Қазақстан Республикасы Тәуелсіздігінің 30 жылдығына арналған «Сейфуллин оқулары – 17: «Қазіргі аграрлық ғылым: цифрлық трансформация» атты халықаралық ғылыми – тәжірибелік конференцияға материалдар = Материалы международной научно – теоретической конференции «Сейфуллинские чтения – 17: «Современная аграрная наука: цифровая трансформация», посвященной 30 – летию Независимости Республики Казахстан.- 2021.- Т.1, Ч.4 - С.224-226

КЛАСТЕРИЗАЦИЯ И КЛАССИФИКАЦИЯ БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ КЛАССИЧЕСКИХ АЛГОРИТМОВ АНАЛИЗА ДАННЫХ

С. Кадыркеш

Наборы данных последовательности обычно встречаются вокруг нас. Например, клики, история просмотров видео и история посещение веб-страниц. В биоинформатике есть большие базы данных последовательностей белка, такие как UniProt. Белковая последовательность состоит из некоторой комбинации из 20 аминокислот. Обычная последовательность белка (рис. 1) состоит из набора букв, где каждая буква соответствует аминокислоте.

```
TGYHDNTWPIHIHPMRVPSKRHQRYGSVLNFAHMAGHMCHSVYGTSWLTMKKIWLWQAEF
WAECRHVWAICKYCIRLNNWNTFDYTMVTQEEYGQEQYYYLTEGHYSHMEENERYVNVHW
AFSGGFYQAGRVAYNIFTTGHVVDACYKCDCQRVMANDWMFSFRGFSWNPFHKEEHSDPQ
RPFADPMWPYPRQCIEVYDWRHITARWNMAPNSKQTPGMHWIYYLSDPIDHTTSHWVYEV
CGHEWQIHHCAQTENSACRHEPNTMTMNDCFMIFSCDQILQVIHTQDLPRGMQDINWMTE
FIPTEMIIWKLDSNLGCMAYRLQDVNHDDYMRFLLWVHDIYWARSCFSNRYSHSKFLSEV
YFLAFTAQMKQDALYFYQAGHMFKHQRASWDDGLYHADLRLMINSKFCWNPPEHLPSAEP
LNDRKCYLFSNFIVQELMVDCVGMWHMYVNMDVFCGWYGGWICKVCGYKKGFDAASKCFV
KRLLSTWIVCDFISFDVGIACRFRAKESIDQGYCKGAEPGPIFNHLWRGDQNFNWRPPVT
NNNNITMQSPSERPYNFQTGVRVRVWKYHWYNYPNTRYANEGAEVEYAQNMDADDHGSNF
NLMQHFSRTNYILHKVHVYKMMRYYYRGEVCNYKPCNTQPQESSLFRCSMWGQRYRWITQ
IPSQSQMMPHLCCKNASKGKHMWCRYDWGIHFRTICWYPLANGNMLEPSTAKGCPLHVMC
YKFWCTNDIWPTWYWSSCEPTCQKKKYGHRGCCGVSYLAAAKLNTSQSAPMYLFQMYENA
SPKTCEDGKSNAQECQAYMDQRIIPFESDAAQFIFWKLHTYQAMKWNKKHNCTGGQGKYE
NNVQGMADILMWNNRYNKANNIYMLCVELWIQTYYRPCAFMVKMTFATNAIGIWVMWWFH
CKYNAWNASKIVGHMGNNRDYKENLTMYQYVKFNTFGWSTTKFTLPTIMAPDGADYKCNQ
TLGYCNTMTHCRYDLNNCAAAVWTAPWKSGCQVEGFRHCQ
```

Рисунок 1 - Пример последовательности белка

Белковая последовательность может содержать не все 20 аминокислот, а некоторое ее подмножество. Корпус последовательностей обычно содержит от тысяч до миллионов последовательностей. Кластеризация и классификация чаще всего требуется, если есть размеченные или неразмеченные данные. Однако сделать это непросто из-за произвольных последовательностей произвольной длины.

Данные о последовательностях встречаются очень часто, что привело к развитию различных методов последовательного анализа. Исследования последовательного анализа можно в общих чертах можно разделить на:

- поиск частых паттернов или подпоследовательностей;
- обнаружение мотивов;
- выравнивание;
- моделирование потока данных;
- встраивание признаков.

Среди них особенно важно встраивание функций, поскольку оно обеспечивает машинно-понятное представление последовательностей. Их можно использовать непосредственно для вычисления или «расстояния» между последовательностями или другие модели машинного обучения. Подобный подход (word2vec) популярен в интеллектуальном анализе текста для преобразования текста в векторные разложения. Это позволяет построить модели классификации последовательностей и кластеризации, которые имеют огромное применение в онлайн-индустрии, биоинформатике и здравоохранении.

Однако внедрение функций является сложной задачей, потому что:

- последовательности произвольны, а строки произвольной длины;
- долгосрочные зависимости трудно уловить (влияние далеких элементов последовательности друг на друга).

Помимо других ограничений большинство существующие методы либо ограничиваются извлечением только краткосрочных шаблонов, либо страдают от увеличения объема вычислений при извлечении долгосрочных паттернов.

Анализ последовательностей – это широко изучаемая проблема. Несколько работ были сделаны для оценки сходства последовательностей и представлений признаков для классификации последовательностей, кластеризации и т.д.

Можно классифицировать методы для анализа последовательностей следующим образом:

- Выравнивание;
- Строковые ядра;
- Классификации временных рядов;
- Глубокое обучение;
- Обнаружение паттернов;
- Хэш-карты;
- Генеративные;
- Временные графы.

Проблемы и актуальность нашего исследования обуславливаются тем, что для современной молекулярной биологии характерно экспоненциальное увеличение объемов доступных геномных данных. Несмотря на постоянный рост производительности вычислительной техники, одного этого недостаточно для удовлетворения растущих потребностей биоинформатики. Поэтому создание и использование быстрых вычислительных алгоритмов в задачах биоинформатики все еще актуально.

В рамках диссертационного исследования нами были изучены белковые последовательности из различных видов живых организмов, проведен обзор научно-исследовательских работ в области протеомики и биоинформатики. А также проводятся исследования классических алгоритмов анализа данных для кластеризации и классификации белковых последовательностей с целью выполнения практической части диссертационного исследования, так как основной задачей является разработка программного обеспечения с использованием классических алгоритмов анализа данных для классификации и кластеризации белковых последовательностей.

Список литературы

1. Любецкий В.А., Селиверстов А.В., Зверков О.А. Построение разделяющих паралоги семейств гомологичных белков, кодируемых в пластидах цветковых растений // Математическая биология и биоинформатика. 2013. Т. 8, № 1. С. 225–233.;

2. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins // Journal of Molecular Biology. 1970. V. 48, No. 3. P. 443– 453.;

3. Ranjan C., Ebrahimi S., Paynabar K. Sequence graph transform (SGT): A feature extraction function for sequence data mining //arXiv preprint arXiv:1608.03533. – 2016.;

4. Aggarwal, C.C., Han, J.: Frequent pattern mining. Springer (2014);

5. Altschul, S.F., Madden, T.L., Sch¨affer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids research 25(17), 3389–3402 (1997);

6. Chiu, D.Y., Wu, Y.H., Chen, A.L.: An efficient algorithm for mining frequent sequences by a new strategy without support counting. In: Data Engineering, 2004. Proceedings. 20th International Conference on, pp. 375–386. IEEE (2004);

7. Eskin, E., Weston, J., Noble, W.S., Leslie, C.S.: Mismatch string kernels for svm protein classification. In: Advances in neural information processing systems, pp. 1441–1448 (2003);

8. Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S.: Mismatch string kernels for discriminative protein classification. Bioinformatics 20(4), 467–476 (2004)

Научный руководитель Исмаилова А.А., PhD