КОМП'ЮТЕРНІ НАУКИ

ISSN 2707-1898 (print)

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2021.02.027
Article received 07.11.2021 p.
Article accepted 23.11.2021 p.
UDK 004.[48+94]

Correspondence author
V. I. Lytvynenko
immun56@gmail.com

І. М. Фефелова, В. І. Литвиненко, А. О. Фефелов
Херсонський національний технічний університет, м. Херсон, Україна

ПРОГНОЗУВАННЯ ТРЕТИННОЇ СТРУКТУРИ БІЛКА НА ДВОМІРНІЙ ТРИКУТНІЙ ҐРАТЦІ ГІБРИДНИМ ЕВОЛЮЦІЙНИМ АЛГОРИТМОМ

Розглянуто завдання прогнозування третинної структури білка з урахуванням його первинної послідовності. Проблема в тому, що науковці, з усією своєю обчислювальною потужністю і набором експериментальних даних, не навчилися будувати моделі, які описують процес згортання молекул білка і прогнозують третинну структуру білка на основі його первинної структури. Однак неправильно вважати, що в цій галузі науки нічого не відбувається. Відомо закономірності складання (згортання) білка, розроблено методи його моделювання. Аналіз поточного стану досліджень щодо цих проблем свідчить про наявність недоліків, пов'язаних із точністю прогнозування і часом, необхідним для отримання оптимального рішення. Отже, розроблення нових обчислювальних методів, позбавлених цих недоліків, є актуальним. Зосереджено увагу на моделі ґратки, що є особливим випадком відомого гідрофобно-полярного кропу. запропоновано конформацію білка за обраною моделлю, гібридні алгоритми клонального відбору, диференціал. Оскільки процеси згортання білка до кінця не вивчені, дослідники запропонували ряд спрощених моделей, заснованих на фізичних властивостях молекул, що призводить до проблем комбінаторної оптимізації. Як модель білка обрано гідрофобно-полярну спрощену модель на плоскій трикутній ґратці. З погляду задачі оптимізації, проблема фолдингу білка зводиться до пошуку конформації з мінімальною енергією. У ґратчастих моделях конформацію представлено у вигляді шляху, що не має самоперетинів. Для вирішення цієї проблеми запропоновано гібридну штучну імунну систему у формі комбінації алгоритмів клонального відбору та диференціальної еволюції. Розроблений гібридний алгоритм використовує спеціальні способи кодування та декодування індивідуумів, а також функцію афінності, що дають змогу зменшити кількість некоректних конформацій (рішень з самоперетинами). Доведено, що в цій рецептурі завдання складання білка є NP-повним. Тому загалом точні методи не здатні впоратися з поставленим завданням у прийнятний час. Для перевірки ефективності алгоритму проведено експериментальні дослідження на тестових послідовностях. Для тестування алгоритмів обрано гідрофобно-полярну модель Ділла на двомірній трикутній ґратці. Здійснено експериментальні дослідження на тестових послідовностях, які показали переваги розроблених алгоритмів перед іншими методами.

Ключові слова: фолдинг білка; гідрофобно-полярна модель; клональний відбір; диференціальна еволюція; штучні імунні системи; гідрофобно-полярна модель.

Вступ

Білки – це довгі послідовності з 20 різних амінокислот. Білки, як відомо, мають багато важливих функцій у клітині, такі як: ферментативна активність, зберігання і транспортування матеріалу, трансдукція сигналу, антитіла і багато інших. Амінокислотний склад білка однозначно визначає його тривимірну структуру, з якою безпосередньо пов'язана функціональність білка.

Проблема прогнозування структури білка сьогодні є однією з найскладніших проблем у біохімії та біоінформатиці. Це просто визначається як задача розуміння і прогнозування того, як інформація, закодована в амінокислотній послідовності білків, перетворюється в тривимірну структуру біологічно активного білка. Вирішення проблеми прогнозування структури білка дало б змогу значно спростити задачу розуміння механізму спадкових та інфекційних захворювань, розроблення препаратів зі специфічними лікувальними властивостями та вирощування біологічних полімерів зі специфічними властивостями матеріалу. Згортання білка (фолдінг білка) варто відрізняти від проблеми прогнозування структури білка. Згортанням білка називають процес спонтанного згортання поліпептидного ланцюга в унікальну нативну (природну, від англ. *Native*) просторову структуру. У проблеми прогнозування структури білка дослідника цікавить не процес складання (динамічна особливість), а тільки досягнута кінцева структура (статична особливість). Загальні методи пошуку білкових тривимірних структур, таких як рентгенівська кристалографічна і ядерний магнітний резонанс, є повільними

і дорогими, і можуть зайняти до декількох місяців лабораторної роботи. Як наслідок, постійно зростає інтерес до розроблення спеціальних алгоритмів для вирішення проблеми прогнозування структури білка. Є два основних типи обчислювальних стратегій, які використовують сьогодні: на основі знань і ab initio. Гіпотеза, що знаходяться в основі методів, заснованих на знаннях, полягає в тому, що подібні послідовності будуть складатися аналогічно. Ab initio стратегії необхідні, коли немає гомології, так що один змушений складати білки з нуля.

Абстрактні (спрощені) моделі білків часто застосовують під час дослідження найважливіших характеристик їх згортання, а також у процесі розроблення та апробації обчислювальних методів, які дають змогу прогнозувати структуру білка в згорнутому стані. Часто, як абстрактні, виступають моделі на плоских або просторових ґратках. Це дає змогу істотно зменшити кількість можливих конформацій білка завдяки тому, що елементи амінокислотної послідовності можуть розміщуватися тільки у вузлах ґратки, обмежуючи конфігурацію згортки. Проте, навіть у такому спрощеному вигляді задача прогнозування третинної структури є NP-повною [1]. Для її вирішення загалом точні методи виявляються малоефективними.

Об'єкт дослідження – процеси моделювання фолдингу білка на основі амінокислотних послідовностей, передбачення його просторової конфігурації.

Предмет дослідження – еволюційні та імунні методи оптимізації та прогнозування.

Мета роботи – розроблення нового, швидкого та точного методу пошуку, що використовує переваги популяційного підходу та технології гібридизації, заснованих на штучних гібридних імунних системах.

Для досягнення зазначеної мети визначено такі *основні завдання дослідження*:

- аналіз наявних методів прогнозування третинної структури білка;
- розроблення методу кодування індивідуумів на основі методу внутрішніх координат та його програмна реалізація;
- розроблення методу декодування індивідуума та її програмна реалізація;
- розроблення та програмна реалізація цільової функції (функція афінності);
- розроблення гібридного алгоритму в основі алгоритмів клонального відбору та диференціальної еволюції;
- дослідити ефективність розробленого алгоритму.

Наукова новизна отриманих результатів дослідження – розроблено новий гібридний імунний алгоритм клонального відбору та диференційної еволюції для прогнозування третинної структури білка. Запропонований підхід дає змогу підвищити якість HP-моделі на плоскій трикутній ґратці.

Практична значущість результатів дослідження – можливість застосування розробленого методу для вирішення завдання фолдингу білка з використанням HP-моделі на плоскій трикутній ґратці, дають змогу будувати високоякісні HP-моделі на плоскій трикутній ґратці завдяки високій швидкодії та збіжності запропонованого алгоритму.

Аналіз останніх досліджень та публікацій. Прогнозування білкових структур із використанням HP-моделі є складною комбінаторною задачею оптимізації, для якої було показано, що вона є NP-повною. Це означає, що загалом точні методи неспроможні впоратися з нею за прийнятний час. Тому для вирішення завдання фолдингу білка на сьогодні запропоновано цілу низку евристичних алгоритмів, таких як: алгоритми мурашиних колоній [6], штучні імунні системи [3], алгоритми оптимізації методом рою частинок [12], оптимізація табу-пошуком [9] тощо. Проте, аналіз сучасного стану досліджень у цій галузі свідчить про наявність недоліків, пов'язаних із точністю прогнозу, часом, який потрібний для отримання оптимального рішення. Окрім цього, ефективність застосовуваних методів знижується внаслідок того, що поверхня пошуку має багато локальних оптимумів. Це збільшує ймовірність попадання алгоритму в один із таких оптимумів з неможливістю вибратися з нього. Тому перспективною є ідея гібридизації наявних методів з метою об'єднання їх кращих сторін і підвищення ефективності пошуку.

Результати дослідження та їх обговорення

Під час використання ґратчастих моделей оптимальною вважають згортку, що має мінімальну енергію і не має самоперетинів. Згортання послідовності проводиться на плоскій трикутній ґратці, а всі амінокислотні залишки розділені на два класи: *P* – водолюбні (гідрофільні) амінокислоти та *H* – водовідштовхувальні (гідрофобні) амінокислоти [5]. Отже, амінокислотну послідовність можна розглядати як вектор $S = (s_1, s_2, \ldots, s_n)$, $s_i \in \{H, P\}$, $i = \overline{1, n}$. Амінокислотні залишки s_i розташовуються у вузлах ґратки так, що сусідні в послідовності елементи відповідають суміжним вузлам, утворюючи шлях. Отож, в якому вузлі буде знаходитись елемент s_i, залежить від прагнення гідрофобних залишків притягуватися один до одного так, щоб максимально уникати контакту з водою. Між контактуючими гідрофобними залишками виникають *H-H* зв'язки. Енергія одного такого зв'язку дорівнює -1, а загальна енергія структури дорівнює сумі енергій всіх *H-H* зв'язків [7]:

$$E(S) = -\sum_{1 \le j \le j-2 \le n-2} I(N_i, N_j) h(s_i, s_j)$$

$$I(N_i, N_j) = \begin{cases} 1, \textit{if node } N_i \textit{ is adjacent to } N_j \\ 0, \textit{otherwise} \end{cases}, \qquad (1)$$

$$h(s_i, s_j) = \begin{cases} 1, \textit{if } s_i \textit{ and } s_j \textit{ are hydrophobic} \\ 0, \textit{otherwise} \end{cases}$$

де $N_i = (x_i, y_i)$ – координати вузла двомірної трикутної ґратки, в якому знаходиться елемент s_i.

Завдання полягає у виборі такого шляху без самоперетину, який мінімізує енергію згортки, тобто $E(S) \to \min$.

Для кодування згортки (кодування необхідне подання індивідуумів популяційного алгоритму) використано метод внутрішніх координат [3]. Шлях задається послідовністю переміщень від вузла до вузла за тими напрямами, де ґратка має ребро. У роботі застосовується відносне кодування, коли кожний наступний напрямок залежить від попереднього переміщення. На рис. 1 штриховою лінією зі стрілкою показано п'ять можливих напрямків, які утворюють алфавіт, що використовується під час кодування згортки: $A_{rel} = \{LD, LU, F, RU, RD\}$

(відповідно: вліво-вниз, вліво-вгору, вперед, вправо-вгору та вправо-вниз).

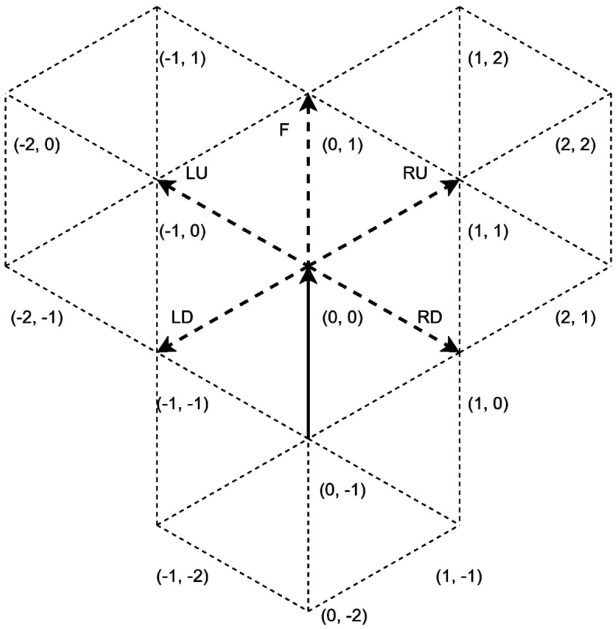

Рис. 1. Двовимірна трикутна ґратка з можливими напрямками під час відносного кодування

З погляду програмної реалізації алгоритму всі можливі напрямки зручно розмістити у таблиці (табл. 1).

Табл. 1. Відносні напрямки переміщень у плоскій трикутній ґратці

Попередній напрямок	Можливі напрямки переміщень				
	LD	LU	F	RU	RD
(-1, -1)	(1, 0)	(0, -1)	(-1, -1)	(-1, 0)	(0, 1)
(-1, 0)	(0, -1)	(-1, -1)	(-1, 0)	(0, 1)	(1, 1)
(0, -1)	(1, 1)	(1, 0)	(0, -1)	(-1, -1)	(-1, 0)
(0, 1)	(-1, -1)	(-1, 0)	(0, 1)	(1, 1)	(1, 0)
(1, 0)	(0, 1)	(1, 1)	(1, 0)	(0, -1)	(-1, -1)
(1, 1)	(-1, 0)	(0, 1)	(1, 1)	(1, 0)	(0, -1)

При цьому рядок індивідуума кодується зміщеннями у цій таблиці (рис. 2).

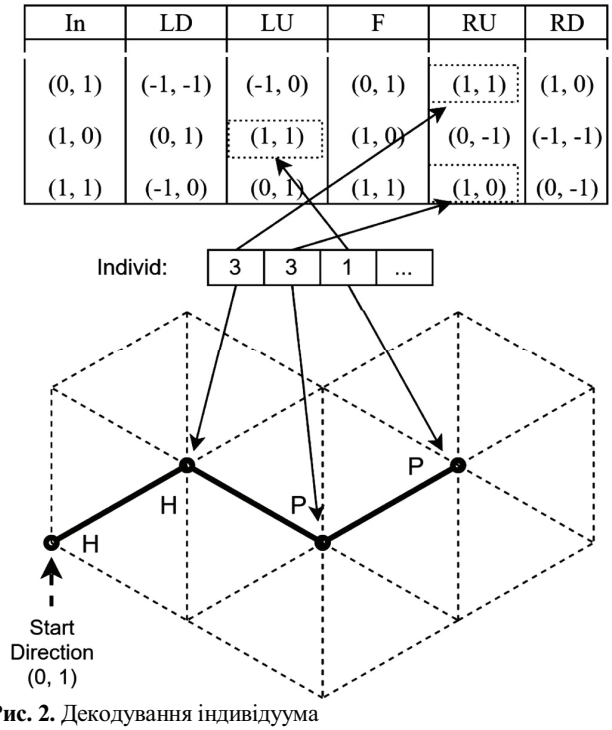

Рис. 2. Декодування індивідуума

Під час формування шляху (декодування індивідуума) деякі напрямки можна ігнорувати, якщо вони призводять до самоперетину. Отже, вилучається можливість побудови некоректного шляху. Якщо всі можливі напрямки призводять до самоперетину, то побудова шляху закінчується. Виходить неповна послідовність, що враховується під час її оцінювання у вигляді штрафних санкцій. Індивідууми оцінюються за допомогою наступної функції афінності:

$$f = \frac{1}{\left|E^*(S)\right| + l^*}, f \to \min, \quad (2)$$

де: $E^*(S)$ – енергія тієї частини HP-послідовності S, яку вдалося декодувати; l^* – довжина декодованої частини послідовності.

Представлений у роботі гібридний алгоритм комбінує переваги штучної імунної системи у формі алгоритму клонального відбору (CLONALG) [4] та диференціальної еволюції (ДЕ) [10].

У клональному алгоритмі значення афінності виражають міру близькості індивідуума до оптимального рішення та обчислюються на підставі наступної цільової функції:

$$f(\bar{x}) \to \min, \bar{x} = (x_1, \ldots, x_l), \quad (3)$$

де: $\bar{x}$ – вектор параметрів задачі, за яким будуються індивідууми популяції рішень $\bar{x}_i^G, i = 1, \ldots, P$; P – розмір популяції рішень; G – поточна ітерація (покоління).

Основною "рушійною силою" в CLONALG є мутація, представлена простою схемою, згідно з якою значення кожного елемента рядка антитіла змінюється випадковим чином (із заданою ймовірністю p_m) за такою формулою:

$$x_{ij}^{G+1} = \begin{cases} randInit(), if\ randEvent(p_m) = 1 \\ x_{ij}^G, otherwise \end{cases}, j = 1,2,\ldots,l, \quad (4)$$

де: x_{ij} – j-ий елемент i-го індивідуума у популяції; l – довжина рядка індивідуума; p_m – ймовірність зміни елемента рядка, що названо тут інтенсивністю мутації; *rndInit*() – функція випадкової ініціалізації елемента рядка індивідуума; *rndEvent*() – бінарна функція генерації випадкової події із заданою ймовірністю.

На відміну від інших еволюційних алгоритмів, мутація у ДЕ виглядає так:

$$\bar{v}_i^{G+1} = \bar{x}_{r_3}^G + F\left(\bar{x}_{r_1}^G - \bar{x}_{r_2}^G\right), \quad (5)$$

де: $\bar{v}_i^{G+1}, i = 1, \ldots, P$ – індивідуум, отриманий внаслідок мутації; $r_1, r_2, r_3 \in \{1, \ldots, P\}$ – індекси індивідуумів, які вибираються випадково з популяції рішень у поточному поколінні, такі, що $r_1 \neq r_2 \neq r_3 \neq i$; F – коефіцієнт масштабування ($F \geq 0$).

Не всі компоненти вектора $\bar{v}_i^{G+1}$ переходять до нового покоління. Формування кандидатів відбувається за таким виразом:

$$x_{ij}^{G+1} = \begin{cases} v_{ij}^{G+1}, if\ randEvent\left(p_{DE}\right) = 1 \vee j = k \\ x_{ij}^G, otherwise \end{cases}, j = 1,2,\ldots,l, \quad (6)$$

де: $k \in \{1, \ldots, l\}$ – випадковий індекс параметра, який вибирається одноразово для кожного індивідуума, сенс якого в тому, щоб гарантувати перехід хоча б одного компонента вектора $\bar{v}_i^{G+1}$ у вектор $\bar{x}_i^{G+1}$; p_{DE} – ймовірність переходу j-го компонента вектора $\bar{v}_i^{G+1}$ у вектор $\bar{x}_i^{G+1}$.

Псевдокод гібридного алгоритму CLONALG і ДЕ представили на рис. 3. $Ab^{(G)}$ – основна популяція антитіл, що має розмір P. Її розмір підтримується постійним у кожному поколінні G. У нульовому поколінні (G=0) основна популяція $Ab^{(0)}$ генерується випадково за допомогою процедури *initPop*(). Афінність антитіл $\overline{x}_i^G \in Ab^{(G)}$ повертається процедурою *evaluate* (), яка проводить обчислення на підставі функції афінності (2). Обчислення афінності містить декодування антитіл (рис. 2) та побудову конформацій протеїну на ґратці з подальшим розрахунком енергії.

У цій роботі зупинка алгоритму *terminationCondition*() відбувається під час виконання однієї з умов: для тестових завдань якщо знайдено мінімальне значення енергії ($E^{\min}$), то алгоритм зупиняється та фіксується кількість пусків процедури *evaluate*(); якщо мінімальне значення енергії не може бути знайдено протягом $G^{\max}$ поколінь, то алгоритм зупиняється після досягнення максимальної кількості поколінь.

Тіло основного циклу починається процедурою клонування *cloning*() антитіл, що мають високу афінність до антигену (тобто низькі значення функції афінності). Антитіла для клонування відбираються за допомогою турніру, під час якого з популяції випадково вибираються k_{Ab} індивідуумів (k_{Ab} – розмір турніру) і порівнянням їх афінностей вибирається переможець, що переходить у проміжну популяцію клонів $Ab^{(C)}$ розміру P_C.

```
1:  ClonalgDE Algorithm( P, P_c, k_Ab, p_c, p_m, p_DE, F)
2:  G := 0;
3:  Ab(G) := initPop(P);
4:  evaluate(Ab(G));
5:  while(!terminationCondition()) do
6:      Ab(C) := cloning(Ab(G), P_c);
7:      Ab(M) := mutation(Ab(C), p_c, p_m);
8:      Ab(DE) := deMutation(Ab(m), 1-p_c, p_DE, F);
9:      evaluate(Ab(DE));
10:     Ab(G+1) := deSelection(Ab(DE), k_Ab);
11:     G := G + 1;
12: end_while
```

Рис. 3. Псевдокод гібридного алгоритму клонального відбору та диференційної еволюції

Мутація впливає на популяцію клонів двома способами. Спочатку виконується проста мутація *mutation*() за виразом (4). Імовірність її застосування до індивідуума дорівнює $p_c!\sim f$, її інтенсивність – $p_m!\sim f$. Тут зворотно пропорційна залежність від значень функції афінності визначається тим фактом, що антитіла, що знаходяться ближче до оптимального рішення, мутують менше, ніж ті, які розташовані далі від оптимуму. Ті антитіла, для яких не була застосована проста мутація, піддаються ДЕ-мутації *deMutation*(). ДЕ-мутація виконується за формулами (5) та (6).

Після всіх змін популяція клонів оцінюється за допомогою процедури *evaluate* ($Ab^{(DE)}$). При цьому в основну популяцію переходить тільки частина антитіл популяції клонів, які вибираються за такою формулою:

$$\overline{x}_i^{G+1} = \begin{cases} \overline{u}_i^{G+1}, if\ f\left(\overline{u}_i^{G+1}\right) \le f\left(\overline{x}_i^G\right) \\ \overline{x}_i^G, otherwise \end{cases}, \qquad (7)$$

де $\overline{u}_i^{G+1} \in Ab^{(DE)}$ – індивід після мутації.

Обговорення результатів дослідження. Експериментальні дослідження проводили з використанням восьми стандартних тестових послідовностей, які часто трапляються в літературних джерелах як зразок під час перевірення продуктивності методів оптимізації, що розробляються [8]. Структура послідовностей та їхня довжина наведені у табл. 2. З метою скорочення запису в *hp*-послідовності додані індекси до типів залишків h_i, p_i та їх груп $(\ldots)_i$. Індекси i означають кількість повторень відповідного символу чи групи символів.

Табл. 2. Тестові послідовності для HP-моделі

№	Довжина	Послідовність
1	20	$hphp_2h_2php_2hph_2p_2hph$
2	24	$h_2p_2(hp_2)_6h_2$
3	25	$p_2hp_2(h_2p_4)_3h_2$
4	36	$p_3h_2p_2h_2p_5h_7p_2h_2p_4h_2p_2hp_2$
5	48	$p_2h(p_2h_2)_2p_5h_{10}p_6(h_2p_2)_2hp_2h_5$
6	50	$h_2(ph)_3ph_4p(hp_3)_2hp_4(hp_3)_2hph_4(ph)_3ph_2$
7	60	$p_2h_3ph_8p_3h_{10}php_3h_{12}p_4h_6ph_2php$
8	64	$h_{12}(ph)_2(p_2h_2)_2p_2h\,(p_2h_2)_2p_2h\,(p_2h_2)_2p_2(hp)_2h_{12}$

Значення параметрів гібридного алгоритму наведено у табл. 3.

Табл. 3. Параметри гібридного алгоритму CLONALG + ДЕ

Назва параметра	Значення параметра
Спосіб кодування індивідуумів	дійсні числа
Кількість поколінь (G^{max})	500
Коефіцієнт відбору (q/P)	0.9
Тип відбору	турнірний
Розмір турніру (k_Ab)	5
Умова зупинки	кількість поколінь або знайдений E^{min}
Розмір основної популяції (P)	3000
Розмір популяції клонів (P_C)	9000
Ймовірність простої мутації (pc)	0.1
Інтенсивність простої мутації (p_m)	0.05
Інтенсивність ДЕ-мутації (p_{DE})	1.0
Коефіцієнт масштабування (F)	0.8
ДЕ-селекція	(7) від'єдано, тільки турніри

Для кожної із восьми послідовностей проведено 30 незалежних пусків алгоритму. Отримані кращі та середні результати порівняли з аналогічними результатами, отриманими іншими обчислювальними методами [2], і навели у табл. 4.

Порівняно із запропонованим алгоритмом (CLONALG + DE) беруть участь:

- HHGA (англ. *Hybrid of Hill-Climbing and Genetic Algorithm*) – гібридний генетичний алгоритм з операторами алгоритмів в гору [11];
- IMOG (англ. *Hybrid of Ions Motion Optimization With a Greedy Algorithm*) – гібридний алгоритм оптимізації руху іонів за допомогою жадібного алгоритму [13];
- GALSTS (англ. *Hybrid of: Genetic Algorithm, Local Search, Tabu Search Strategy*) – гібридний алгоритм, який об'єднує генетичний алгоритм, локальний пошук і стратегію пошуку табу [2].

Результати із поміткою "Кращий" демонструють мінімальні значення енергії, досягнуті за 30 пусків, тоді як результати із поміткою "Середній" демонструють стійкість алгоритму до досягнення мінімумів. Як видно з таблиці, розроблений алгоритм у перших п'яти тестах

досяг таких самих мінімальних показників енергії, як і інші обчислювальні методи, а в тестах 6, 7 і 8 були виявлені нові глобальні мінімуми. Що ж до стійкості, то й тут гібридний алгоритм демонструє перевагу, показуючи вищі результати.

Табл. 4. Порівняльні результати тестів розробленого алгоритму

№	HHGA		IMOG		GALSTS		CLONALG+DE	
	кращий	середній	кращий	середній	кращий	середній	кращий	середній
1	-15	-14.73	-15	-14.73	-15	-14.86	-15	-15.00
2	-17	-14.93	-17	-14.93	-17	-15.53	-17	-16.83
3	-12	-11.57	-12	-11.57	-12	-12	-12	-11.95
4	-23	-21.27	-23	-21.27	-24	-21.93	-24	-23.18
5	-41	-37.30	-41	-37.30	-43	-39.86	-43	-40.74
6	-38	-34.10	-38	-34.10	-40	-37.6	-41	-39.30
7	-66	-1-61.83	-66	-61.83	-70	-68.26	-71	-68.12
8	-63	-56.53	-63	-56.53	-67	-58.46	-72	-70.22

Висновок

У роботі запропонували гібридний метод та алгоритм вирішення завдання фолдингу білка з використанням HP-моделі на плоскій трикутній ґратці. Метод заснували на комбінації штучної імунної системи у формі алгоритму клонального відбору та алгоритму диференціальної еволюції. Відповідно до запропонованого підходу, до клонального алгоритму додали оператор ДЕ-мутації. Це дало змогу значно поліпшити пошукові можливості клонального алгоритму і прискорити його збіжність. З метою перевірки ефективності запропонованого підходу виконали експерименти на тестових *hp*-послідовностях. Порівняльний аналіз результатів роботи запропонованого алгоритму на тестових послідовностях з аналогічними результатами інших опублікованих методів дає змогу зробити висновок про високу ефективність розробленого методу. Зокрема, вдалося досягти більшої стійкості результату і в деяких випадках отримати конформації, які мають меншу енергію.

References

[1] Berger, B., & Leighton, T. (1998). Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. of Computational Biology, 5*(1), 27-40. https://doi.org/10.1089/cmb.1998.5.27

[2] Boumedine, N., & Bouroubi, S. (2021). A new hybrid genetic algorithm for protein structure prediction on the 2D triangular lattice. Turkish J. Electr. Eng. Comput. Sci., 29, 499-513. https://doi.org/10.3906/elk-1909-31

[3] Cutello, V., Niscosia, G., Pavone, M., & Timmis, J. (2007). An immune algorithm for protein structure prediction on lattice models. *IEEE Transactions on Evolutionary Computation, 11*(1), pp. 101–117. https://doi.org/10.1109/TEVC.2006.880328

[4] De Castro, L. N., & Von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation, 6*(3), 239–251. https://doi.org/10.1109/TEVC.2002.1011539

[5] Dill, K. A. (1985). "Theory for the folding and stability of globular proteins," *Biochemistry, 24*(6), 1501-1509. https://doi.org/10.1021/bi00327a032

[6] Fidanova, S., & Lirkov, I. (2008). Ant colony system approach for protein folding, *2008* International Multiconference on Computer Science and Information Technology, 887–891. https://doi.org/10.1109/IMCSIT.2008.4747347

[7] Gulyanitskiy, L. F., & Rudyik, V. A. (2010). Simulation of protein coagulation in space. *Computer mathematics, 1*, 128-137. [In Russian].

[8] Krasnogor, N., Blackburne, B. P., Burke, E. K., & Hirst, J. D. (2002). Multimeme algorithms for protein structure prediction. In *Proc. Int. Conf. Parallel Problem Solving from Nature* (PPSN VII), *Granada, Spain, Sep*. 2002, 769-778. https://doi.org/10.1007/3-540-45712-7_74

[9] Liu, J., Sun, Y., Li, G., Song, B., & Huang, W. (2013). Heuristic-based tabu search algorithm for folding two-dimensional AB off-lattice model proteins. *Computational biology and chemistry, 47*, 142-148. https://doi.org/10.1016/j.compbiolchem.2013.08.011

[10] Storn R., & Price, K. V. (1997). Price Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization, 11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

[11] Su, S. C., Lin, C. J. & Ting, C. K. (2011). An effective hybrid of hill climbing and genetic algorithm for 2D triangular protein structure prediction. *Proteome Sci,* 9, S19. https://doi.org/10.1186/1477-5956-9-S1-S19

[12] Yang, C. H., Lin, Y. S., Chuang, L. Y., & Chang, H. W. (2017). A Particle Swarm Optimization-Based Approach with Local Search for Predicting Protein Folding. *Journal of computational biology: a journal of computational molecular cell biology, 24*(10), 981-994. https://doi.org/10.1089/cmb.2016.0104

[13] Yang, C. H., Wu, K. C., Lin, Y. S. *et al.* (2018). Protein folding prediction in the HP model using ions motion optimization with a greedy algorithm. *BioData Mining,* 11, 17. https://doi.org/10.1186/s13040-018-0176-6

I. M. Fefelova, V. I. Lytvynenko, A. O. Fefelov

Kherson National Technical University, Kherson, Ukraine

PREDICTION OF THE TERTIARY STRUCTURE OF A PROTEIN ON A TWO-DIMENSIONAL TRIANGULAR LATTICE BY A HYBRID EVOLUTIONARY ALGORITHM

This work discusses the problem of forecasting the tertiary structure of a protein, based on its primary sequence. The problem is that science, with all its computing power and a set of experimental data, has not learned to build models that describe the process of protein molecule coagulation and predict the tertiary structure of a protein, based on its primary structure. However, it is wrong to assume that nothing is happening in this field of science. The regularities of folding (convolution) of the protein are known, methods for its modelling have been developed. Analysis of the current state of research in the field of these problems indicates the presence of shortcomings associated with the accuracy of forecasting and the time necessary to obtain the optimal solution. Consequently, the development of new computational methods, deprived of these shortcomings, seems relevant. In this

work, the authors focused on the lattice model, which is a special case of the known hydrophobic-polar dill. protein conformation according to the chosen model, hybrid algorithms of cloning selection, differential are proposed. Since the processes of protein coagulation have not been fully understood, the researchers proposed several simplified models based on the physical properties of molecules and which leads to problems of combinatorial optimization. A hydrophobic-polar simplified model on the planar triangular lattice is chosen as a protein model. From the point of view of the optimization problem, the problem of protein folding comes down to finding a conformation with minimal energy. In lattice models, the conformation is represented as a non-self-cutting pathway. A hybrid artificial immune system in the form of a combination of clonal selection and differential evolution algorithms is proposed to solve this problem. The paper proposes a hybrid method and algorithm to solve the protein folding problem using the HP model on a planar triangular lattice. In this paper, a hybrid method and algorithm for solving the protein folding problem using the HP model on a planar triangular lattice are proposed. The developed hybrid algorithm uses special methods for encoding and decoding individuals, as well as the affinity function, which allows reducing the number of incorrect conformations (self-cutting solutions). Experimental studies on test hp-sequences were conducted to verify the effectiveness of the algorithm. The results of these experiments showed some advantages of the developed algorithm over other known methods. Experiments have been taught to verify the effectiveness of the proposed approach.

The results labelled "Best" show the minimum energy values achieved over 30 runs, while the results labelled "Medium" show the robustness of the algorithm to achieve minima. Regarding robustness, the hybrid algorithm also offers an advantage, showing higher results. A comparative analysis of the performance results of the proposed algorithm on test sequences with similar results of other published methods allows us to conclude the high efficiency of the developed method. In particular, the result is more stable, and, in some cases, conformations with lower energy are obtained.

Keywords: protein folding; hydrophobic-polar model; clonal selection; differential evolution; artificial immune systems; hydrophobic-polar model.

Інформація про авторів:

Фефелова Ірина Михайлівна, аспірантка, кафедра інформатики і комп'ютерних наук. **Email:** fim2019@ukr.net; http://orcid.org/0000-0002-3857-3811

Литвиненко Володимир Іванович, д-р техн. наук, професор, завідувач кафедри інформатики і комп'ютерних наук. **Email:** immun56@gmail.com

Фефелов Андрій Олександрович, канд. техн. наук, доцент, кафедра інформатики і комп'ютерних наук. **Email:** fao1976@ukr.net; http://orcid.org/0000-0003-1140-0985

Цитування за ДСТУ: Фефелова І. М., Литвиненко В. І., Фефелов А. О. Прогнозування третинної структури білка на двомірній трикутній ґратці гібридним еволюційним алгоритмом. *Український журнал інформаційних технологій.* 2021, т. 3, № 2. С. 27–32.

Citation APA: Fefelova, I. M., Lytvynenko, V. I., & Fefelov, A. O. (2021). Prediction of the tertiary structure of a protein on a two-dimensional triangular lattice by a hybrid evolutionary algorithm. *Ukrainian Journal of Information Technology, 3*(2), 27–32. https://doi.org/10.23939/ujit2021.02.027