Теорія оптимальних рішень

УДК 51-76

В.В. ГОРІН

ГОМОЛОГІЧНЕ МОДЕЛЮВАННЯ ТРЕТИННОЇ СТРУКТУРИ БІЛКІВ

Гомологіче моделювання базується на знаходженні білків, первинна структура яких схожа на структуру студійованого білка, і на створенні процедури групування. Остання може бути глобальною і локальною. Перша – це процедура глобальної оптимізації, яка намагається згрупувати кожну амінокислоту з кожною. Локальні групування виявляють схожі регіони всередині довгих послідовностей. Для розв'язання проблеми групування, використовуються точні методи, такі як динамічне програмування, та швидкі евристичні алгоритми або ймовірнісні методи.

Вступ. Передбачення трьохвимірної структури білка на основі послідовності його амінокислот, іншими словами, передбачення третинної структури білка на основі його первинної структури – це одна з найважливіших задач, яка стоїть перед біоінформатикою та теоретичною хімією. Незважаючи на сучасний прогрес експериментальної техніки, визначення структурних та динамічних властивостей білків – дуже трудомістка задача, що відіймає немало часу [1]. Тому комп'ютерне моделювання відіграє ключову роль у розв'язанні таких задач.

Водночас, практична роль моделювання структури білків сьогодні є важливішою ніж будь-коли. Величезні обсяги даних первинних структур білків створюються сьогодні великими, сучасними проектами по вивченню послідовностей ДНК, такими як, наприклад, проект геному людини. Незважаючи на великі зусилля в області структурної геноміки, швидкість отримання третинних структур (як правило, експериментальними методами) дуже сильно відстає від швидкості отримання первинних структур. Оптимальним вирішенням проблеми відставання отримання третинних структур від первинних є побудова моделей та обчислювальних систем, здатних передбачувати третинну структуру білків на основі існуючих структурних даних уже вивчених білків.

Існує ряд факторів, які роблять передбачення структури білка дуже непростою задачею. Дві основні проблеми – це надзвичайно велика кількість можливих третинних

структур, і те, що фізичні основи структурної стабільності білка досі не відомі в повній мірі. Як наслідок, будь-який метод передбачення третинної структури повинен мати механізм, що дозволяв би вивчати простір можливих структур ефективно (стратегія пошуку) та спосіб ідентифікації найбільш правдоподібних структур (функція енергії).

При порівняльному передбаченні (також відомому як гомологічне моделювання), пошук скорочується шляхом припущення, що студійований білок частково запозичує свою структуру в хоча б одного відомого білка. В методах «de novo» або «ab initio» (без використання відомих структур) таке припущення не використовується, і як результат задача стає набагато важчою. В обох випадках функція енергії потрібна для того, щоб вибрати найбільш ймовірну структуру. Нажаль, пошук цієї функції є у великій мірі відкритою проблемою.

Гомологічне моделювання, по-перше, базується на знаходженні одного або декількох відомих структур білків (структур-зразків), первинна структура яких схожа на структуру студійованого білка, по-друге – на створенні процедури групування (alignment), яка зіставляє залишки амінокислотної послідовності студійованого білка із залишками у структурі зразка. Процедура групування і зразки використовуються потім для створення структурної моделі студійованого білка. У зв'язку з тим, що білкові структури є більш консервативними ніж структури ДНК, знайдені подібності у первинній структурі білків, як правило, означають суттєві подібності у третинній структурі [2].

Якість гомологічної моделі залежить від якості процедури групування та від самої структури зразка (ступені її подібності до структури студійованого білка). Гомологічний підхід може бути ускладнений наявністю, по-перше, інделів (indels – групувальні шпарини, вставки або пропуски), які представляють собою структурну область, яка присутня в шуканій структурі і відсутня у структурі зразка, і, по-друге – структурних шпарин у структурі зразка на тій чи іншій ділянці, які виникають внаслідок низької роздільної здатності експериментального методу (наприклад, рентгенівської кристалографії). Регіони моделі, сконструйовані без використання зразку, як правило методом циклічного моделювання, є менш точними. Помилки у розрахунках структури заступних груп і їх розташування також зростають із зменшенням ступені ідентичності між студійованою структурою і зразком; є припущення, що різні види конфігурацій цих угрупувань – основна причина низької якості моделі при низьких ступенях ідентичності конфігурацій [3].

Гомологічне моделювання може дати структурні моделі високої якості у випадку, коли шукана структура і структура зразка дуже схожі. Це стало спонукальним мотивом для створення консорціуму структурної геноміки, присвяченого створенню експериментальних структур типових класів білків.

Процедура гомологічного моделювання може бути розбита на чотири етапи:

вибір структури-зразка, групування шуканої структури до структури зразка, побудова моделі та оцінка її якості [2].

Вибір структури-зразка, процедура групування. Перший вирішальний крок у гомологічному моделюванні – це знаходження найкращої структури-зразка, якщо взагалі така доступна. Найпростіший спосіб ідентифікації структури-зразка базується на попарних групуваннях послідовності амінокислот первинної структури студійованого та структур потенційних зразків. Пошукові методи FASTA, BLAST, та інші, серед яких одночасне групування декількох структур (наприклад, PSI-BLAST), метод розпізнавання згорток білків (protein threading), тут дуже корисні [4–7].

Якщо дві послідовності при групуванні мають спільне походження, їх розбіжності можуть бути трактовані як точкові мутації, а шпарини – як індели (indels – вставки чи пропуски), що з часом виникли в одній або в обох послідовностях, оскільки вони обидві пішли одна від одної, або від спільного предка. При вирівнюванні білків, рівень схожості між амінокислотами послідовностей може бути розцінений як груба оцінка того, на скільки консервативним щодо мутацій є той чи інший регіон послідовності.

Дуже короткі або дуже схожі послідовності можуть бути згруповані вручну. Проте, найбільша кількість цікавих проблем вимагає групування довгих або багатьох послідовностей, що дуже відрізняються одна від одної, і не можуть бути згруповані вручну.

Локальне та глобальне групування. Обчислювальні підходи до вирівнювання послідовностей, як правило, поділяють на дві категорії: глобальні та локальні. Розрахунок глобального вирівнювання – це різновид глобальної оптимізації, яка «змушує» групування розповсюджуватись на всю довжину послідовності (амінокислот первинної структури). Напроти, локальні групування виявляють схожі регіони всередині довгих послідовностей, які часто несхожі загалом. Локальні групування дають завжди більш якісну картину, але можуть бути більш складними з точки зору розрахунків через необхідність розв'язувати додаткову задачу із виявлення схожих регіонів. Велика кількість алгоритмів була використана для розв'язання проблеми групування послідовностей, включаючи повільні методи, такі як динамічне програмування, та швидкі, але не такі точні евристичні алгоритми або імовірнісні методи, спроектовані для широкомасштабного пошуку за базами даних.

Глобальні вирівнювання, які прагнуть знайти пару кожному залишку в кожній послідовності, найбільш продуктивні, коли послідовності в наборі схожі одна на одну і приблизно однакової довжини. Типовим прикладом глобального вирівнювання є алгоритм Нідлмана – Вунча (Needleman – Wunsch), який базується на динамічному програмуванні [8]. Локальні вирівнювання найбільш підходять для несхожих у цілому послідовностей, які, як очікується, мають схожі регіони. Прикладом методу локального групування може слугувати алгоритм

Сміта – Вотермана (Smith – Waterman) і також базується на динамічному програмуванні [9]. У випадку достатньо схожих послідовностей немає принципової різниці між локальним та глобальним вирівнюваннями.

Гібридні методи, також відомі як напівглобальні, намагаються знайти найкраще можливе вирівнювання, що включає у себе початок і кінець однієї чи іншої послідовності. Це може бути особливо корисно коли нижній потік однієї послідовності перекривається з верхнім потоком іншої. У цьому випадку, ні глобальне, ні локальне вирівнювання не є підходящим: глобальне вирівнювання буде намагатися розширити себе за межі регіону перекриття, водночас, як локальне може не охопити цей регіон повністю [10].

Попарне групування. Методи попарного групування використовуються для знаходження схожих ділянок лише двох послідовностей. Ці методи ефективні та часто використовуються для випадків коли не потрібна дуже висока точність (наприклад, пошук послідовностей з високим ступенем гомологічності до даної). Основні три методи попарного групування – це растрові (dot-matrix) методи, динамічне програмування, та «словесні» методи (word methods) [11]. Хоча всі три методи мають свої сильні і слабкі сторони, всі вони мають проблеми з послідовностями, що мають ділянки з низьким рівнем інформації, які часто повторюються, особливо, коли кількість таких ділянок неоднакова у двох послідовностей, що групуються.

Растровий підхід – якісний і простий, але повільний, якщо послідовність велика. Щоб сконструювати растр, дві послідовності записуються одна в горизонтальний ряд зверху двовимірної матриці, інша у вертикальний – зліва від неї. Матриця заповнюється крапками у клітинах, що знаходяться на перетині колонок і рядків з однаковими літерами. Растри схожих послідовностей будуть виглядати як лінія, яка проходить неподалік від основної діагоналі матриці.

Техніка динамічного програмування може бути застосована для створення глобальних групувань за методом Нідлмана – Вунча і локальних групувань за методом Сміта – Вотермана. Як правило, використовується матриця підстановок щоб призначити питому вагу фактам збіжності чи розбіжності амінокислот у послідовності й штраф на прогалини для сполучення амінокислоти в одній послідовності з прогалиною в іншій. Загальноприйнятим розширенням стандартних лінійних штрафів на прогалини – використання двох різних штрафів на відкриваючу прогалину (першу) і на її продовження (прогалини, що слідують зразу за першою). Як правило, перший набагато більший за другий. Це має сенс з біологічної точки зору, бо прогалини і залишки при використанні такої моделі тримаються поруч.

«Словесні» методи, також відомі як k-tuple методи – це евристичні методи, які не гарантують знаходження оптимального розв'язку задачі групування, але працюють набагато швидше ніж методи динамічного програмування. Вони особливо корисні для пошуку за великими базами даних, де зрозуміло, що переважна більшість послідовностей-кандидатів не матиме суттєвих сполучень із студі-

йованою послідовністю. «Словесні» методи реалізовані в таких алгоритмах як FASTA і BLAST. Ці методи виявляють серії коротких підпослідовностей, що не перекриваються («слів») у студійовані послідовності, які потім зіставляються з послідовностями-кандидатами із баз даних.

Метод FASTA дозволяє задати параметр k – довжину слова, яка використовується при пошуку по базі. Метод працює повільніше і більш точно при менших значеннях k. FASTA швидкий і селективний; методи FASTP і FASTA були розроблені для знаходження білкових послідовностей, які походять від спільного предка і вони довели свою виняткову корисність для цієї задачі [4]. Метод BLAST був розроблений щоб надати більш швидкий альтернативний алгоритм з меншою точністю; так само, як і FASTA, BLAST використовує «словесний» пошук довжини k, але оцінює тільки найбільш значимі збіги слів, на відміну від FASTA. Більша частина реалізацій BLAST використовує фіксовану стандартну довжину слова, яка оптимізована під запит і тип бази даних, і змінюється тільки в залежності від деяких особливих обставин, таких як пошук дуже коротких послідовностей [7]. Реалізації цих алгоритмів можуть бути знайдені в мережі Інтернет, наприклад EMBL FASTA, NCBI BLAST, EMBOSS-Align та ін [5, 6].

Оцінка ваги гомологічного моделювання. Вирівнювання послідовностей є корисним у біоінформатиці для з'ясування ступеню схожості послідовностей амінокислот, для створення пілогенетичних дерев [12] та для розробки гомологічних моделей структури білків. Тим не менше, біологічна значимість вирівнювання послідовностей не завжди чітко зрозуміла. Вирівнювання часто розглядаються як спосіб відображення еволюційних змін, що відбулися у двох білків, які мають спільного предка; але формально можливо, також, що конвергентна еволюція може спричинити до виникнення схожостей у білків, що не мають спільного предка, але виконують схожі функції [13].

При використанні пошукових систем по базам даних, таких як BLAST, результати вирівнювання можуть відрізнятися в залежності від складу бази даних, по якій проводиться пошук. Ймовірність знаходження хорошого вирівнювання тим збільшується, якщо база даних складається тільки з послідовностей того ж самого організму, що й шукана послідовність. Послідовності, які повторюються (в базі чи в запиті), також можуть викривляти отримані результати або оцінку їх важливості; BLAST автоматично фільтрує такі послідовності-повтори у запиті щоб уникнути видимих збігів, які є штучними (статистичними артефактами). Методи оцінки статистичної ваги для вирівнювання послідовностей з прогалинами доступні в літературі, наприклад в [14].

Функції підрахунку. Вибір функції підрахунку ступені подібності двох послідовностей (біологічних чи статистичних) є дуже важливим для побудови якісних вирівнювань. Білкові послідовності часто вирівнюються за допомогою підстановочних матриць, які відображають ймовірності даних «символ-до-сиволу» замін. Ряд матриць PAM (Point Accepted Mutation matrices – матриці точкових

мутацій, уперше визначені Маргаретою Дейхоф; інколи їх називають матрицями «Дейхов») явно кодують еволюційні апроксимації щодо частот та ймовірностей мутацій окремих амінокислот [15]. Інший розповсюджений клас матриць підрахунку, відомий як BLOSUM (Blocks Substitution Matrix – матриця заміни блоків), кодує ймовірності заміни, отримані емпірично. Варіації обох видів матриць використовуються для знаходження послідовностей з різним рівнем невідповідності, таким чином, даючи можливість користувачам BLAST або FASTA обмежити пошуки до більш близьких степеней відповідності або навпаки розширити пошук для знаходження більш далеких за схожістю послідовностей [16].

Дуже корисним може бути використання декількох матриць для одного вирівнювання. Порівнюючи результати, можна визначити регіони, в яких знайдене рішення є неточним або не єдиним – це регіони де вирівнювання дуже відрізняються в залежності від використаних матриць оцінки та їх параметрів.

Висновки. Використання вищезгаданих методів не вимагає наявності бази білкових структур або технічної наявності доступу до неї, оскільки багато з них впроваджені й доступні як відкриті інтернет-сервери розпізнавання структур. Використання таких серверів дозволяє скоротити зусилля на пошук структур-зразків і зосередитись на побудові моделі білків. Одним з таких серверів є FASTA-сервер Європейського Інституту Біоінформатики. За його допомогою планується створити алгоритм для побудови третинних білкових структур на базі первинних. Очікується, що алгоритм матиме хорошу точність для структур, близьких за походженням до вивчених нині експериментально.

Послідуючим кроком може бути вдосконалення алгоритму шляхом використання інших методів попарного групування, також доступних як відкриті Web-сервіси (наприклад, EMBOSS-Align). Також планується використання методів структурного вирівнювання на базі вторинних структур білків.

В.В. Горин

ГОМОЛОГИЧЕСКОЕ МОДЕЛИРОВАНИЕ ТРЕТИЧНОЙ СТРУКТУРЫ БЕЛКОВ

Гомологическое моделирование базируется на нахождении белков, первичная структура которых похожа на структуру изучаемого белка, и на создании процедуры группирования. Последняя может быть глобальной и локальной. Первая – это процедура глобальной оптимизации, стремящаяся сгруппировать каждую аминокислоту с каждой. Локальные группирования выявляют похожие регионы внутри длинных последовательностей. Для решения задачи группирования используются точные методы, такие как динамическое программирование и быстрые эвристические алгоритмы или вероятностные методы.

V.V. Gorin

HOMOLOGICAL MODELING OF TERTIARY PROTEIN STRUCTURE

Homology modeling relies on the identification of protein structures likely to resemble the structure of the query sequence, and on the production of an alignment. Alignment can be global or local. Global alignment is a procedure of global optimization and attempts to align every residue in every sequence. Local alignments reveal similar subsequences inside long amino acid sequences. Precise methods are used to make alignments; among them are slow dynamic programming and efficient heuristic algorithms or probabilistic methods.

1. *Slabinski L., Jaroszewski L., Rodrigues A.P.C. and other.* The challenge of protein structure determination – lessons from structural genomics // Protein Sci. – 2007. – **16** (11). – P. 2472–2482.
2. *Marti-Renom M.A., Stuart A.C., Fiser A. and other.* Comparative protein structure modeling of genes and genomes // Annu Rev Biophys Biomol Struct. – 2000. – **29.** – P. 291–325.
3. *Chung S.Y., Subbiah S.* A structural explanation for the twilight zone of protein sequence homology // Structure. – 1996. – **4** (10). – P. 1123–1127.
4. *Pearson W.R.* Rapid and Sensitive Sequence Comparison with FASTP and FASTA // Methods in Enzymology. – 1990. – **183.** – P. 63–98.
5. *Pearson W.R., Lipman D.J.* Improved Tools for Biological Sequence Comparison // Proceedings of the National Academy of Sciences. – 1988. – **85** (8). – P. 2444–2448.
6. *Altschul S.F., Madden T.L., Schaffer A.A. and other.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Research. – 1997. – **25** (17). – P. 3389–3402.
7. *Altschul S.F., Warren G., Miller W. And other.* Basic local alignment search tool // J. Mol. Biol. – 1990. – **215.** – P. 403–410.
8. *Needleman S.B., Wunsch C.D.* A general method applicable to the search for similarities in the amino acid sequence of two proteins // J. Mol. Biol. – 1970. – **48 (3)**. – P. 443–53.
9. *Smith T.F., Waterman M.S.* Identification of Common Molecular Subsequences // J. Mol. Biol. – 1981. – **147.** – P. 195–197.
10. *Brudno M., Malde S., Poliakov A. and other.* Glocal alignment: finding rearrangements during alignment // Bioinformatics. – 2003. – **19**, Suppl. 1 – P. 54–62.
11. *Mount D.M.* Bioinformatics: Sequence and Genome Analysis. – NY: Cold Spring Harbor Laboratory Press, 2004. – 600 p.
12. *Felsenstein J.* Inferring Phylogenies. – Sunderland, MA: Sinauer Associates, 2004. – 580 p.
13. *Zhang, J., Kumar S.* Detection of convergent and parallel evolution at the amino acid sequence level // Mol. Biol. Evol. – 1997. – **14.** – P. 527–536.
14. *Newberg L.A.* Significance of gapped sequence alignments // J. Comput Biol. – 2008. – **15**. – P. 1187–1194.
15. *Dayhoff M.O., Schwartz R., Orcutt B.C.* A model of Evolutionary Change in Proteins // Atlas of protein sequence and structure. – 1978. –**5** (3). – P. 345–358.
16. *Henikoff S., Henikoff J.G.* Amino Acid Substitution Matrices from Protein Blocks // PNAS. – 1992. – **89** (22). – P. 10915–10919.

Отримано 02.04.2010