ПОБУДОВА ДЕРЕВА РІШЕНЬ ДЛЯ ПЕРЕДБАЧЕННЯ ФОРМИ БІОГЕННИХ МАГНІТНИХ НАНОЧАСТИНОК У МАГНІТОТАКСИСНИХ БАКТЕРІЯХ

Хахно К.Ю., Горобець С.В.
КПІ ім. Ігоря Сікорського, khakhnokyryl@gmail.com

Abstract
Magnetic nanoparticles (MNPs) play key roles in targeted therapies and other fields. Magnetotactic bacteria (MTB) produce diverse biogenic magnetic nanoparticles (BMNs), but characterizing their shapes is challenging. The study develops advanced classification models combining decision tree analysis and protein characterization, improving our understanding of BMN shapes.

Keywords: *biogenic magnetic nanoparticles; decision tree classifier; mam-proteins.*

Вступ. Магнітні наночастинки (МНЧ) використовуються в різноманітних галузях, проте ключову роль вони відіграють у медичних дослідженнях та практиці, включаючи магнітно-резонансну томографію як контрастні агенти для покращення якості зображення тканин та гіпертермічну терапію для боротьби з пухлинами. Окрім того, МНЧ можуть бути використані в таргетованій доставці ліків, зменшуючи побічні ефекти та підвищуючи ефективність лікування, а також в тераностиці як терапевтичні та діагностичні агенти, що забезпечують візуалізацію пухлин та контроль їх реакції на лікування [1, 2].

Основною властивістю МНЧ є здатність рухатись за напрямком магнітного поля, що робить їх подібними до керованих векторів. Така здатність залежить не лише від матеріалу наночастинок, але й від їх розмірів та форми [2]. Під час штучного синтезу наночастинок магнетиту, щоб контролювати форму, підбирають відповідні умови синтезу (температуру, тиск, розчинники) [1, 2], але для виділення чистих, безпечних для використання в медицині, наночастинок потрібно використовувати складні методи. Натомість заміною можуть бути біогенні магнітні наночастинки (БМН), які є біологічно сумісними. Найбільш відомими продуцентами БМН є магнітотаксисні бактерії, представники яких є різноманітними за морфологією прокаріотами. Основною проблемою дослідження форми утворених такими бактеріями наночастинок є те, що більшість з них неможливо або складно культивувати в лабораторних умовах [3], тому зараз спостерігається недостатня кількість експериментальних підтверджень форми БМН у таких організмах. Передбачення форми на основі механізму біомінералізації також неможливо на даний час, оскільки відомо лише про дію основних білків з магнітосомного острівця, що відповідають за утворення магнітних наночастинок у МТБ. За останній час з’явилось безліч доступних методів секвенування геному та його анотації, але дослідження функцій білків все ще залишається довгим та дороговартісним. Тому метою цієї роботи є створення класифікатора магнітотаксисних бактерій на основі даних про наявність чи відсутність генів, що кодують відповідальні за біомінералізацію магнетиту та грейгіту білки.

Розуміння впливу білків на форму магнітних наночастинок також може сприяти розвитку методів штучного синтезу таких наночастинок, оскільки

експериментально підтверджено, що білки, наприклад, Mms6, мають важливе значення для контролю форми МНЧ, утворених *in vitro* [1].

Матеріали і методи. У якості початкових даних було знайдено інформацію про 29 геномів магнітотаксисних бактерій у базі даних GenBank NCBI (https://www.ncbi.nlm.nih.gov/), а також відповідні амінокислотні послідовності у БД Protein NCBI. Додатково було використано дані зі статті [4], де міститься інформація про ще 18 геномів МТБ. Крім пошуку вручну, було використано інструмент Entrez у складі бібліотеки Biopython для мови програмування Python для автоматизованого пошуку записів про амінокислотні послідовності білків. Для аналізу було використано біоінформаційні методи, а саме: мову програмування Python і бібліотеки Biopython, pandas та sklearn.

З метою аналізу даних використано пакет sklearn, який містить класифікатор для побудови дерева рішень (DecisionTreeClassifier) – структури у вигляді бінарного дерева, де кожен вузол відображає точку рішення, ребра відповідають можливим варіантам наслідків цього рішення, враховуючи відповідність або невідповідність поставленій умові, а листки вказують на окремі класи [5].

Результати та обговорення. У першій моделі відповідно до правил побудови статистичних моделей дані було розділено на тренувальні та тестові. Їх співвідношення було обрано як 2:1, щоб забезпечити належний рівень тестування. Таким чином у тренуванні дерева рішення було задіяно 31 запис із таблиці, яка була сформована на основі експериментальних даних, що узяті зі статті [4]. У результаті отримуємо графічне представлення (рис. 1).

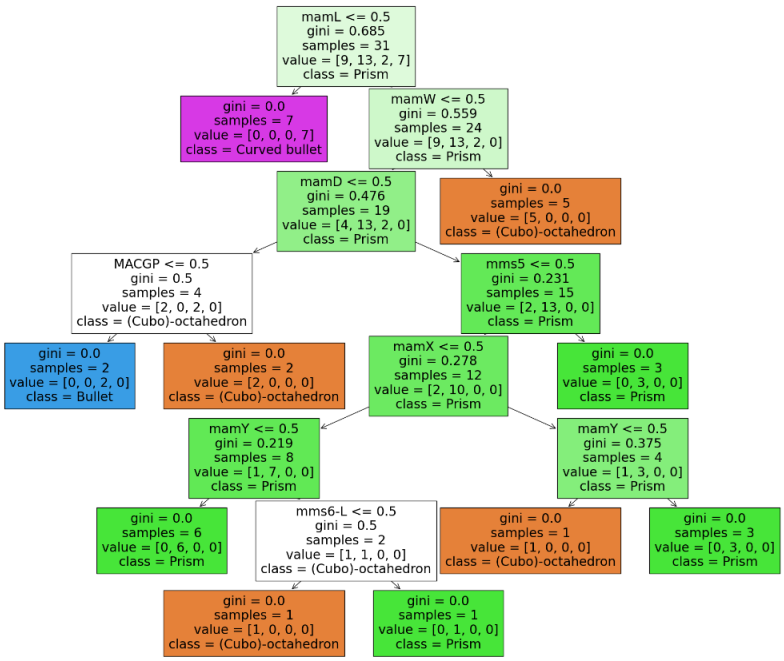

Рис. 1. Графічне представлення дерева рішень на основі даних статті [4].

Для тестування отриманої моделі було використано решту 16 записів про геноми магнітотаксисних бактерій. Всю необхідну інформацію для порівняння реальної та передбаченої форми БМН занесено до табл. 1.

Таблиця 1. Результати тестування першої моделі.

№	Штам	Експериментально встановлена форма	Передбачена форма	Чи є передбачення правильним
27	YQC-5	Prism	Prism	Так
39	XYR	Curved bullet	Curved bullet	Так
26	YQC-3	Prism	(Cubo)-octahedron	Ні
43	LBB02	Curved bullet	Curved bullet	Так
24	YQC-9	Prism	Prism	Так
36	BW-1	Bullet	Bullet	Так
12	MC-1	(Cubo)-octahedron	(Cubo)-octahedron	Так
19	LM-1	Prism	Prism	Так
4	SO-1	(Cubo)-octahedron	(Cubo)-octahedron	Так
25	UR-1	Prism	Prism	Так
8	LBB-42	(Cubo)-octahedron	Prism	Ні
3	ME-1	(Cubo)-octahedron	(Cubo)-octahedron	Так
6	BB-1	(Cubo)-octahedron	Prism	Ні
40	YQR-1	Curved bullet	Curved bullet	Так
33	RS-1	Bullet	Bullet	Так
13	IT-1	(Cubo)-octahedron	(Cubo)-octahedron	Так

Оскільки використаний алгоритм машинного навчання є нестабільним [6], тобто залежить від розподілу на тренувальні і тестові набори, то отримуємо дані, що розроблена модель помиляється у 15-20% випадків. Щоб зрозуміти причини хибних передбачень, можна скористатись однією з переваг дерев рішень над моделями глибокого навчання, тобто можливістю детермінувати принцип розділення за допомогою графічного представлення навченої моделі [5]. Тому, проаналізувавши рис. 1, можна зробити висновок, що модель має конкретні умови для розділення магнітотаксисних бактерій, що формують біогенні магнітні наночастинки у формі кулі (Bullet) чи вигнутої кулі (Curved bullet). Натомість розділення двох інших класів ((Cubo)-octahedron та Prism) вимагає ще трьох додаткових рівнів глибини, умови в яких варіюються за різних випадкових станів розподілу. Тобто, моделі не вистачає даних для однозначної класифікації.

Другу модель було вирішено будувати з доповненням інформації про білки, що кодуються присутніми в обох класах генами – mamL та mamO. Для 14 штамів з відомими послідовностями відповідних протеїнів було розраховано властивості за допомогою інструменту ProtParam бібліотеки Biopython мови програмування Python. Розглянуто такі ознаки як відсоток різних амінокислот, молекулярну вагу, ароматичність, гнучкість, ізоелектричну точку, індекс нестабільності, розподіл вторинних структур та відносну вагу вирівнювання. Для кожної потенційної властивості було розраховано значення p-value за допомогою критерію Стьюдента. Статистично значуща відмінність (p-value < 0,05) у двох вибірках була знайдена для двох ознак mamL: значення ізоелектричної точки та відносної ваги вирівнювання. Відповідні значення від ProtParam були перенесені у таблицю з даними для тренування.

Оскільки кількість досліджуваних штамів у другій моделі зменшилась до 29, то використовувати розділення на тренувальний та тестовий набір не завжди дає змогу виділити всі 4 класи. Тому можливо два варіанти: підібрати відповідні

випадкові стани та провести тестування (рис. 2, а) чи залишити модель без тестування (рис. 2, б).

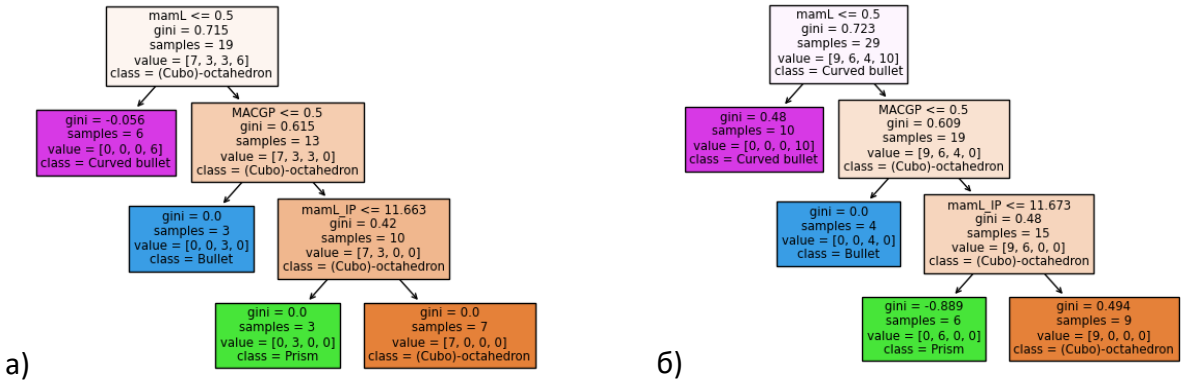

Рис. 2. Графічне представлення дерева рішень на основі розширених даних про білки mamL та mamO: а) із розділенням на тренувальні та тестові дані; б) без розділення.

На основі побудованих дерев рішень видно, що спочатку з усіх проаналізованих штамів МТБ відділяються ті, що формують наночастинки магнетиту у вигляді вигнутих куль (Curved bullet). Якщо у геномі не виявлено ген mamL, то ці організми відносяться до класу Curved Bullet. Тоді решта переходить до наступного вузла дерева, де ключовою ознакою є наявність/відсутність специфічних генів для типу Pseudomonadota (MACGP), до яких відносяться білки mamH, mamF, mamS та mamT. Якщо вони відсутні, то ці організми відносять до класу Bullet. А решта розділяється за значенням ізоелектричної точки білка, який кодується геном mamL. За відповідним пороговим значенням виділяється два останні класи Prism та (Cubo)-octahedron.

Висновки. На основі анотованих геномів із бази даних GenBank NCBI та даних зі статті [4] побудовано дерева рішень для класифікації магнітотаксисних бактерій за формою біогенних магнітних наночастинок, які вони утворюють. Розроблені моделі можна використати для передбачення форми БМН для нових магнітотаксисних бактеріях лише за секвенованим геномом.

Список використаної літератури:

1. Горобець С.В., Горобець О.Ю., Горбик П.П., Уварова І.В. Функціональні біо- та наноматеріали медичного призначення: монографія. Київ. 2018. 480 с. http://www.materials.kiev.ua/publications/5.pdf

2. Shape-, size- and structure-controlled synthesis and biocompatibility of iron oxide nanoparticles for magnetic theranostics / W. Xie et al. Theranostics. 2018. Vol. 8. No. 12. P. 3284–3307. https://doi.org/10.7150/thno.25220

3. Le Nagard L., Morillo-López V., Fradin C., Bazylinski D. A. Growing Magnetotactic Bacteria of the Genus Magnetospirillum: Strains MSR-1, AMB-1 and MS-1. Journal of Visualized Experiments. 2018. No. 140. https://doi.org/10.3791/58536

4. Liu P., Zheng Y., Zhang R., Bai J., Zhu K., Benzerara K., Menguy N., Zhao X., Roberts A. P., Pan Y., Li J. Key gene networks that control magnetosome biomineralization in magnetotactic bacteria. National Science Review. 2023. Vol. 10. No. 1. https://doi.org/10.1093/nsr/nwac238

5. Decision trees: from efficient prediction to responsible AI / H. Blockeel et al. Frontiers in Artificial Intelligence. 2023. Vol. 6. https://doi.org/10.3389/frai.2023.1124553

6. Mirzamomen Z., Kangavari M. R. A framework to induce more stable decision trees for pattern classification. Pattern Analysis and Applications. 2016. Vol. 20. No. 4. P. 991–1004. https://doi.org/10.1007/s10044-016-0542-2