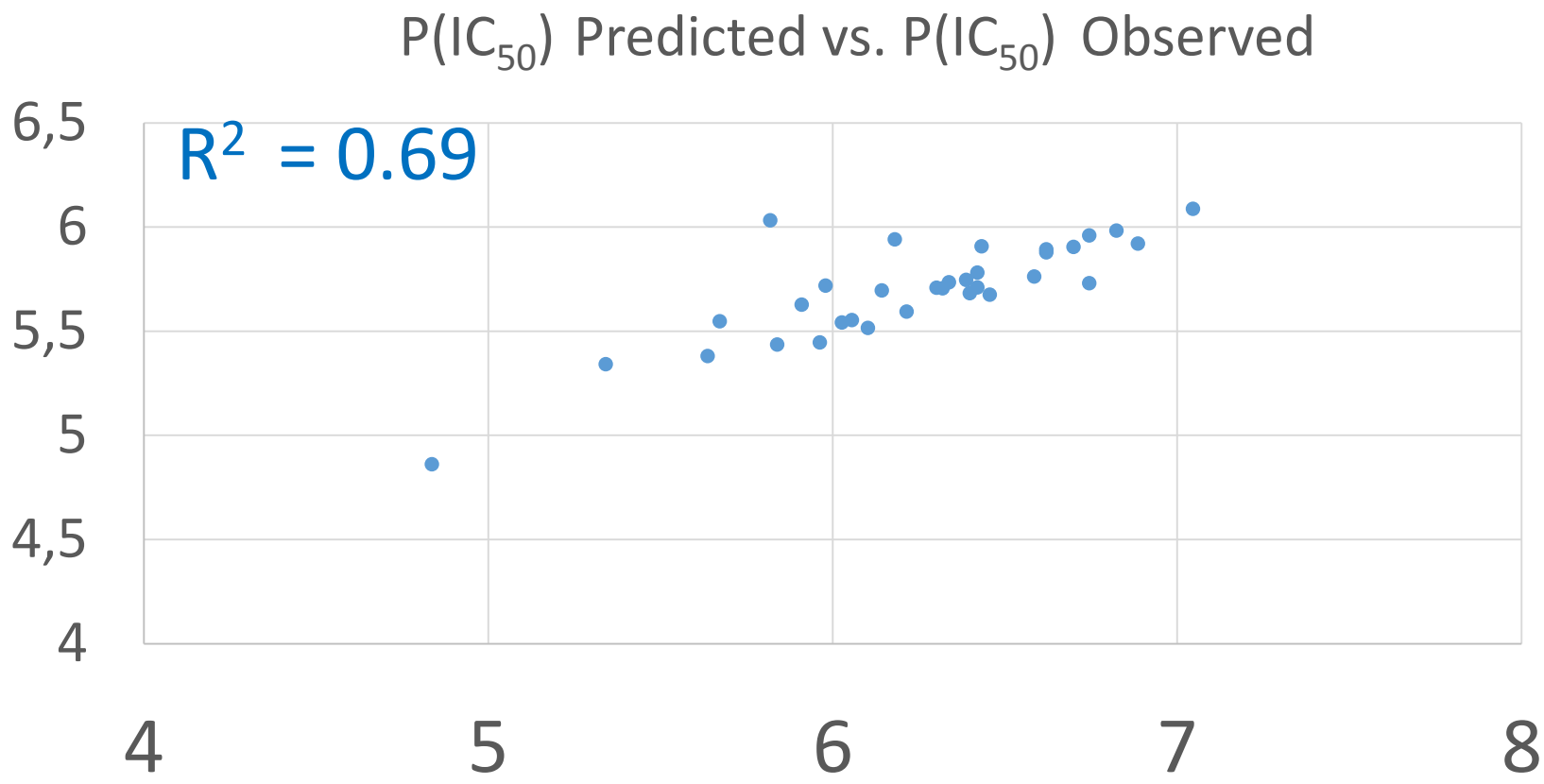
Прогноз выполнен для 33 соединений

Вне домена применимости модели: 4 соединения

Обучающая выборка: 1848 ингибиторов ОТ из БД ChEMBL

Тестовая выборка: 37 соединений из публикации(D.L. Romero et al., 1996, [PMID: 8809615])

Результаты прогноза для выборки с учетом данных о тестировании



Прогноз выполнен для 33 соединений

Вне домена применимости модели: 4 соединения

Обучающая выборка: 71 соединение из: Sluis-Cremer N., 2006 (PMID: 16884295) & Cantrell A.S., 1996 (PMID: 8863804)

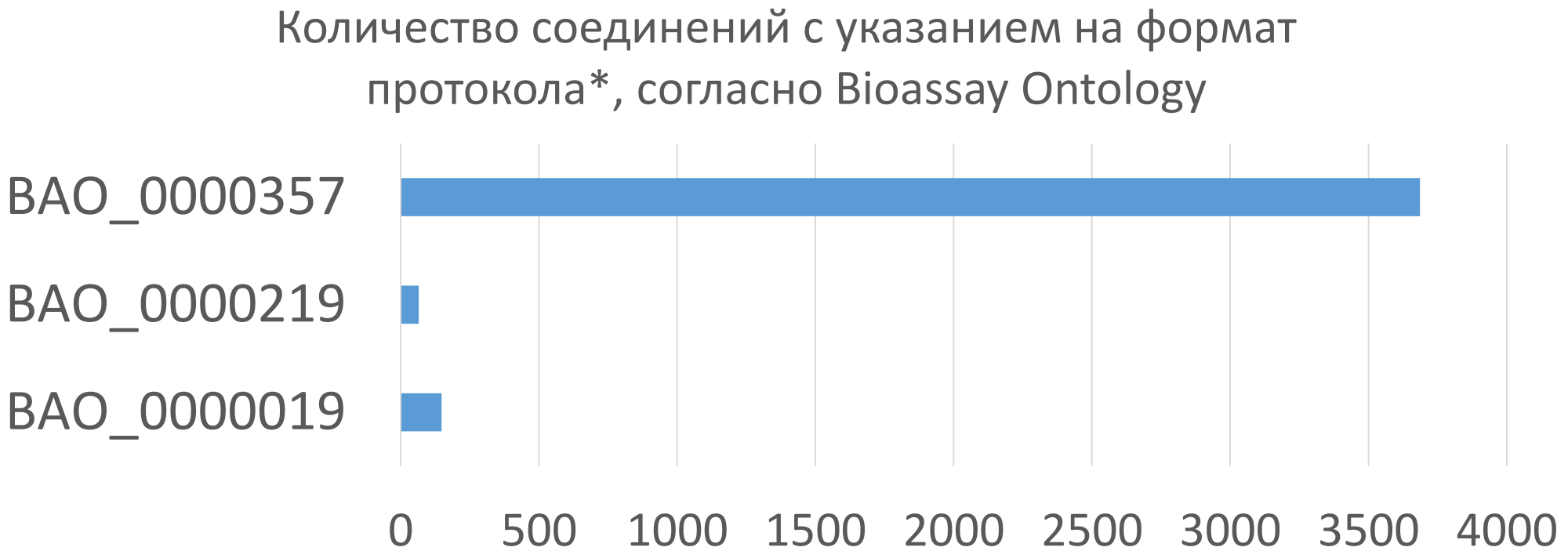
Тестовая выборка: 37 соединений из D.L. Romero et al., 1996 (PMID: 8809615)

Описание протоколов исследования

Проект	Характеристика	Авторы
Bioassay Ontology (BAO)	Семантические характеристики протоколов экспериментального тестирования	U. Visser, 2011
Assay Annotation	Аннотация протоколов тестирования из PubChem на основе структуры BAO	S.C. Schürer, 2012
BioAssay Express	Преобразование данных о протоколах исследования из PubChem в машиночитаемый формат	Alex M. Clark, 2014

ChEMBL

Ингибиторы обратной транскриптазы ВИЧ-1, IC50 (3899 соединений)



BAO_0000357 – Biochemical/protein-based
BAO_0000219 – Cell-based
* BAO_0000019 – Assay format

- Анализ полных текстов научных публикаций может быть применён к определению и классификации протоколов экспериментального тестирования органических соединений

- Отмечена высокая вариабельность описаний протоколов экспериментального тестирования ингибиторов обратной транскриптазы ВИЧ-1 в научных публикациях

- Разработанный алгоритм позволяет создать дескрипторы протоколов экспериментального тестирования, которые можно впоследствии применять для создания выборок органических соединений для анализа взаимосвязей «структура-активность»

Ограничения алгоритма

- Схожесть терминологии описания протоколов тестирования интегразы и обратной транскриптазы ВИЧ-1
- Неполнота описания протоколов тестирования в научных публикациях

- Необходимость учета синонимов для более точной классификации (пример - “cytotoxicity-cytopathicity”, “cell-based - cellular” и т. п.)

- Один абзац в большинстве случаев содержит описание одного эксперимента, но это не всегда так

Благодарности

Дмитрий Филимонов, ИБМХ

Владимир Поройков, ИБМХ

Алексей Захаров, NCATS, NCI, US

Андрей Ржецкий, University of Chicago, IL, US

Илья Майзус, University of Chicago, IL, US

Работа выполнена при поддержке гранта РФФИ № **16-34-60187**

Спасибо за внимание