# Разработка Windows-версии программы SigmoID для анализа регуляторной информации в последовательностях бактериальных геномов

В. Н. Яцков, В. В. Скакун, Е. А. Николайчик

*Белорусский государственный университет, Минск;*
*e-mail: vlad18742@gmail.com*

Важной задачей аннотации геномных последовательностей является идентификация регуляторных последовательностей, контролирующих экспрессию генов. Для коррекции аннотаций бактериальных геномов на основании анализа регуляторной информации был разработан программный комплекс SigmoID. Начиная с версии 2 в силу необходимости запуска сторонних POSIX программ данный комплекс потерял статус кроссплатформенного программного продукта вследствие невозможности его запуска на ОС Windows. Целью настоящей работы является исследование различных способов запуска POSIX программ непосредственно в ОС Windows и реализация одного из них в рамках проекта SigmoID.

**Ключевые слова:** регуляторные последовательности, аннотирование бактериальных геномов, SigmoID, Cygwin, WSL, кроссплатформенное приложение

## Введение

Интенсивное развитие геномных технологий определяет необходимость разработки программного обеспечения для автоматизации обработки огромных объемов генерируемых этими технологиями данных, в первую очередь данных высокопроизводительного секвенирования (NGS). И если задача сборки геномных последовательностей из данных NGS к настоящему времени в значительной мере решена, программные средства аннотации геномных последовательностей существенно отстают в своем развитии. В частности, до настоящего времени ни один из конвейеров аннотации геномов, включая наиболее популярные PGAP [1], RAST [2] и PROKKA [3], не аннотирует регуляторные последовательности (промоторы, операторы и терминаторы), что скрывает информацию об условиях экспрессии генов и существенно снижает практическую ценность автоматически аннотированных геномных последовательностей.

Программа SigmoID [4] была разработана для коррекции аннотаций бактериальных геномов на основании анализа регуляторной информации, доступной в специализированных базах данных RegPrecise, RegulonDB и CollecTF. SigmoID разрабатывался с использованием среды Xojo (https://www.xojo.com) как графическое кроссплатформенное приложение для трех основных десктопных систем, включая Linux, macOS и Windows, с основной идеей сделать доступными сложные алгоритмы анализа регуляторной информации для биологов с минимальным опытом в области биоинформатики. Вторая версия SigmoID [5, 6] была существенно модифицирована путем добавления конвейера анализа операторов и промоторов, распознаваемых неизученными транскрипционными факторами, что позволяет осуществлять практически полный анализ регуляторной информации для любых бактериальных геномов. Однако новые возможности потребовали более тесной интеграции со сторонними программами для анализа промоторных и операторных мотивов, что затруднительно в операционных системах, не поддерживающих стандарт POSIX, к каковым относится и большинство версий OS Windows. Поэтому поддержка SigmoID для OS Windows была приостановлена начиная с версии 2.0.

Появление в последних версиях Windows полноценного слоя совместимости для запуска GNU/Linux-приложений в виде Windows Subsystem for Linux (WSL) позволило

рассмотреть вопрос о возобновлении поддержки SigmoID для этой операционной системы. Целью настоящей работы является исследование различных способов запуска POSIX программ непосредственно в Windows и реализация одного из них в рамках проекта SigmoID.

**Результаты**

Приложение SigmoID является кроссплатформенным. Препятствием к полноценному запуску SigmoID на ОС Windows является необходимость вызова программ из пакетов HMMER (http://hmmer.org/), MEME Suite (https://meme-suite.org/meme/), MeShClust (https://github.com/BioinformaticsToolsmith/MeShClust), NCBI EDirect (https://www.ncbi.nlm.nih.gov/books/NBK179288/), разработанных для POSIX совместимых операционных систем (Linux, macOS и др.) и доступ к переменным окружения этих же ОС. Для запуска программ, написанных под другие операционные системы, можно использовать технологии виртуализации и эмулирования. Простое решение, заключающееся в создании виртуальной машины, используя, например, Oracle VM Virtual Box, установки на ней одной из POSIX совместимых ОС, например Ubuntu, и развертывании SigmoID, а также всех сторонних программ внутри нее, не является удачным, так как это не является средством создания кроссплатформенного приложения, а лишь позволяет виртуализовать другую операционную систему в рамках этой же машины. Запуск оконного приложения, установленного на виртуальной машине, невозможен на основной операционной системе. Соответственно, следует искать иные способы решения данной проблемы. Для исследования были выбраны два зарекомендовавших себя с лучшей стороны решения, позволяющие запуск POSIX программ непосредственно в среде Windows: Cygwin и WSL.

Cygwin – набор инструментов GNU с открытым исходным кодом, которые обеспечивают в Windows функциональность, аналогичную дистрибутиву Linux (https://cygwin.com/). В проекте исследовались следующие его возможности: запуск Bash-скриптов и компиляция POSIX программ с возможностью их использования в среде Windows (в результате компиляции создаются исполняемые в ОС Windows EXE-версии требуемых программ). Механизм использования Cygwin представлен на рис. 1.

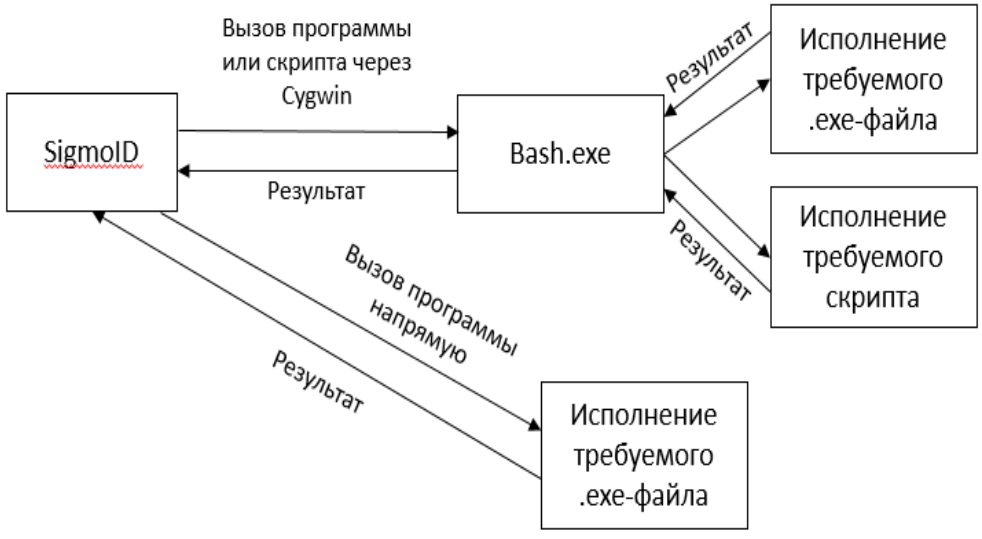


Рис. 1. – Схема использования Cygwin в проекте SigmoID.

Из рис. 1 видно, что скрипты можно запускать только через Bash.exe – один из инструментов, предоставляемых Cygwin, скомпилированные же программы возможно запускать как напрямую, так и через Bash.exe.

WSL (Windows Subsystem for Linux), прослойка для ОС Windows, написанная непосредственно самой корпорацией Microsoft, позволяет запускать среду GNU/Linux, включая инструменты командной строки, утилиты и приложения, не используя привычного механизма создания изолированных виртуальных машин, а используя инструменты виртуализации более низкого уровня [7].

В работе использовалась WSL версии 2. В качестве рабочей среды была установлена Ubuntu 20.04 LTS. Механизм использования WSL в проекте SigmoID представлен на рис. 2.

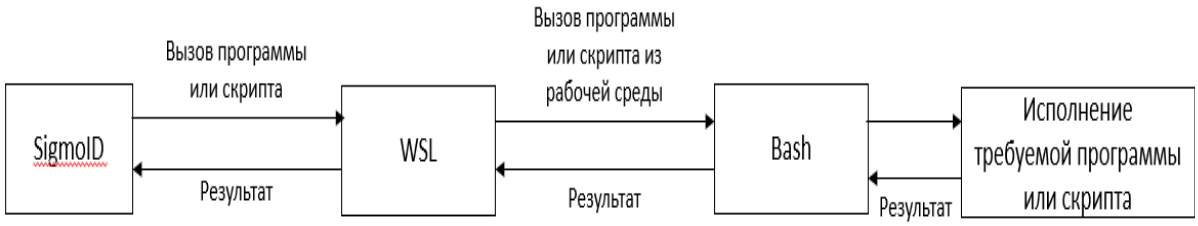


Рис. 2. – Схема использования WSL в проекте SigmoID.

Из рис. 2 видно, что все требуемые сторонние программы работают в той же рабочей среде, для которой они и были созданы, то есть в ОС Ubuntu 20.04 LTS. SigmoID же только делает вызовы в WSL, который, в свою очередь, транслирует данные вызовы в рабочую среду Ubuntu.

Реализация обоих подходов показала их полную состоятельность. Реализована возможность запуска проекта SigmoID и использование всей его функциональности непосредственно в ОС Windows.

Приведем результаты сравнительного анализа применения Cygwin и WSL в рамках портирования проекта SigmoID под Windows. Cygwin работает на большинстве ОС Windows: Windows Vista, 7, 8, 10, 11 и других. WSL доступен только для Windows 10 и 11 [7]. WSL имеет значительно более высокие показатели по эффективности взаимодействия с рабочей средой и скорости выполнения POSIX программ внутри нее [8]. Распространение полученного образа системы в WSL полностью свободно, что для конечного пользователя означает только установку WSL и ее рабочей среды с установленными пакетами программ. Лицензия Cygwin запрещает распространение ее образа, что означает, что пользователю потребуется самому устанавливать и компилировать из исходного кода все требуемые программные пакеты. Установка WSL и рабочей среды, например Ubuntu, не представляет большой сложности. Пакет Ubuntu для WSL можно поставить, просто выбрав его в магазине Microsoft (Microsoft Store).

Считая удобство пользователя по установке и настройке приложения SigmoID более приоритетным, для запуска сторонних (POSIX) программ, требуемых для его полной функциональности, в конечном итоге была выбрана WSL. Обновленный код программы SigmoID доступен через репозиторий github.cob/nikolaichik/SigmoID.

**Выводы**

Проведен сравнительный анализ различных подходов к запуску POSIX программ непосредственно в ОС Windows. В рамках проекта SigmoID реализованы две технологии: Cygwin и WSL. В результате проведенного сравнительного анализа предпочтение было отдано WSL, так как данный подход значительно упрощает установку и конфигурирование SigmoID, что является очень немаловажным, учитывая то, что пользователями SigmoID преимущественно являются люди, не имеющие специальных навыков программирования и администрирования операционных систем. Вызов POSIX программ посредством WSL происходит с высокой скоростью, лишь ненамного меньшей, чем в нативной среде, что является еще одним плюсом к выбору WSL.

**Литература**


1. Tatusova T. et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016. Vol. 44, № 14. P. 6614–6624.
2. Aziz R.K. et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008. Vol. 9, № 1. P. 75.
3. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014. Vol. 30, № 14. P. 2068–2069.
4. Nikolaichik Y., Damienikan A.U. SigmoID: a user-friendly tool for improving bacterial genome annotation through analysis of transcription control signals. PeerJ. 2016. Vol. 4. P. e2056.
5. Nikolaichik Y., Vychik P. New approach to genome-wide automated inference of bacterial transcription factor binding sites. Bioinformatics of Genome Regulation and Structure. Systems Biology. Institute of cytology and genetics, siberian branch of the russian academy of science, Novosibirsk State University, 2020. DOI: 10.18699/BGRS/SB-2020-046
6. Nikolaichik Y., Vychik P. Genome-wide inference of bacterial transcription factor binding sites: new method and its applications. BMC Bioinformatics. 2020. Vol. 21, № S20. P. O2.
7. Microsoft Docs [Электронный ресурс]. – Режим доступа: https://docs.microsoft.com/en-us/windows/wsl/. – Дата доступа: 24.09.2021.
8. Performance of WSL, Cygwin and Bare-Bone Linux on Generating This Website [Электронный ресурс]. – Режим доступа: https://thedrwu.com/posts/shell-perf-2019/. – Дата доступа: 04.10.2021.


## Development of Windows version of SigmoID application for the analysis of regulatory information in bacterial genome sequences


V.M. Yatskou, V.V. Skakun, Y.A. Nikolaichik

*Belarusian State University, Minsk; e-mail: vlad18742@gmail.com*



An important task in the annotation of genomic sequences is the identification of regulatory sequences that control gene expression. To correct the annotations of bacterial genomes based on the analysis of regulatory information, the SigmoID software package was developed. Starting from version 2, due to the need to run third-party POSIX programs, this package has lost the cross-platform status due to the impossibility of its proper execution on Windows OS. The purpose of this work is to study various ways to run POSIX programs directly in Windows and implement one of them within the SigmoID project.



**Keywords:** regulatory sequences, bacterial genomes annotation, SigmoID, Cudwin, WSL, cross-platform application.