DOI: 10.58168/CISMP2024_245-251
УДК 004.046

МЕТОДЫ РЕАЛИЗАЦИИ АЛГОРИТМА КРИПТОГРАФИЧЕСКОГО ХЕШИРОВАНИЯ ХИМИЧЕСКИХ СОЕДИНЕНИЙ

А.Г. Абрасимовская, И.С. Голубятников

ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

Аннотация. В статье рассмотрены способы оцифровки химических соединений с последующим выбором оптимального метода хеширования, отвечающего необходимым условиям для поставленных задач. Материал нацелен на создание базы тестовых данных для обучения нейросети для автоматического анализа состава различных веществ.

Ключевые слова: хемоинформатика, линейная нотация, InChI, криптографическое хеширование, SHA-2, Python.

METHODS OF ALGORITHM IMPLEMENTATION CRYPTOGRAPHIC HASHING OF CHEMICAL CONNECTIONS

A.G. Abrasimovskaya, I.S. Golubyatnikov

Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. The article discusses ways to digitize chemical compounds with the following choice of the optimal hashing method that meets the necessary conditions for the tasks set. The material is aimed at creating a database of test data for training a neural network for automatic analysis of the composition of various substances.

Keywords: chemoinformatics, linear notation, InChI, cryptographic hashing, SHA-2, Python.

Введение

На современном рынке информационных технологий для сельского хозяйства большую популярность приобретают беспилотные авиационные системы (БАС) с встроенной нейронной сетью для принятия различного рода решений на основе качественно-количественных показателей окружающей среды. В этом свете фрагментальная реализация тех или иных алгоритмов для последующего построения или обучения системы с искусственным интеллектом обретает значение для общего понимания существующих проблем на пути к реализации готового программно-технического продукта.

Следующим этапом после создания БАС для зондирования местности на предмет состояния посевов [1] является создание систем для автоматизации

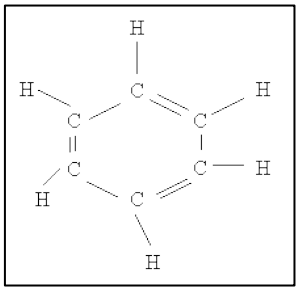
анализа пригодности местности для возделывания тех или иных культур. Для успешной реализации такого рода систем необходимым условием является существование программного обеспечения, способного на основе полученных данных о составе почвы и атмосферного воздуха дать прогноз о релевантности разработки местности.

Одной из ключевых функций проектируемого ПО является оцифровка и упорядочивание полученных химических данных. Рассмотрим существующие методы хемоинформатики и выберем оптимальный, отвечающий поставленным задачам.

1 Международный химический идентификатор (InChl)

Самым распространенным видом внутреннего представления структур химических соединений являются молекулярные графы (см. рис. 1). При необходимости, данная модель может быть дополнена сведениями о трёхмерных координатах атомов, а также о их динамике во времени.

Рисунок 1 – Пример молекулярного графа



К простейшим типам внешнего представления относятся линейные нотации. Пионером в данном направлении стал Вильгельм Висвессер, в 1948г запатентовавший линейное представление молекул для компьютерного анализа - Wiswesser line notation (WLN). [2]

Алфавит нотации Висвессера состоит из латинских букв, цифр и символов «&», «/», «-» и пробела. С учетом валентности атомов и ближайшего окружения, всем химическим элементам, основным типам связей, циклов и фундаментальных групп присвоены буквенные обозначения, образуя достаточно обширный и в некотором роде избыточный словарь.

Настоящее время большой популярностью пользуется Упрощенная Молекулярная Система Ввода Входной Строки (SMILES). Данная система относится к нотации однозначного описания состава и структуры соединения, в основе которой лежит кодировка ASCII. [3]

В такой записи неводородные атомы пишутся согласно обозначения в периодической системе химических элементов. Для элементов органической химии (B, C, N, O, S и пр.) скобки могут быть опущены, а также допускается отсутствие указание количества присоединенных атомов водорода, количество которых полагается равным минимальной нормальной валентности элемента.

Однако, на ряду с SMILES используются такие линейные нотации как СИБИЛ (Sybyl Line Notation), SMARTS (расширенная нотация SMILES для работы с базами данных), ROSDAL. Поэтому в 2005г Международным союзом теоретической и прикладной химии (IUPAC) было принято решение об унификации кодировки химических структур и разработана универсальная линейная нотация International Chemical Identifier (InChI) [4] (см. рис. 2).

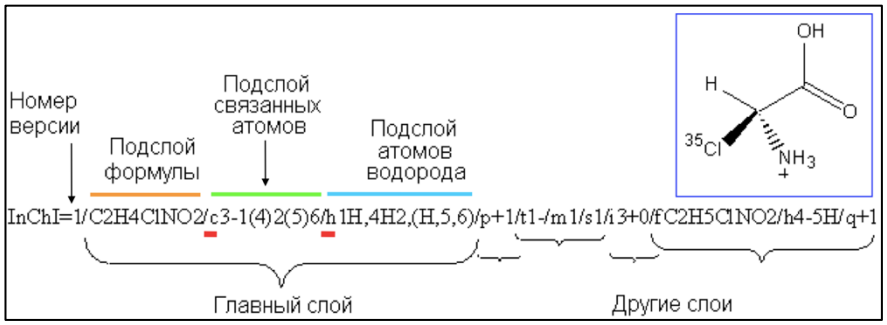


Рисунок 2 -Модульная структура InChI

Такая запись облегчает поиск информации в интернете, но ее главным недостатком является нефиксированная, и, зачастую, большая длина записи. Поэтому на основе InChI одновременно была разработана линейная нотация InChIKey, длина которой составляет 27 символов, что значительно упрощает ее индексацию в базах данных (см. рис. 3).

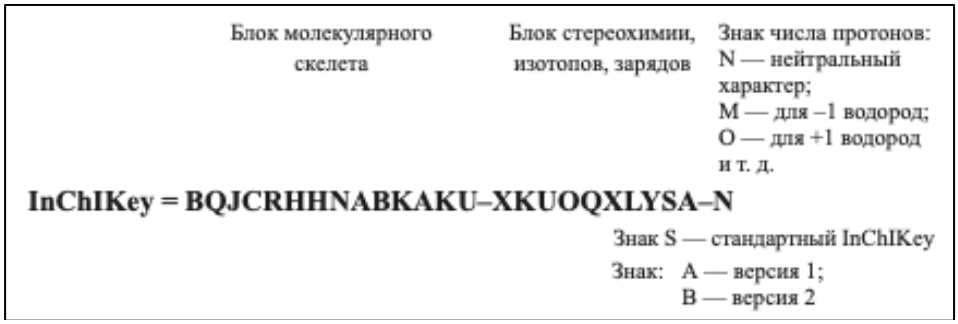


Рисунок 3 – Модульная структура InChIKey

Запись InChIKey создается на основе записи InChI методом хеширования второй алгоритмом криптографического хеширования Secure Hash Algorithm 2 (SHA-2).

2 Алгоритм криптографического хеширования SHA-2

Хеширование подразумевает под собой кодировку входного сообщения произвольной длины в битовую строку фиксированной длины, для последующего заполнения базы данных.

SHA-2 представляет собой семейство однонаправленных хеш-функций, включающее в себя алгоритмы SHA-224, SHA-256, SHA-384, SHA-512, SHA-512/256 и SHA-512/224 (см. табл. 1).

Таблица 1 - Основные характеристики хеш-функций семейства SHA-2

Хеш-функция	**Длина дайджеста сообщения (бит)**	**Длина блока (бит)**	**Макс длина сообщения (бит)**	**Длина слова (бит)**	**Количество итераций в цикле**	**Скорость (MiB/s)**
SHA-256, SHA-224	256/224	512	$2^{64} - 1$	32	64	139
SHA-512, SHA-384, SHA-512/256, SHA-512/224	512/384/256/224	1024	$2^{128} - 1$	64	80	154

Семейство SHA-2 разработано Агентством Национальной Безопасности США и опубликовано Национальным институтом стандартов и технологий (NIST) в федеральном стандарте обработки информации FIPS PUB 180-2 в августе 2002г. В этот стандарт также вошла хеш-функция SHA-1, разработанная в 1995г, однако, уже в 2005г она была признана небезопасной из-за высокого риска коллизий. [5]

В настоящее время данное семейство претерпело большое количество доработок, в виду чего остается одним из самых популярных способов хеширования, отвечающих современным требованиям безопасности. [6] Тем не менее, специалистам в области информационной безопасности уже удалось реализовать успешные методы конструирования коллизий для SHA-256 и SHA-512. [7]

Для реализуемого алгоритма наиболее оптимальной будет являться функция SHA-256 в виду скорости ее работы, а также достаточной длине дайджеста сообщения.

Принцип работы SHA-2 заключается в разбиении исходного дополненного сообщения на блоки, которые, в свою очередь, разбиваются на 8 слов. Каждый блок пропускается через цикл с 64 или 80 итерациями, на каждой из которых 2 из 8 слов преобразуются, а остальные слова задают функцию преобразования. Сумма конечных результатов обработки каждого блока является значением хеш-функции.

3 Примерный код функции

Для разработки кода функции был выбран высокоуровневый язык Python, в виду того, что он достаточно часто эксплуатируется в качестве языка написания нейросетей.

Рассмотрим код, преобразующий запись InChI в InChIKey с использованием алгоритма SHA-256. Результат записывается в динамическую таблицу соответствий InChI – InChIKey, для формирования базы знаний для обучения.

```
# Библиотека hashlib позволяет использовать хеш-функции, включая SHA-256
import hashlib

def inchi_to_inchikey(inchi):
    # Преобразуем InChI в InChIKey с помощью SHA-256
    hash_object = hashlib.sha256(inchi.encode('utf-8'))
    hash_digest = hash_object.hexdigest()

    # В соответствии с форматом InChIKey, разбиваем хэш на части
    inchikey = f"{hash_digest[:14]}-{hash_digest[14:18]}-{hash_digest[18:22]}-{hash_digest[22:26]}-{hash_digest[26:]}"

    return inchikey

# Динамическая таблица для хранения InChI и соответствующих InChIKey
inchi_key_table = []

# Пример записи InChI
inchi_examples = [
    "InChI=1S/H2O/h1H2",
    "InChI=1S/CH4/h1H4",
    "InChI=1S/C2H6O/c1-2-3/h2H,1H3"
]

for inchi in inchi_examples:
    inchikey = inchi_to_inchikey(inchi)
    inchi_key_table.append((inchi, inchikey))

# Вывод результатов
for inchi, inchikey in inchi_key_table:
    print(f"InChI: {inchi} -> InChIKey: {inchikey}")
```

Представленная функция будет полезна для создания различных наборов данных, которые могут быть использованы для обучения нейросети по обработке химических данных.

Заключение

Данный материал является первым шагом на пути к созданию нейросети для обработки полученным химических данных. Выбор описанных выше методов обработки информации обусловлен их популярностью, а также наличием действительного сертификата качества. Представленная функция на языке Python может быть использована в качестве инструмента для создания базы тестовых данных для машинного обучения.

Список литературы

1. Газета «Комерсантъ». – URL: tps://www.kommersant.ru/doc/6096816?ysclid=m1y5tcs8o037445764 (дата обращения: 07.10.2024).
2. Гелберг, Алан. Некролог: Уильям Джозеф Висвессер (1914-1989). // Журнал химической информации и компьютерных наук. – 1990. - №35. – С. 38-40.
3. Вейнингер, Дэвид. SMILES, химический язык и информационные системы. 1. Введение в методологию и правила кодировки. // Журнал химической информации и компьютерных наук. – 1988. – Т. 28, № 1. – С. 31-36.
4. Международный химический идентификатор (InChI) // Международный союз теоретической и прикладной химии. – URL: http://www.iupac.org/inchi/ (дата обращения: 07.10.2024).
5. Фергюсон Нильс, Шнайер Брюс. Практическая криптография. — М. : Диалектика, 2004. — 432 с. — 3000 экз. — ISBN 5-8459-0733-0, ISBN 0-4712-2357-3.
6. Титов, М. Ю. Пути повышения эффективных способов защиты информации в беспилотных летательных аппаратах двойного назначения / М. Ю. Титов // Моделирование систем и процессов. – 2024. – Т. 17, № 2. – С. 82-92. – DOI 10.12737/2219-0767-2024-17-2-82-92. – EDN IQOQPI.
7. Мендель Ф., Над Т., Шлэффер М. Улучшение локальных коллизий: новые атаки на сокращенный SHA-256 (англ.) // Достижения в криптологии - EUROCRYPT 2013: 32–я ежегодная международная конференция по теории и применению криптографических методов, Афины, Греция, 26-30 мая 2013 г. Труды / Т. Йоханссон, П. К. Нгуен – Шпрингер Берлин, Гейдельберг, 2013. – С. 262-278. – 736 с. – ISBN 978-3-642-38347-2 – doi:10.1007/978-3-642-38348-9_16.

References

1. The newspaper "Kommersant". – URL: https://www.kommersant.ru/doc/6096816?ysclid=m1y5tcs8o037445764 (date of access: 07.10.2024).
2. Gelberg, Alan. Obituary: William Joseph Wiswesser (1914-1989). // Journal of Chemical Information and Computer Science. – 1990. - No.35. – pp. 38-40.
3. David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules // J. Chem. Inf. Comput. Sci.. — 1988. — Т. 28, № 1. — P. 31—36.

4. International Chemical Identifier (InChI) // INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRYTHE (IUPAC). – URL: http://www.iupac.org/inchi / (date of access: 07.10.2024).

5. Ferguson Niels, Schneier Bruce. Practical Cryptography: Designing and Implementing Secure Cryptographic Systems. — Moscow: Dialectics, 2004. — 432 p. — 3000 copies. — ISBN 5-8459-0733-0, ISBN 0-4712-2357-3.

6. Titov, M. Yu. Ways to improve effective ways of protecting information in dual-purpose unmanned aerial vehicles / M. Yu. Titov // Modeling of systems and processes. - 2024. – Vol. 17, No. 2. – pp. 82-92. – DOI 10.12737/2219-0767-2024-17-2-82-92. – EDN IQOQPI.

7. Mendel F., Nad T., Schläffer M. Improving Local Collisions: New Attacks on Reduced SHA-256 (англ.) // Advances in Cryptology – EUROCRYPT 2013: 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings / T. Johansson, P. Q. Nguyen — Springer Berlin Heidelberg, 2013. — P. 262—278. — 736 p. — ISBN 978-3-642-38347-2 — doi:10.1007/978-3-642-38348-9_16.