УДК 575.224

ПРЕДСКАЗАНИЕ ПАТОГЕННЫХ ВАРИАНТОВ MYH7 МЕТОДОМ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ ДАННЫХ ОБ АГРЕГАЦИИ И СТРУКТУРНОЙ СТАБИЛЬНОСТИ БЕЛКА

Кокорина М.А. (ИТМО)

Научный руководитель – Пьянков И.А. (ИТМО, СПбГУ)

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 21-74-20093-П от 28.05.2025).

Введение. Патогенные варианты в гене MYH7 вызывают широкий спектр нарушений, которые существенно влияют на работу сердечной и скелетных мышц. Варианты в стержневом домене белка приводят к его агрегации и формированию внутриклеточных включений, что лежит в основе миозиновой накопительной миопатии (MSM) [1]. Механизмы и последствия мутаций такого типа изучены недостаточно по сравнению с мутациями в головном домене миозина. Существующие вычислительные методы плохо приспособлены для прогнозирования патогенных вариантов, специфично связанных с нарушением стабильности и агрегацией белка. В данной работе мы представляем предсказательную модель RDSM-MYH7, созданную на основе машинного обучения с использованием ранее разработанного подхода [2].

Основная часть. В качестве структуры для анализа была впервые получена полноразмерная модель димера MYH7. Из-за большого размера белка (1935 а.о.) и ограничений современных методов предсказания структуры, модель была собрана из семи перекрывающихся фрагментов, смоделированных в AlphaFold3 [3], с последующей релаксацией и оптимизацией.

Отбор патогенных вариантов, связанных с MSM проводился по трем главным критериям:

1) расположение в стержневом домене белка;
2) классификация вариантов как Pathogenic или Likely Pathogenic в базе данных ClinVar;
3) ассоциация варианта с MSM в базе данных ClinVar.

Мутации не связанные с MSM также отбирались из различных баз данных по расположению в стержневом домене, и по оценке патогенности (Benign/ Likely benign; Uncertain significance с аннотацией, позволяющей отнести мутацию к непатогенным). По итогам отбора был сформирован датасет из 40 вариантов, 24 из которых были классифицированы как связанные с MSM, оставшиеся 16 – как не связанные с MSM.

Оценка эффекта варианта на стабильность димера проводилась с помощью инструментов, опирающихся на структурные данные (FoldX, ICM, DDGun), а также основанных на методах машинного обучения (mCSM, DynaMut2) [4-8]. Статистическая значимость различий в предсказаниях между группами патогенных и нейтральных вариантов оценивалась для каждого инструмента. Кроме того, была проведена оценка патогенности вариантов с помощью инструментов прогнозирования AlphaMissence и PolyPhen-2 [9,10]. Был произведен анализ агрегационного потенциала вариантов с помощью конвейера TAPASS, интегрирующего три комплементарных алгоритма (ArchCandy, TANGO, PASTA) [11].

Для выбора наиболее подходящего метода мы провели сравнительный анализ семи классификаторов, представленных в библиотеке Scikit-learn на Python [12], представляющих разные подходы машинного обучения: линейные методы, ансамблевые методы на основе деревьев и простейшую нейронную сеть (многослойный перцептрон). Надежное обучение и оценка модели на небольшом наборе данных была обеспечена схемой вложенной 5x3 кросс-валидации с автоматическим подбором гиперпараметров через фреймворк Optuna. Это позволило избежать переобучения и получить объективную оценку эффективности.

Выводы. Впервые методом гибридного моделирования на основе AlphaFold3 получена и валидирована полноразмерная атомарная модель димера MYH7 человека, что позволило провести направленный анализ эффектов мутаций. Установлено, что патогенные варианты, ассоциированные с миозиновой накопительной миопатией (MSM), статистически значимо чаще приводят к дестабилизации структуры димера и располагаются в локально разупорядоченных участках стержневого домена. Разработан специализированный алгоритм машинного обучения RDSM-MYH7, который по совокупности метрик (F1-score = 0.869, AUC-ROC = 0.908) превосходит инструменты AlphaMissense и PolyPhen-2 в задаче предсказания патогенных вариантов MYH7. Предложенный подход позволяет целенаправленно выявлять патогенные варианты, механизм действия которых связан со снижением структурной стабильности и повышенной склонностью белка к агрегации.

Список использованных источников:

1. Naderi N. et al. A novel heterozygous missense MYH7 mutation potentially causes an autosomal dominant form of myosin storage myopathy with dilated cardiomyopathy // BMC Cardiovascular Disorders. – 2023. – Vol. 23. – № 1. – P. 1–8. DOI: 10.1186/s12872-023-03538-8.
2. Pyankov I.A. et al. A computational approach to predict the effects of missense mutations on protein amyloidogenicity // Journal of Structural Biology. – 2025. – Vol. 217. – P. 108176. DOI: 10.1016/j.jsb.2025.108176.
3. Abramson J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3 // Nature. – 2024. – Vol. 630. – P. 493–500. DOI: 10.1038/s41586-024-07487-w.
4. Schymkowitz J. et al. The FoldX web server: an online force field // Nucleic Acids Research. – 2005. – Vol. 33. – Web Server issue. – P. W382–W388. DOI: 10.1093/nar/gki387.
5. Pires D.E.V. et al. mCSM: predicting the effects of mutations in proteins using graph-based signatures // Bioinformatics. – 2014. – Vol. 30. – № 3. – P. 335–342. DOI: 10.1093/bioinformatics/btt691.
6. Abagyan R., Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins // Journal of Molecular Biology. – 1994. – Vol. 235. – № 3. – P. 983–1002. DOI: 10.1006/jmbi.1994.1052.
7. Montanucci L. et al. DDGun: An untrained method for the prediction of protein stability changes upon point variations // BMC Bioinformatics. – 2019. – Vol. 20. – № 1. – P. 1–10. DOI: 10.1186/s12859-019-2923-1.
8. Rodrigues C.H.M. et al. DynaMut2: Assessing changes in stability and flexibility upon missense mutations // Protein Science. – 2021. – Vol. 30. – № 1. – P. 60–69. DOI: 10.1002/pro.3942.
9. Tordai H. et al. Analysis of AlphaMissense data in different protein groups and structural context // Scientific Data. – 2024. – Vol. 11. – № 1. – P. 1–12. DOI: 10.1038/s41597-024-03327-8.
10. Adzhubei I. et al. Predicting functional effect of human missense mutations using PolyPhen-2 // Current Protocols in Human Genetics. – 2013. – Vol. 76. – № 1. – P. 7.20.1–7.20.41. DOI: 10.1002/0471142905.hg0720s76.
11. Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.
12. Ahmed A. B. et al. A structure-based approach to predict predisposition to amyloidosis // Alzheimer's & Dementia. – 2015. – Vol. 11. – № 7. – P. 681–690. DOI: doi.org/10.1016/j.jalz.2014.06.007.