STUDIA INFORMATICA 2012
Volume 33 Number 2A (105)

Adam KRYGOWSKI, Bożena MAŁYSIAK-MROZEK, Dariusz MROZEK
Politechnika Śląska, Instytut Informatyki

DWUFAZOWY ALGORYTM DOPASOWANIA W POSZUKIWANIU PODOBIEŃSTWA STRUKTUR BIAŁKOWYCH

Streszczenie. Poszukiwanie podobieństwa strukturalnego białek jest jednym z kluczowych, a zarazem trudnych zadań współczesnej bioinformatyki strukturalnej. Bogata przestrzeń poszukiwań oraz różnorodność cech strukturalnych i funkcjonalnych białek sprawiają, że powstaje wiele algorytmów poszukiwania podobieństwa białek, których działanie opiera się na różnych cechach reprezentatywnych. W niniejszym artykule przedstawiono nowy, dwufazowy algorytm dopasowania struktur białkowych, wykorzystywany w poszukiwaniu podobieństwa białek.

Słowa kluczowe: bioinformatyka, dopasowanie, białka, struktura, podobieństwo

TWO-PHASE ALIGNMENT ALGORITHM FOR PROTEIN STRUCTURE SIMILARITY SEARCHING

Summary. Protein structure similarity searching is one of the key, and yet difficult tasks of modern structural bioinformatics. A reach exploration space and a wide variety of structural and functional features of proteins caused the development of many algorithms of protein similarity searching, whose activity is based on various representative characteristics. In this paper, we present a new, two-phase algorithm for matching protein structures used in the protein similarity searching.

Keywords: structural bioinformatics, alignment, protein structure, similarity

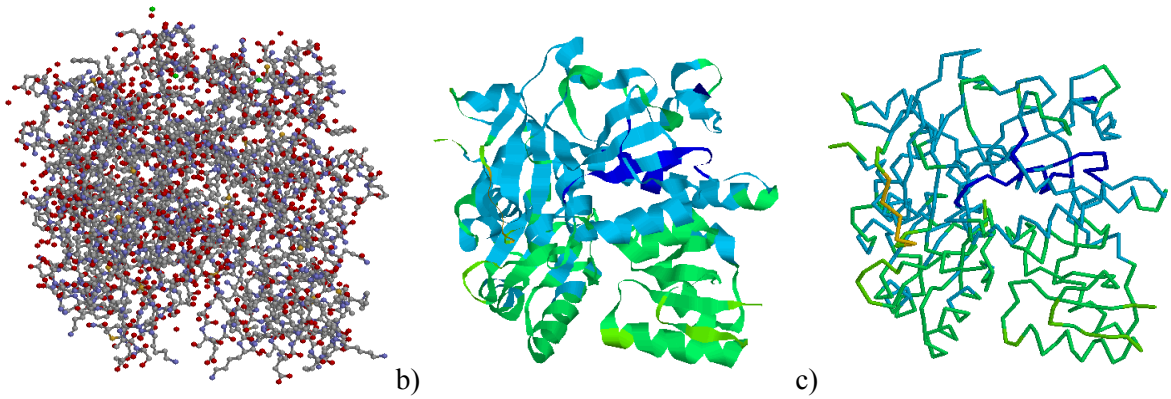
1. Wprowadzenie

Bioinformatyka strukturalna jest gałęzią klasycznej bioinformatyki zajmującą się tworzeniem rozwiązań informatycznych w odniesieniu do przestrzennych struktur molekularnych cząstek biologicznych, takich jak białka, kwasy DNA i RNA. Koncentruje się ona zatem na tworzeniu narzędzi ogólnego przeznaczenia, operujących na bardzo dużym poziomie szcze-

gółowości jeśli chodzi o budowę samych cząstek biologicznych, a mianowicie na poziomie poszczególnych atomów. Przynosi to ogromną nadzieję, że w niedalekiej przyszłości możliwe stanie się bardzo precyzyjne analizowanie skomplikowanych systemów biologicznych oraz diagnozowanie efektów mutacji genetycznych na poziomie molekularnym. Podczas gdy analiza genetyczna pozwala śledzić, jaki jest związek pomiędzy sekwencją genetyczną a wynikającymi z tego skutkami dla funkcjonowania organizmu, bioinformatyka strukturalna umożliwia wgląd w mechanizmy tych konsekwencji biologicznych, a co za tym idzie – w daleko idące zrozumienie, jakie funkcje biologiczne wypływają ze struktury cząsteczki [1].

Poszukiwanie podobieństwa strukturalnego białek jest jednym z trudniejszych zadań współczesnej bioinformatyki strukturalnej. Jednocześnie jest to zadanie niezwykle ważne. O ile poszukiwanie podobieństwa sekwencyjnego białek (na poziomie struktury pierwszorzędowej) sprowadza się zwykle do przeprowadzenia operacji na łańcuchach tekstowych, o tyle problematyczność porównania struktur wynika ze skomplikowanej budowy białek na poziomie molekularnym. Na rysunku 1 zaprezentowano przykładową, skomplikowaną strukturę atomową domen białkowych BARD1 BRCT [2] (PDB ID: 2NTE) z bazy Protein Data Bank [3]. Jeśli przyjmiemy, że średniej wielkości białko jest zbudowane z kilkuset aminokwasów, a każdy z aminokwasów jest zbudowany z kilkunastu atomów, to porównanie wyłącznie pary struktur białkowych staje się nie lada wyzwaniem. Jeśli dodatkowo chcielibyśmy porównać strukturę danego białka z całą bazą danych białek, na przykład w celu porównania zmutowanych struktur, to – biorąc pod uwagę rosnącą liczbę struktur białkowych w bazach danych, takich jak Protein Data Bank (PDB) – zadanie to komplikuje się jeszcze bardziej.

Poszukiwanie podobieństwa strukturalnego białek jest jednak zadaniem bardzo ważnym dla współczesnej biologii porównawczej.



Rys. 1. Różne sposoby reprezentacji struktury przykładowego białka 2NTE z bazy PDB: a) atomy i wiązania, b) struktury drugorzędowe, c) szkielet poprowadzony przez pozycje węgla C_α

Fig. 1. Various representations of sample protein structure 2NTE from PDB: a) atoms and bonds, b) secondary structures, c) backbone through C_α carbon positions

Na podstawie informacji o zbliżonej budowie białek można wnioskować o wspólnym pochodzeniu organizmów i w konsekwencji poznać ewolucję organizmów na przestrzeni milionów lat. Analiza struktur przez ich porównanie umożliwia również poszukiwanie substytutów dla molekuł biologicznych o kluczowym znaczeniu dla wybranych procesów komórkowych, a których niedobór lub nieodpowiednia budowa może powodować dysfunkcje organizmu lub poważne choroby.

W niniejszym artykule przedstawiono nowy algorytm dwufazowego dopasowania pary struktur białkowych, który jest używany w poszukiwaniu podobieństwa strukturalnego białek. Przedstawiony algorytm wciąż pozostaje w fazie testowania, jednakże na chwilę obecną daje bardzo dobre wyniki działania. Autorzy przedstawili przebieg algorytmu oraz przeprowadzone testy nad skutecznością jego działania.

2. Konkurencyjne rozwiązania

W ostatnich dwóch dekadach powstało kilka algorytmów poszukiwania podobieństwa strukturalnego białek, np. VAST [4], DALI [5], LOCK2 [6], FATCAT [7], CTSS [8], CE [9]. Ze względu na złożoność struktur białkowych, zbudowanych z tysięcy atomów, algorytmy te opierają się na różnej reprezentacji struktur przestrzennych w procesie poszukiwania podobieństwa.

Na przykład w algorytmie CTSS uwzględniono lokalne cechy geometryczne oraz wybrane cechy biologiczne. Dla każdego białka konstruowana jest sygnatura kształtu (ang. *shape signature*), na którą składa się krzywa aproksymująca pozycje węgli C_α, natomiast dla każdej reszty aminokwasowej obliczana jest krzywizna (ang. *curvature*) oraz wartości kątów torsyjnych, włączana jest również informacja o typie struktury drugorzędowej.

Algorytm DALI stosuje w procesie porównania tzw. macierze odległości, które buduje dla każdego z pary porównywanych białek. W każdej komórce macierzy przechowywana jest odległość pomiędzy węglami C_α aminokwasów *i* oraz *j* w białku. Macierze odległości są następnie dekomponowane na tzw. wzorce kontaktów (ang. *contact patterns*), będące fragmentami macierzy o wymiarach 6 x 6, i porównywane w celu znalezienia najlepszego dopasowania.

Z kolei w algorytmie VAST w określeniu podobieństwa wykorzystuje się elementy struktury drugorzędowej (ang. *secondary structure elements*), tworzące rdzenie porównywanych białek (α-helisy i β-kartki). Elementy SSE są następnie odwzorowywane na wektory reprezentatywne, dzięki czemu upraszcza się proces analizy. Podczas porównania algorytm podejmuje próbę dopasowania zbioru wektorów dla par struktur białkowych. Podczas takiego dopasowania brany jest pod uwagę: typ struktury reprezentowanej przy pomocy wektora, względna orientacja wektorów oraz uporządkowanie wektorów.

Kompletnie inne podeście zastosowano w rozwijanym przez autorów w ubiegłych latach algorytmie EAST [10] i jego późniejszych modyfikacjach [11, 12, 13]. Algorytm EAST opiera się na reprezentacji struktury białek w postaci tzw. charakterystyk energetycznych, wyznaczanych na podstawie położenia atomów w strukturze. Następnie dopasowanie wybranych charakterystyk energetycznych odzwierciedla dopasowanie struktur białkowych.

3. Opis algorytmu dopasowania

Poszukiwanie podobieństwa białek jest zwykle realizowane przez porównanie zadanego przez użytkownika białka kwerendowego z kolejnymi białkami z bazy danych. W niniejszym rozdziale zostanie przedstawione działanie algorytmu porównania dwóch struktur białkowych. Porównanie to zostanie zrealizowane przy wzięciu pod uwagę różnych poziomów opisu struktury białka. W trakcie porównania wśród cech reprezentatywnych struktur znalazły się cechy trzech poziomów organizacji struktur białkowych – struktury: pierwszorzędowej, drugorzędowej i trzeciorzędowej.

3.1. Reprezentacja białek w procesie porównania

Przyjmijmy następujące oznaczenia:

Q – struktura białka kwerendowego o długości *q* reszt (aminokwasów);

D – struktura białka z bazy danych o długości *d* reszt (aminokwasów).

Struktury w pierwszym etapie działania algorytmu opisywane są za pomocą zredukowanych łańcuchów struktur drugorzędowych, ujawniających rodzaje struktur drugorzędowych występujących w białku:

$$Q = \left(SE_1^Q, SE_2^Q, ..., SE_n^Q\right), \tag{1}$$

gdzie: SE_i^Q to *i*-ty element struktury drugorzędowej białka Q, a *n* to liczba struktur drugorzędowych w łańcuchu kwerendowym, $n \le q$,

$$D = \left(SE_1^D, SE_2^D, ..., SE_m^D\right), \tag{2}$$

gdzie: SE_j^D to *j*-ty element struktury drugorzędowej białka D, a *m* to liczba struktur drugorzędowych w łańcuchu z bazy danych, $m \le d$.

Elementy SE_i^Q i SE_j^D, zwane dalej także regionami lub fragmentami *SE*, zbudowane są z grup sąsiadujących ze sobą aminokwasów o tej samej strukturze drugorzędowej, np. 6 aminokwasów o strukturze helisy alfa i występujących obok siebie tworzy jedną strukturę SE

typu helisa alfa. Stąd też na tym etapie całkowite struktury białek są określone przez *zredukowane łańcuchy struktur drugorzędowych*.

W drugim etapie dopasowywania białka opisywane są w bardziej szczegółowy sposób i na wyższym poziomie rozdzielczości – kolejne aminokwasy białek opisywane są przy pomocy wybranych, charakterystycznych cech strukturalnych. Białka reprezentowane są w postaci łańcuchów:

$$Q = \left(s_1^Q, s_2^Q, ..., s_q^Q\right), \tag{3}$$

gdzie: s_i^Q to *i*-ta sygnatura strukturalna w strukturze białka kwerendowego Q, odpowiadająca *i*-temu aminokwasowi w łańcuchu tego białka, a *q* to długość łańcucha kwerendowego w resztach (aminokwasach);

$$D = \left(s_1^D, s_2^D, ..., s_d^D\right), \tag{4}$$

gdzie: s_j^D to *j*-ta sygnatura strukturalna w strukturze białka z bazy danych D, odpowiadająca *j*-temu aminokwasowi w łańcuchu tego białka, a *d* to długość łańcucha z bazy danych w resztach (aminokwasach).

Każda dowolna sygnatura strukturalna s_i określona jest przez następujące parametry:

$$s_i = \left\langle \vec{C}_i^r, SSE_i, r_i \right\rangle, \tag{5}$$

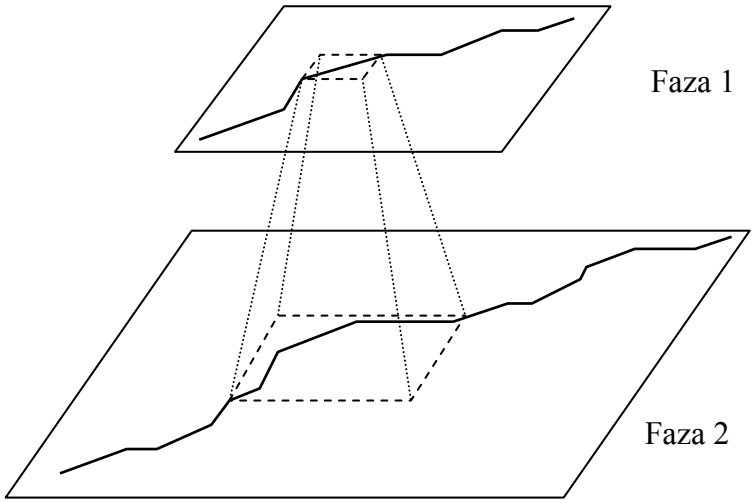
gdzie: $\vec{C}_i^r$ to wektor pomiędzy atomami węgla C_α *i*-1 oraz *i*-tego aminokwasu w łańcuchu białka, SSE_i to rodzaj struktury drugorzędowej, jaką współtworzy dana reszta aminokwasowa, a r_i to typ aminokwasu.

3.2. Ogólny przebieg działania algorytmu

Algorytm dopasowania został podzielony na dwie fazy (rys. 2):

1. W pierwszej fazie prowadzone jest zgrubne dopasowanie struktur przestrzennych, reprezentowanych przez elementy struktury drugorzędowej. Jest to faza dopasowania niskiej rozdzielczości (ang. *low resolution alignment*), ponieważ grupy aminokwasów występujące w każdej strukturze są w niej zgrupowane do jednego elementu reprezentatywnego. Faza ta pozwala na prowadzenie szybkich dopasowań, w których budowane są macierze podobieństwa o niewielkich rozmiarach. Eliminuje to konieczność prowadzenia kosztownych dopasowań dla białek zupełnie do siebie niepodobnych. Białka, które wykazują podobieństwo struktur drugorzędowych, zostają poddane bardziej wnikliwej analizie w fazie 2.
2. W drugiej fazie prowadzone jest szczegółowe dopasowanie struktur przestrzennych, reprezentowanych przez tzw. sygnatury strukturalne. Dopasowanie to opiera się na wyni-

kach dopasowania zgrubnego z fazy 1, jest to natomiast faza dopasowania wysokiej rozdzielczości (ang. *high resolution alignment*), ponieważ aminokwasy nie są w niej grupowane. Zamiast tego każdy aminokwas występujący w strukturze jest reprezentowany przez odpowiednią sygnaturę strukturalną. Następuje zatem dopasowanie łańcuchów sygnatur z wykorzystaniem znacznie większych macierzy podobieństwa, niż to miało miejsce w przypadku fazy 1. Ponadto w tej fazie jest analizowanych więcej cech opisujących struktury białka, a samo białko jest reprezentowane bardziej szczegółowo.



Rys. 2. Ogólny przebieg dwufazowego algorytmu dopasowania
Fig. 2. Overview of two-phase alignment algorithm

W obu fazach dopasowanie prowadzone jest z użyciem, specjalnie zaadaptowanej do tego celu, metody Smitha-Watermana [14]. Przedstawiony sposób dopasowania stanowi podstawę działania algorytmu poszukiwania podobieństwa strukturalnego CASSERT. Nazwa algorytmu jest akronimem od słów określających cechy reprezentatywne struktur białkowych, wziętych pod uwagę w procesie porównania – C_α Atom, Secondary Structure Element, Residue Type (CASSERT). Dokładny sposób reprezentacji struktur oraz przebieg obu faz dopasowania przedstawiono w kolejnych podrozdziałach.

3.3. Pierwszy etap – porównanie struktur drugorzędowych

W pierwszym etapie działania algorytmu struktury białek Q i D są porównywane przez dopasowanie na podstawie ich zredukowanych łańcuchów struktur drugorzędowych, zbudowanych z elementów struktur drugorzędowych SEi. Tę fazę algorytmu nazwano także dopasowaniem niskiej rozdzielczości (ang. *low resolution alignment*). Każdy element SEi, czyli fragment łańcucha wyodrębniony na podstawie jego struktury drugorzędowej, charakteryzują dwie wartości:

$$SE_i = [SSE_i, L_i], \tag{6}$$

gdzie: SSE_i określa rodzaj struktury drugorzędowej, a L_i to długość i-tego elementu SE_i mierzona w resztach (aminokwasach).

W prezentowanej metodzie wyróżnia się trzy podstawowe rodzaje struktur drugorzędowych:

- α-helisa,
- β-harmonijka lub β-nić,
- struktura nieokreślona, która reprezentuje również tzw. pętle i zakręty.

W celu dopasowania struktur Q i D zastosowano algorytm optymalnego dopasowania Smitha-Watermana. W przebiegu algorytmu budowana jest macierz podobieństwa SSE o wymiarach $n \times m$, gdzie n i m określają liczbę struktur drugorzędowych w łańcuchach Q i D, czyli liczbę fragmentów łańcuchów Q i D o rozpoznanej strukturze drugorzędowej. Kolejne komórki macierzy SSE są wypełniane zgodnie z następującymi regułami:

dla $0 \le i \le n$ oraz $0 \le j \le m$:

$$SSE_{i,0} = SSE_{0,j} = 0, \tag{7}$$

$$SSE_{i,j}^{(1)} = SSE_{i-1,j-1} + \delta_{ij}, \tag{8}$$

$$SSE_{i,j}^{(2)} = \max_{1 \le k \le n} \{SSE_{i-k,j} - \omega_k\}, \tag{9}$$

$$SSE_{i,j}^{(3)} = \max_{1 \le l \le m} \{SSE_{i,j-l} - \omega_l\}, \tag{10}$$

$$SSE_{i,j}^{(4)} = 0, \tag{11}$$

$$SSE_{i,j} = \max_{v=1..4} \{SSE_{i,j}^{(v)}\}, \tag{12}$$

gdzie: δ_{ij} jest nagrodą za podobieństwo, określającą stopień podobieństwa dwóch elementów SE_i^Q i SE_j^D białek Q i D, ω_k, ω_l to ewentualne (odpowiednio pozioma i pionowa) kary za wprowadzenie przerwy o długości k i l.

Nagroda δ_{ij} przyjmuje wartości z przedziału <0;1>, gdzie 0 oznacza brak podobieństwa, natomiast 1 oznacza najwyższe możliwe podobieństwo. Stopień podobieństwa wyliczany jest na podstawie wzoru:

$$\delta_{ij} = \sigma_{ij} - \left(\sigma_{ij} * \frac{\left| L_j^D - L_i^Q \right|}{\left(L_j^D + L_i^Q \right)} \right), \tag{13}$$

gdzie: L_i^Q, L_j^D to długości porównywanych fragmentów łańcuchów SE_i^Q i SE_j^D, natomiast σ_{ij} określa stopień podobieństwa struktur drugorzędowych, budujących fragmenty i oraz j porównywanych łańcuchów. Parametr ten może przyjmować trzy możliwe wartości zgodnie z poniższymi regułami:

1) $\sigma_{ij} = 1$, gdy oba elementy SE mają tę samą strukturę drugorzędową, czyli oba łańcuchy mają budowę α-helisy lub β-harmonijki;
2) $\sigma_{ij} = 0{,}5$, gdy przynajmniej jeden z łańcuchów ma nieokreśloną strukturę drugorzędową;
3) $\sigma_{ij} = 0$, gdy jeden z łańcuchów ma budowę α-helisy, a drugi β-harmonijki.

Na przykład, jeżeli oba fragmenty łańcuchów mają tę samą długość, np. 8, i tę samą określoną strukturę drugorzędową, to zgodnie z zależnością (13) stopień ich podobieństwa to: δ_{ij} = 1 – 1*(|8-8|/(8+8)) = 1 – 1*(0/8) = 1 – 0 = 1.

Gdyby jednak jeden z łańcuchów miał długość 4, a drugi długość 8, przy założeniu że wciąż mają tę samą strukturę drugorzędową, to ich stopień podobieństwa byłby następujący: δ_{ij} = 1 – 1*(|4-8|/(4+8)) = 1 – 1*(4/12) = 2/3.

Zmodyfikowany algorytm Smitha-Watermana określa – na podstawie wartości innych komórek – czy stopień podobieństwa jest wystarczający, czy jednak bardziej optymalne będzie umieszczenie w tym miejscu przerwy w dopasowaniu.

3.4. Etap drugi – dopasowanie sygnatur strukturalnych

Dopasowanie sygnatur strukturalnych przebiega analogicznie do pierwszego etapu, jednak wykonywane jest na podstawie elementów struktur drugorzędowych SE_i, dopasowanych na etapie pierwszym. Dla każdej dopasowanej pary regionów SE_i^Q i SE_j^D z poprzedniego etapu tworzona jest nowa macierz podobieństwa o rozmiarach L_i^Q i L_j^D (lub zbliżonych) i wykonywane jest dopasowanie sygnatur strukturalnych, odpowiadających regionom SE_i^Q i SE_j^D (dopasowanie wyższej rozdzielczości, ang. *high resolution alignment*).

Samo dopasowanie przebiega w sposób analogiczny do pierwszego etapu, różnica polega jednak na sposobie oceny podobieństwa dwóch porównywanych elementów, czyli w tym wypadku dwóch sygnatur strukturalnych s_i i s_j. Przy ocenie podobieństwa sygnatur uwzględnia się struktury pierwszo-, drugo- i trzeciorzędowe, zgodnie ze wzorem:

$$ss_{ij} = w_C * \sigma_{ij}^C + w_{SSE} * \sigma_{ij}^{SSE} + w_r * \sigma_{ij}^r, \tag{14}$$

gdzie: σ_{ij}^C jest podobieństwem wektorów $\vec{C}_i^Q$ i $\vec{C}_j^D$ opisujących położenie reszt aminokwasowych w łańcuchach Q i D, σ_{ij}^{SSE} to podobieństwo struktur drugorzędowych (wyliczane zgodnie z regułami 1-3, jak w pierwszym etapie), σ_{ij}^r to podobieństwo reszt aminokwasowych, określane za pomocą zunitaryzowanej macierzy substytucji BLOSUM62, $w_C, w_{SSE}, w_r \in [0;\ 1]$ to wagi dla poszczególnych składowych (domyślnie wszystkie przyjmują wartość 1).

Podobieństwo wektorów $\vec{C}_i^Q$ i $\vec{C}_j^D$ określamy zgodnie z zależnością:

$$\sigma_{ij}^C = 1 - \left| d_{QD} \right|, \tag{15}$$

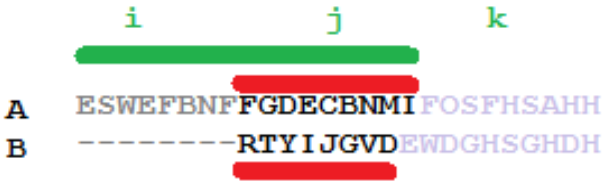
gdzie $\left| d_{QD} \right|$ to moduł z różnicy pomiędzy długościami wektorów $\vec{C}_i^Q$ i $\vec{C}_j^D$:

$$d_{QD} = \left| \vec{C}_i^Q \right| - \left| \vec{C}_i^D \right|. \tag{16}$$

Wartość podobieństwa sygnatur strukturalnych ss_{ij} (zależność (14)) jest uwzględniana jako nagroda za podobieństwo (parametr δ_{ij}, zależność (8)).

Na początku niniejszego rozdziału wspomniano, że macierz podobieństwa ma przybliżone rozmiary L_i^Q i L_j^D. Dokładny rozmiar macierzy podobieństwa sygnatur zależy od dopasowania regionów *SE* sąsiadujących z aktualnie porównywanymi, ponieważ wszystkie ewentualne reszty aminokwasowe, które nie zostały dopasowane w poprzednim kroku (czyli na lewo od porównywanych w bieżącym kroku), zostają dołączone do aktualnych regionów. Jeśli na przykład kilka regionów *SE* poprzedza przerwa w jednym z łańcuchów, to wszystkie reszty wchodzące do przerwy, znajdujące się we fragmencie *SE* drugiego łańcucha, są dołączone do regionu *SE*, sąsiadującego z nimi z prawej strony.

Zasadę tę zilustrowano na rysunku 3. Opisu dokonano na poziomie reszt aminokwasowych ze względu łatwość wizualizacji, w rzeczywistości jednak dopasowaniu ulegają całe sygnatury strukturalne. Na rysunku 3 różnymi odcieniami szarości oznaczono osobne regiony *SE* łańcuchów A i B (regiony *i*, *j*, *k*). Dopasowanie *j* obejmuje fragmenty oznaczone czerwonymi liniami (linie bezpośrednio nad i pod elementami sekwencji). Fragment *j* łańcucha A składa się z 9 reszt, natomiast fragment *j* łańcucha B składa się z 8 reszt. Jednak ponieważ po lewej stronie od *j*-tego dopasowania w łańcuchu B znajduje się przerwa, więc wszystkie aminokwasy z łańcucha A, przyporządkowane do tej przerwy (z części *i*), zostaną włączone do *j*-tego dopasowania.



Rys. 3. Rozszerzanie regionu dopasowania o dodatkowe elementy
Fig. 3. Extending alignment region with additional elements

W konsekwencji takiego scalenia z fragmentem łańcucha B o długości 8 reszt (oznaczonym kolorem czerwonym) porównywany będzie region łańcucha A o długości 17 reszt (oznaczony kolorem zielonym – długa linia tuż pod indeksami regionów *i* oraz *j*).

Załóżmy również, że po wykonaniu *j*-tego dopasowania na poziomie reszt byłyby one ułożone tak jak na rysunku 3. Wówczas we fragmencie łańcucha A po prawej stronie pozosta-

łaby jedna reszta (dokładnie I) niedopasowana do żadnej reszty z łańcucha B. W takiej sytuacji zostałaby ona dołączona do *k*-tego regionu łańcucha A przy wykonywaniu *k*-tego dopasowania.

3.5. Ocena podobieństwa struktur białkowych

Zauważmy także, że zależność (14), opisująca stopień podobieństwa sygnatur strukturalnych, uwzględnia każdy z poziomów organizacyjnych struktury białka – struktury pierwszo-, drugo- i trzeciorzędowe. Ma to duży wpływ na ocenę porównania pary sygnatur z dwóch struktur. Jeżeli dwa aminokwasy opisane przez sygnatury jednocześnie wykazują duże podobieństwo reszty aminokwasowej (np. identyczność), zgodność struktury drugorzędowej oraz zgodność struktury przestrzennej, to istnieje duże prawdopodobieństwo, że oba elementy są strukturalnie podobne, co znajduje odzwierciedlenie w wysokim stopniu podobieństwa sygnatur strukturalnych. Algorytm jednocześnie zakłada, że duży stopień podobieństwa można osiągnąć, jeśli elementy struktury nie wykazują podobieństwa sekwencyjnego (struktury pierwszorzędowej). Co prawda, wartość podobieństwa sygnatur wówczas spada, ale wciąż może pozostać na wysokim poziomie, jeśli pozostałe składowe wciąż wykazują podobieństwo. Oznacza to, że algorytm określa jako podobne również struktury, które nie wykazują podobieństwa sekwencyjnego, wykazują natomiast podobieństwo kształtu. Znajduje to swoje uzasadnienie biologiczne, struktury drugorzędowe są bowiem bardziej konserwatywne od struktur pierwszorzędowych. Jednak, niska zgodność na poziomie struktur trzeciorzędowych może zostać pominięta, jeśli istnieje duże podobieństwo struktur drugorzędowych i pierwszorzędowych, które daje duże szanse na podobieństwo ogólnej struktury. Dużo zależy od podobieństwa nie tylko pojedynczych reszt opisanych przez sygnatury strukturalne, lecz także od tego, czy zostaje zachowana tendencja podobieństwa lub jego braku w otoczeniu kilku kolejnych elementów porównywanych białek. Tendencja ta znajduje odzwierciedlenie w skumulowanych wartościach komórek w macierzy podobieństwa Smitha-Watermana. Wartość miary podobieństwa Score jest otrzymywana dla optymalnej ścieżki dopasowania w macierzy podobieństwa (oznaczonej w tej fazie algorytmu jako S); kumuluje ona wszystkie możliwe nagrody za dopasowania, kary za niedopasowania oraz kary za przerwy w dopasowaniu dla optymalnej ścieżki (zgodnie z zależnościami (7)-(12)) i jest równa najwyższej wartości w macierzy S:

$$Score = \max\{S_{ij}\}, \tag{17}$$

gdzie $i = 1,...,q$, $j = 1,...,d$, q to długość białka kwerendowego, a d to długość białka z bazy danych.

Udział każdej składowej w procesie poszukiwania podobieństwa można regulować za pomocą wag udziału ustawianych przez użytkownika. Na przykład naukowcy poszukujący wyłącznie zaskakujących podobieństw strukturalnych wśród białek niewykazujących podobieństwa sekwencji mogą wyłączyć składową struktury pierwszorzędowej przez określenie wartości 0 dla tej właśnie składowej.

4. Porównanie wyników działania algorytmu CASSERT z rozwiązaniami konkurencyjnymi

Skuteczność działania algorytmu CASSERT poddano różnym testom. W pierwszej kolejności porównano ten algorytm z popularnie wykorzystywanymi algorytmami poszukiwania podobieństwa strukturalnego białek DALI oraz VAST. W drugiej części testów porównano dopasowania generowane przez algorytm CASSERT i algorytm DALI. Trzecia część testów poświęcona była badaniu szybkości działania algorytmu CASSERT, jednakże ze względu na jej objętość nie będzie ona prezentowana w niniejszym artykule.

4.1. Porównanie zbiorów wynikowych

W pierwszym teście porównane zostały zbiory najbardziej podobnych łańcuchów wskazanych przez algorytmy CASSERT, DALI i VAST. Utworzono 6 zapytań dla sześciu różnych białek kwerendowych. Każde zapytanie wykonano za pomocą każdego z trzech algorytmów. Białka kwerendowe zadane w zapytaniu różniły się wielkością (długością) – od białek o krótkich łańcuchach (do 100 aminokwasów), poprzez średniej wielkości (do 500 aminokwasów), aż po długie łańcuchy (powyżej 500 aminokwasów). W ten sposób otrzymywano trzy zbiory wynikowe, z których każdy zawierał 100 struktur białkowych (lub ich łańcuchów) wskazanych przez każdy z algorytmów jako najbardziej podobne. Otrzymane zbiory można podzielić na trzy grupy w zależności od rozmiarów białka kwerendowego. W tabelach 1-3 przedstawiono zestawienie opisujące, ile procent tych samych białek (łańcuchów) znaleziono w zbiorach wynikowych testowanych algorytmów. Analizując wskazane zestawienia, można zauważyć duże podobieństwo pomiędzy wynikami generowanymi przez algorytm DALI i autorski algorytm CASSERT. Można także stwierdzić, że występują dość duże rozbieżności pomiędzy wynikami tych dwóch algorytmów a wynikami generowanymi przez algorytm VAST. Jednak Powodów należy szukać nie w zasadach działania algorytmów, ale w wykorzystywanych bazach danych. Testowanie algorytmów DALI i VAST odbyło się przez dedykowane serwisy internetowe, które wykorzystują swoje własne bazy danych struktur. Baza danych wykorzystywana przez algorytm CASSERT jest uproszczoną wersją bazy danych używanej przez al-

gorytm DALI – baza ta zawiera mniejszą liczbę rekordów oraz mniej dokładnie opisuje struktury przestrzenne białek.

Algorytm VAST wykorzystuje do obliczeń bazę danych MMDB [15], dlatego wśród wyników pojawiają się takie białka, które nie mają swych odpowiedników w bazie danych stosowanej podczas testowania algorytmu CASSERT. Niestety autorzy nie mieli zbyt dużego wpływu na bazy danych używane przez konkurencyjne algorytmy właśnie ze względu na fakt, iż ich testowanie odbywa się za pośrednictwem dedykowanych serwisów internetowych, które uniemożliwiają wybranie bazy danych. Niemniej jednak pojawianie się wspólnych białek w każdym zbiorze należy uznać za dobrą prognozę przed kolejnymi testami.

Tabela 1

Procent tych samych łańcuchów w wynikach działania testowanych algorytmów dla krótkich łańcuchów kwerendowych

%	VAST	DALI	CASSERT
VAST	100	32.5	29
DALI	32,5	100	86
CASSERT	29	86	100

Tabela 2

Procent tych samych łańcuchów w wynikach działania testowanych algorytmów dla średnich łańcuchów kwerendowych

%	VAST	DALI	CASSERT
VAST	100	35	25
DALI	35	100	77
CASSERT	25	77	100

Tabela 3

Procent tych samych łańcuchów w wynikach działania testowanych algorytmów dla długich łańcuchów kwerendowych

%	VAST	DALI	CASSERT
VAST	100	15	17
DALI	15	100	54
CASSERT	17	54	100

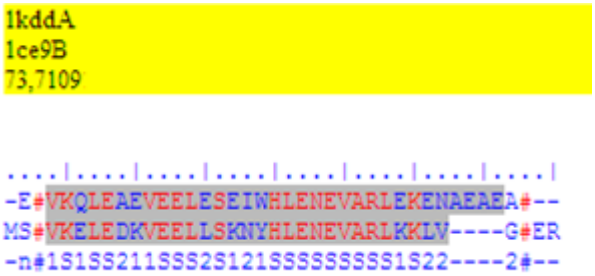
Inną zauważalną prawidłowością jest spadek podobieństwa zbiorów wyników dla rosnących rozmiarów białka określonego w zapytaniu. Wynika to z faktu, iż im łańcuchy białkowe zadane w zapytaniu są dłuższe, tym więcej możliwości dopasowania, dlatego różnice w dopasowaniach poszczególnych algorytmów są większe.

4.2. Porównianie dopasowań

Poszukiwanie podobieństwa strukturalnego polega nie tylko na wskazywaniu podobnych łańcuchów białkowych, ale także na określeniu ich dopasowania, tzn. na wskazaniu wystę-

powania wzajemnej relacji (podobieństwa strukturalnego) pomiędzy poszczególnymi aminokwasami w strukturze porównywanych białek. Przeprowadzone testy, podobnie jak w poprzednim przypadku, pokazały dużą zbieżność wyników generowanych przez algorytmy CASSERT oraz DALI, natomiast mniejszą zbieżność wyników w porównaniu z algorytmem VAST.

Wyniki te mogą się również nieco różnić. Przyjrzyjmy się przykładom dopasowań strukturalnych przeprowadzonych za pomocą algorytmu CASSERT (rysunek 4) oraz algorytmu DALI (rys. 5) dla pary przykładowych struktur 1KDD-łańcuch A [16] oraz 1CE9-łańcuch B [17] z bazy DALI.



Rys. 4. Dopasowanie strukturalne wygenerowane przez algorytm CASSERT
Fig. 4. Structural alignment generated by CASSERT algorithm

```
Query=1kddA Sbjct=1ce9B Z-score=4.1
DSSP  lhHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHL-
Query evKQLEAEVEELESEIWHLENEVARLEKENAECEA-    35
ident          |
Sbjct --MSVKELEDKVEELLSKNYHLENEVARLKKLVGEr    34
DSSP  --LLHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLLl
```

Rys. 5. Dopasowanie strukturalne wygenerowane przez algorytm DALI
Fig. 5. Structural alignment generated by DALI algorithm

W strukturze drugorzędowej obu białek występują α-helisy. Dopasowania wygenerowane przez oba algorytmy są nieco inne – w dopasowaniu strukturalnym algorytmem DALI widać zbieżność tylko jednej reszty aminokwasowej (oznaczona pionową linią), natomiast w dopasowaniu strukturalnym algorytmem CASSERT zbieżność strukturalna reszt aminokwasowych jest znacznie większa (pozycje oznaczone kolorem czerwonym lub literą S w linii podobieństwa). Ponieważ jednocześnie ze zbieżnością strukturalną, oznaczoną przez algorytm CASSERT kolorem czerwonym (lub literą S), obserwujemy również zbieżność rodzajową aminokwasów, pozwala to sądzić, że na tych pozycjach dopasowania strukturalne są poprawne. Podobieństwo lub wręcz identyczność aminokwasów w porównywanych łańcuchach, szczególnie tych obserwowanych na wielu następujących po sobie elementach, bardzo często pociąga za sobą podobieństwo ukształtowania przestrzennego cząsteczek białkowych. Wynika z tego, że uwzględnienie w dopasowaniu strukturalnym również podobieństwa sekwencji korzystnie wpływa na końcowy wynik dopasowania.

Porównując oba rezultaty, można również zauważyć, że dopasowania wygenerowane przez oba algorytmy są przesunięte o trzy reszty aminokwasowe, czyli mniej więcej o tyle, ile wynosi jeden obrót w α-helisie (na jeden obrót przypada dokładnie 3,6 reszty aminokwasowej). Algorytmy, które nie uwzględniają w swoim działaniu podobieństwa sekwencji aminokwasów, są narażone na tego typu przesunięcia, ponieważ spiralna α-helisa zawiera powtarzające się kształty, które można dopasować na różne sposoby.

5. Podsumowanie

Algorytm CASSERT z zaimplementowaną metodą dwufazowego dopasowania struktur białkowych daje bardzo obiecujące wyniki. Jak pokazały przeprowadzone testy, zbiór rezultatów zwracanych przez algorytm CASSERT jest bardzo zbliżony do tego, który jest zwracany przez algorytm DALI. Trudno porównać wyniki algorytmu CASSERT z wynikami algorytmu VAST ze względu na różne zbiory białek, które są przeszukiwane w trakcie prowadzonych testów. Dzieje się tak dlatego, że przeszukiwanie algorytmem VAST było prowadzone przez witrynę internetową, gdzie użytkownik nie ma wpływu na bazę danych, na której prowadzi się poszukiwania. Otrzymane wyniki porównania CASSERT-VAST nie są zatem najbardziej wiarygodne, co w tym wypadku autorzy traktują mimo wszystko jako pozytywną przesłankę, tzn. pokładają głęboką nadzieję, że liczba zgodnych struktur byłaby procentowo wyższa w porównaniu wyników obu algorytmów, gdyby badania zostały przeprowadzone na tej samej bazie danych struktur.

Z tego też powodu w dalszej części badań postanowiono prowadzić porównania algorytmu CASSERT tylko z algorytmem DALI. Dopasowania generowane przez oba algorytmy są podobne, przy czym w trakcie prowadzonych badań autorzy znaleźli przypadki, w których algorytm CASSERT daje lepsze dopasowania niż algorytm DALI.

Wśród potencjalnych zastosowań algorytmu CASSERT można wskazać przede wszystkim identyfikację białek na podstawie ich struktury przestrzennej, a w konsekwencji także przewidywanie pełnionych przez białko funkcji komórkowych. Innym obszarem zastosowań może być porównywanie struktur białkowych, otrzymanych w procesie predykcji, do znanych struktur białek z bazy danych i oszacowanie jakości predykcji.

Dalsze prace autorów będą zmierzały w kierunku poprawy wydajności działania algorytmu CASSERT przez efektywne składowanie cech reprezentatywnych struktur białkowych w bazie danych oraz rozpraszanie obliczeń na wielu komputerach lub procesorach.

Praca naukowa finansowana ze środków na naukę w latach 2008-2011 jako projekt badawczy N N516 265835: Poszukiwanie podobieństwa strukturalnego białek w rozproszonym środowisku wieloagentowym.

BIBLIOGRAFIA

1. Gu J., Bourne P. E.: Structural Bioinformatics (Methods of Biochemical Analysis), 2 edition. Wiley-Blackwell, 2009.
2. Birrane G., Varma A. K., Soni A., Ladias J. A.: Crystal structure of the BARD1 BRCT domains. Biochemistry, Vol. 46(26), 2007, s. 7706÷7712.
3. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H. et al.: The Protein Data Bank. Nucleic Acids Res., No. 28, 2000, s. 235÷242.
4. Gibrat J. F., Madej T., Bryant S. H.: Surprising similarities in structure comparison. Curr. Opin. Struct. Biol., Vol. 6(3), 1996, s. 377÷385.
5. Holm L, Sander C.: Protein structure comparison by alignment of distance matrices. J. Mol. Biol., Vol. 233(1), 1993, s. 123÷138.
6. Shapiro J., Brutlag D.: FoldMiner and LOCK2: protein structure comparison and motif discovery on the web. Nucleic Acids Res., Vol. 32, 2004, s. 536÷541.
7. Ye Y., Godzik A.: Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics, Vol. 19(2), 2003, s. 246÷255.
8. Can T., Wang Y. F.: CTSS: A Robust and Efficient Method for Protein Structure Alignment Based on Local Geometrical and Biological Features. Proceedings of the 2003 IEEE Bioinformatics Conference, 2003, s. 169÷179.
9. Shindyalov I. N., Bourne P. E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering, Vol. 11(9), 1998, s. 739÷747.
10. Mrozek D., Małysiak B., Kozielski S.: EAST: Energy Alignment Search Tool. Springer-Verlag GmbH, LNAI, Vol. 4223, 2006, s. 696÷705.
11. Mrozek D., Małysiak B.: Searching for Strong Structural Protein Similarities with EAST. Journal of Computer Assisted Mechanics and Engineering Sciences, No. 14, 2007, s. 681÷693.
12. Mrozek D., Małysiak-Mrozek B.; Kozielski S.: Alignment of protein structure energy patterns represented as sequences of Fuzzy Numbers. 32th Annual Meeting of the North American Fuzzy Information Processing Society, 2009. IEEE, Cincinnati, USA 2009, s. 1÷6.

13. Mrozek D., Małysiak-Mrozek B.: An Improved Method for Protein Similarity Searching by Alignment of Fuzzy Energy Signatures. International Journal of Computational Intelligence Systems, Vol. 4, No. 1, Atlantis Press 2011, s. 75÷88.
14. Smith T. F., Waterman M. S.: Identification of common molecular subsequences. J. Mol. Biol., Vol. 147, 1981, s. 195÷197.
15. Hogue C., Ohkawa H., Bryant S.: A dynamic look at structures: WWW-Entrez and the Molecular Modelling Database. Trends Biochem. Sci., Vol. 21, 1996, 226÷229.
16. Keating A. E., Malashkevich V. N., Tidor B., Kim P. S.: Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. Proc. Natl. Acad. Sci., Vol. 98(26), USA 2001, s.14825÷14830.
17. Lu M., Shu W., Ji H., Spek E., Wang L., Kallenbach N.R.: Helix capping in the GCN4 leucine zipper. J. Mol. Biol., Vol. 288(4), 1999, s. 743÷752.

Wpłynęło do Redakcji 16 stycznia 2012 r.

Abstract

Protein structure similarity searching is one of the key, but most difficult tasks of modern structural bioinformatics. While searching for protein sequence similarity (similarity at the level of primary structure) is often carried out by using operations on strings, protein structure comparison is more problematic due to the complicated nature of proteins on a molecular level. If we assume that an average size protein is made up of several hundred amino acids, and each amino acid is made up of several atoms, then a comparison of only one pair of protein structures is a challenge. If you also want to compare the structure of the protein with the entire database of proteins, for example, to compare mutant structures to each other, then taking into account the increasing number of protein structures in databases, such as the Protein Data Bank (PDB), this task becomes even more complicated.

However, protein structure similarity searching is a very important task for the modern structural bioinformatics. Based on the information about similar protein structures we can conclude about common descent of organisms and thus, we can study the evolution of organisms over millions of years. The analysis of protein structures by their comparison allows us to search for substitutes for biological molecules critical for certain cellular processes, whose lack or inadequate design can cause dysfunction of the body or serious diseases.

In this paper we present a new, two-phase algorithm for matching protein structures used in the protein similarity searching. Presented algorithm is still in the testing phase, but it already gives promising results. In the paper, we also present tests that we have performed in order to examine the effectiveness of the algorithm.

Adresy

Adam KRYGOWSKI: absolwent Instytutu Informatyki, Politechniki Śląskiej, ul. Akademicka 16, 44-100 Gliwice, Polska, adam.krygowski@gmail.com.
Bożena MAŁYSIAK-MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, bozena.malysiak@polsl.pl.
Dariusz MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, dariusz.mrozek@polsl.pl.