STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Adam SKOWRON, Dariusz MROZEK
Politechnika Śląska, Instytut Informatyki

WIZUALIZACJA ALGORYTMÓW OPTYMALNEGO DOPASOWANIA SEKWENCJI NUKLEOTYDÓW I AMINOKWASÓW

Streszczenie. Celem prac przedstawionych w niniejszym artykule była konstrukcja narzędzia do wizualizacji wybranych algorytmów optymalnego dopasowania sekwencji nukleotydów i aminokwasów. Zasadę działania zbudowanego narzędzia można sprowadzić do trzech kroków. W kroku pierwszym określane są parametry wejściowe. W kroku drugim następuje wizualizacja dopasowania sekwencji biopolimerowych. Na koniec wyznaczane jest optymalne dopasowanie, zobrazowane ścieżką przejścia oraz wartością liczbową.

Słowa kluczowe: algorytmy optymalnego dopasowania sekwencji, algorytm Needlemana-Wunscha, algorytm Smitha-Watermana, algorytm Needlemana-Wunscha-Sellersa

VISUALIZATION OF OPTIMAL ALIGNMENT ALGORITHMS OF NUCLEOTIDE AND PROTEIN SEQUENCES

Summary. The aim of the work reported in this paper was to develop a tool for visualization of optimal alignment algorithms of nucleotide and protein sequences. Functioning of the developed tool is based on three steps. In the first step, input parameters are determined. In the second step, the tool visualizes alignment of biopolymers according to the chosen algorithm. At the end, an optimal alignment is illustrated as an alignment path with appropriate similarity measure.

Keywords: optimal alignment algorithms on two sequences, Needleman-Wunsch algorithm, Smith-Waterman algorithm, Needleman-Wunsch-Sellers algorithm

1. Wprowadzenie

Dopasowanie sekwencji nukleotydów lub aminokwasów to próba ułożenia dwóch sekwencji biopolimerów (DNA, RNA lub białek) w celu zidentyfikowania regionów wykazujących podobieństwo. Na podstawie oceny podobieństwa staramy się wnioskować o występowaniu strukturalnej, funkcjonalnej lub ewolucyjnej relacji pomiędzy tymi sekwencjami. Poznanie funkcji badanych regionów pozwala na ich wykorzystanie do celów praktycznych.

Przy założeniu wspólnego pochodzenia sekwencji znajomość dopasowania na poziomie molekularnym to nie tylko zakładanie podobieństwa, ale także próba wnioskowania o różnicach, które w takim przypadku można nazwać mutacjami. Rozwój w dziedzinie biologii molekularnej zaowocował odkryciem, iż występujące w sekwencjach mutacje mogą: nie mieć konsekwencji genetycznych, są niepożądane (choroby o podłożu genetycznym, np. choroby nowotworowe, choroby układu odpornościowego), ale także pożyteczne (nabycie odporności) [1, 2].

Informacje o dopasowaniu sekwencji biopolimerów mają wiele praktycznych zastosowań między innymi w przemyśle spożywczym, branży medycznej i farmaceutycznej, sądownictwie, archeologii. Badania nad poszukiwaniem podobieństwa poszczególnych genów, całych genotypów lub fenotypów często występują także jako tematy prac naukowych. Z kolei algorytmy dopasowania sekwencji wykorzystywane są również do celów niezwiązanych z biologią, na przykład w kryptografii czy branży muzycznej (wyłapywanie plagiatów na podstawie linii melodycznej) [3].

Narzędzie zaprezentowane w niniejszym artykule zostało zaprojektowane w celu wizualizacji wybranych algorytmów optymalnego dopasowania. W aplikacji o nazwie OPAL zaproponowane zostały trzy tryby działania. Tryb pierwszy – *Pokazowo-szkoleniowy* – służy przedstawieniu kolejnych kroków algorytmów, w celu zrozumienia zasady działania. Tryb drugi – *Szybki wynik* – polega na natychmiastowym wyświetleniu optymalnego dopasowania i może być wykorzystywany do szybkiego porównywania dwóch sekwencji. Ostatni tryb – *Interaktywny* – został zaproponowany w celu zrozumienia zasad działania algorytmów.

W sekcji 2 przedstawiono rozwiązania konkurencyjne dla serwisu OPAL. Elementy wpływające na wynik dopasowania sekwencji oraz ich interpretacje biologiczne są tematem sekcji 3. W sekcji 4 ogólnie opisano wybrane algorytmy optymalnego dopasowania, a następnie przedstawiono uzyskane wyniki prac (sekcja 5).

2. Prace pokrewne

Istniejące aplikacje komercyjne, takie jak CLC Genomics Workbench [4] oraz Geneious [5] pozwalają na prezentacje dopasowania sekwencji z wykorzystaniem wielu dodatkowych opcji i wykresów, takich jak: zmiany kolorów, grafy miejsc konserwatywnych czy wykresy jakościowe dopasowania. Dodatkowo, wyświetlają wiele pomocnych statystyk, pozwalają na zapamiętywanie historii operacji oraz bezpośrednie zaznaczanie interesujących regionów i ich anotacje. Oba programy nie pokazują jednak zasad działania wykorzystywanych algorytmów, a jedynie prezentują ostateczny wynik dopasowania. W obu przypadkach rola użytkownika sprowadzona została do zdefiniowania sekwencji, wybrania jednej z dostępnych macierzy podobieństwa oraz określenia wartości kar za wprowadzenie bądź wydłużenie przerwy.

Istniejące darmowe rozwiązania NW-align [6] oraz B.A.B.A [7] są znacznie uboższe – pod względem prezentacji wyników – od zaprezentowanych aplikacji komercyjnych. W przypadku NW-align użytkownik może jedynie wprowadzić sekwencje, reszta parametrów jest z góry określona przez autorów.

Interesującym rozwiązaniem jest applet Java o charakterze edukacyjnym o nazwie B.A.B.A. Aplikacja pozwala na wprowadzanie sekwencji, macierzy podobieństwa i określenia liniowych wartości kar. W trakcie działania aplikacji użytkownikowi prezentowane są uproszczone wzory, a kolejne komórki oznaczane są w sposób pozwalający na identyfikację etapów. Wynik dopasowania nie zawsze jednak pokrywa się z wynikami uzyskanymi za pomocą aplikacji komercyjnych – przy tym samym algorytmie i wartościach parametrów. Można zatem wnioskować o niedokładnej implementacji prezentowanego algorytmu. Dodatkowo brakuje numerycznej wartości wyniku dopasowania, nie można zdefiniować kary za wprowadzenie przerwy w postaci afinicznej, a wyświetlane obliczenia matematyczne zostały uproszczone do jednej operacji w danym kierunku. W aplikacji edukacyjnej B.A.B.A brakuje również trybu interaktywnego, który pozwalałby użytkownikowi na sprawdzenie stopnia zrozumienia algorytmów.

3. Elementy wpływające na wynik dopasowania sekwencji

Elementami najmocniej wpływającymi na wynik dopasowania są funkcja kosztów oraz macierze substytucji. W związku z tym, uzyskany wynik dopasowania sekwencji biopolimerów może być różny.

3.1. Funkcja kosztów

Jedną z podstawowych miar pozwalających mówić o podobieństwie dwóch sekwencji względem siebie jest miara odmienności napisów, zwana odległością Levenshteina. Taka interpretacja pozwala na określenie wszystkich typów operacji, jakie występują przy porównywaniu kolejnych znaków sekwencji i zamianie jednej z nich w drugą. Są to:

- pozostawienie pary symboli bez zmian (występuje zgodność pomiędzy znakami),
- zamiana litery na inną,
- insercja symbolu w sekwencji 1,
- delecja symbolu w sekwencji 1.

Do każdej z operacji przypisać można jej koszt wykonania, zwany także wagą. Dzięki temu możliwe jest liczbowe określenie miary podobieństwa lub dystansu. Funkcja, która na podstawie porównywanych par zwraca pewną wartość liczbową, nazywana jest funkcją kosztów. Dobierając wartości poszczególnych wag można wymusić zachowanie algorytmu do generowania dopasowań z częściej (lub rzadziej) pojawiającymi się przerwami lub niedopasowaniami. Co więcej zmiana funkcji kosztów może sprawić, że pojawi się nowe optymalne dopasowanie.

Należy brać pod uwagę, iż odpowiednia definicja funkcji kosztów powinna mieć swoje uzasadnienie biologiczne, na przykład, że indele (przerwy spowodowane insercją lub delecją) pojawiają się rzadziej w kodzie genetycznym niż mutacje punktowe zmieniające dany nukleotyd w inny [8-9].

3.2. Macierze substytucji

W przypadku dopasowywania nukleotydów wartości wag zazwyczaj określane są jednakowo dla dopasowania, niedopasowania oraz wprowadzonej przerwy. Należy jednak być świadomym, iż każda zamiana nukleotydu w inny jest z góry traktowana jako niedopasowanie, natomiast wszystkie dopasowania są jednakowo ważne, co niekoniecznie może być prawdą w danej sekwencji. W przypadku gdy porównanie dotyczy sekwencji aminokwasów, należy uwzględnić występujące między nimi biofizyczne i biochemiczne zależności.

W rzeczywistości każdy aminokwas ma mniejsze lub większe prawdopodobieństwo mutacji w inny konkretny aminokwas i tak na przykład, hydrofobowy aminokwas walina podczas mutacji częściej pozostaje hydrofobowy, niż zmienia swoją właściwość. Zatem zmiana aminokwasu niekoniecznie musi wpływać na zmianę funkcji białka, a tego typu podstawienia powinny być lepiej punktowane niż te, które zmieniają funkcję białka [10].

Jako, że współczynniki wpływające na prawdopodobieństwo są liczne i zróżnicowane, bezpośrednia obserwacja rzeczywistych stosunków wszystkich podstawień jest często najlepszym sposobem na uzyskanie wartości podobieństwa. Na podstawie obserwacji powstało

wiele różnych macierzy substytucji dla aminokwasów PAM, BLOSUM, GONNET, JTT, DCMUT, MTREV i inne, jednak do najpopularniejszych należą pierwsze dwie [11, 12].

4. Algorytmy optymalnego dopasowania sekwencji

Pierwszy algorytm do porównania sekwencji aminokwasów zaproponowany został przez Saula B. Needlemana i Christiana D. Wunscha w 1970 r. Algorytm ten, nazwany od nazwisk naukowców algorytmem Needlemana-Wunscha, pozwala odnaleźć optymalne dopasowanie sekwencji w ujęciu globalnym. Działania algorytmu opierają się na wyznaczeniu macierzy podobieństwa M dwóch sekwencji $A=a_1a_2...a_m$ i $B=b_1b_2...b_p$ oraz znalezieniu w niej ścieżki dopasowania. Wartości komórek w macierzy M wyznacza się zgodnie ze wzorem:

$$M_{ij} = \max\begin{cases} M_{i-1,j-1} + s(a_i, b_j) \\ \max\{M_{i-k,j-1} + s(a_i, b_j) - d\}, \\ \max\{M_{i-1,j-l} + s(a_i, b_j) - d\} \end{cases} \tag{1}$$

gdzie: $s(a_i,b_j)$ to nagroda za dopasowanie lub kara za brak dopasowania (w najprostszej postaci jeśli elementy sekwencji są sobie równe, tj. $a_i = b_j$, wartość nagrody równa się 1, w przeciwnym wypadku kara wynosi 0), M_{ij} to komórka w macierzy dopasowania, d kara za wprowadzenie przerwy [9, 13, 14].

Rozszerzeniem algorytmu Needlemana-Wunscha jest algorytm zaproponowany w 1981 r. przez Temple F. Smitha i Michaela S. Watermana. Autorzy zasugerowali modyfikację pozwalającą odnaleźć kilka odcinków, po jednym dla każdej sekwencji, takich, że nie ma innych par odcinków o większym podobieństwie (homologii). A zatem poszukiwanie podobieństwa nie ogranicza się do całych sekwencji, lecz pozwala porównywać odcinki o dowolnej długości wybierając te, które są optymalne ze względu na miarę podobieństwa (dopasowanie lokalne). Działanie algorytmu Smitha-Watermana również oparte jest na skonstruowaniu macierzy podobieństwa, zgodnie ze wzorem:

$$H_{ij} = \max\begin{cases} H_{i-1,j-1} + s(a_i, b_j) \\ \max\{H_{i-k,j} - W_k\} \\ \max\{H_{i,j-l} - W_l\} \\ 0 \end{cases}, \tag{2}$$

gdzie: $s(a_i,b_j)$ to podobieństwo pomiędzy elementami a_i i b_j sekwencji A i B, W_k i W_l to kary za wprowadzenie przerwy o długości k i l [14-16].

Algorytm programowania dynamicznego stosowany w dopasowaniu globalnym sekwencji nazywany jest powszechnie algorytmem Needlemana-Wunscha. W rzeczywistości jego aktualna wersja bliższa jest metody, zaproponowanej przez Petera H. Sellersa w 1974 r. Algorytm

Needlemana-Wunscha-Sellersa to prostsza wersja algorytmu Needlemana-Wunscha, w której zredukowano stopień złożoności obliczeniowej algorytmu z $O(N^3)$ do $O(N^2)$. Macierz podobieństwa w algorytmie Needlemana-Wunscha-Sellersa wyznacza się zgodnie ze wzorem:

$$M_{ij} = \max \begin{cases} M_{i-1,j-1} + s(a_i, b_j) \\ \max\{M_{i-k,j} - G_P\}, \\ \max\{M_{i,j-l} - G_P\} \end{cases} \tag{3}$$

gdzie: $s(a_i,b_j)$ to podobieństwo pomiędzy elementami a_i i b_j sekwencji A i B, G_p to kara za wprowadzenie przerwy. Kara za wprowadzenie przerwy może być stała, liniowa lub może przyjmować postać kary afinicznej $G_p = G_{open} + nG_E$, gdzie: G_E – kara za przedłużenie przerwy, n – długość przerwy, G_{open} – kara za otwarcie przerwy. Ta ostatnia postać kary ma swoje biologiczne uzasadnienie, które wynika z faktu, że niektóre sekwencje są bardziej narażone na dłuższe przerwy, niż na wiele krótkich przerw [17-18].

5. Serwis OPAL

Serwis OPAL (**op**timal alignment **al**gorithms for nucleotide and protein sequences) pozwala na dopasowanie dwóch sekwencji biopolimerowych używając jednego z trzech omówionych algorytmów optymalnego dopasowania. OPAL pozwala także na dokładną wizualizację poszczególnych kroków, prowadzących do dopasowania. Serwis OPAL, w zależności od wybranego trybu działania, może pełnić rolę edukacyjną i pokazową w procesie nauczania popularnych algorytmów, wykorzystywanych w bioinformatyce, może także pełnić rolę czysto użytkową w procesie badania dwóch sekwencji biopolimerowych. Serwis dostępny jest w Internecie pod adresem: *http://www.opal.przyjaznycms.pl*. Główne okno serwisu OPAL, które pełni rolę informacyjną, przedstawiono na rys. 1.

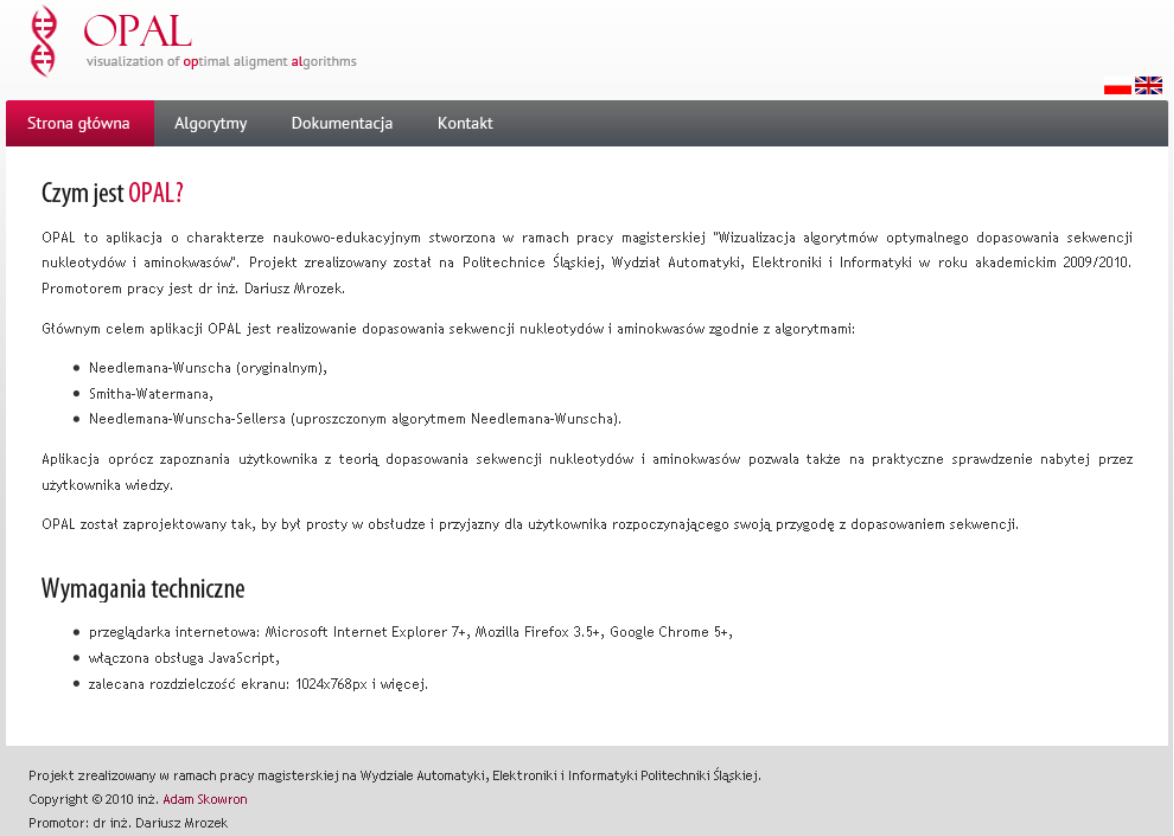
Aby prześledzić działanie algorytmów dopasowania, należy najpierw na stronie *Algorytmy* wybrać algorytm i rodzaj sekwencji, a także określić parametry algorytmu i procesu wizualizacji (rys. 2). Formularz jest podzielony na dwie części: z opcjami podstawowymi, których wypełnienie jest obowiązkowe oraz z opcjami dodatkowymi, które posiadają ustawienia domyślne.

System OPAL pozwala wizualizować następujące algorytmy dopasowania sekwencji:

- Needlemana-Wunscha,
- Needlemana-Wunscha-Sellersa,
- Smitha-Watermana.

Wizualizację można prowadzić dla sekwencji:

- nukleotydów,
- aminokwasów.



Rys. 1. Strona główna aplikacji OPAL
Fig. 1. OPAL homepage

Punktacja w wybranym algorytmie może być:

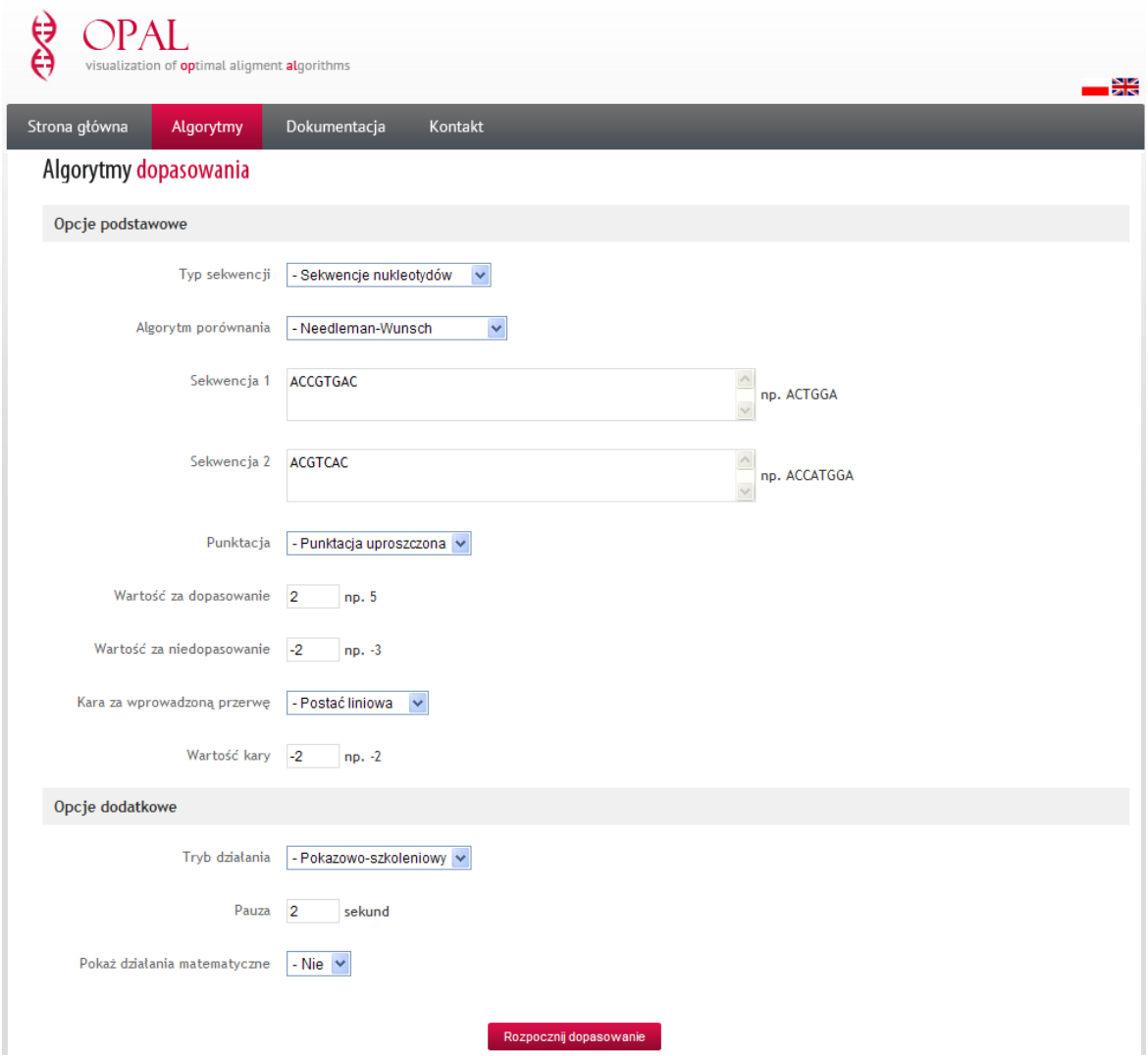
- uproszczona, z podaniem nagrody za dopasowanie i kary za niedopasowanie,
- zgodna z macierzą podobieństwa, zwykle wybraną macierzą substytucji podczas dopasowania sekwencji białkowych.

Kara za wprowadzenie przerwy może mieć postać:

- liniową – stała wartość kary za każdą przerwę,
- afiniczną – kara za otwarcie przerwy oraz kara za rozszerzanie przerwy, proporcjonalna do jej długości.

Serwis OPAL umożliwia trzy tryby wizualizacji dopasowania (sekcja opcje dodatkowe, rys. 2):

- *Pokazowo-szkoleniowy* – wizualizuje kolejne kroki działania algorytmów, w celu pokazania zasady ich działania.
- *Szybki wynik* – natychmiast wyświetla optymalne dopasowanie i może być wykorzystywany do szybkiego porównywania dwóch sekwencji.
- *Interaktywny* – pozwala użytkownikowi samodzielnie wypełnić macierz podobieństwa dla wybranego algorytmu, zgodnie z przyjętym sposobem punktacji.



Rys. 2. Ustawianie parametrów algorytmów dopasowania w serwisie OPAL
Fig. 2. Configuration of parameters of alignment algorithms in OPAL

Dodatkowo, dla trybu *Pokazowo-szkoleniowego* można określić pauzę pomiędzy kolejno wizualizowanymi krokami (np. 2 s), a także czy mają być pokazywane działania matematyczne, prowadzące do określenia wartości danej komórki macierzy podobieństwa.

Biorąc pod uwagę, iż aplikacja daje możliwość dopasowania sekwencji z wykorzystaniem jednego z trzech algorytmów optymalnego dopasowania i dla każdego z nich możliwe są trzy tryby działania, dostępnych jest dziewięć różnych możliwości uruchomienia aplikacji – niezależnie od danych wejściowych, tj. bez względu na to, czy użytkownik przekaże sekwencje nukleotydów czy aminokwasów oraz jaką punktację i rodzaj kary wybierze.

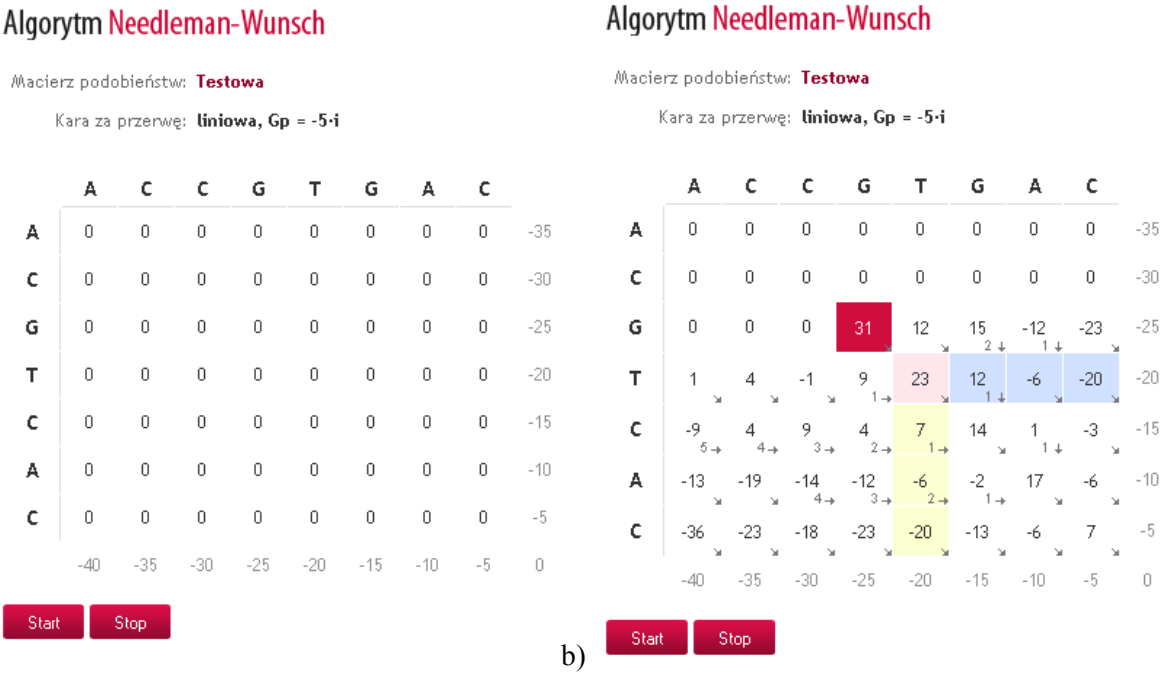
Reguły działania narzędzia różnią się w każdym z 9 wspomnianych przypadków. Najbardziej widoczna dla użytkownika jest zmiana trybu działania algorytmu, jednakże zwracany wynik dopasowania najsilniej zależy od wybranego algorytmu. Mając na uwadze powyższe rozważania w niniejszym artykule zaprezentowano przypadki uruchomienia aplikacji w dwóch trybach działania: *Pokazowo-szkoleniowym* i *Interaktywnym*. Dla trybu *Pokazowo-*

szkoleniowego dopasowaniu poddano sekwencje ACCGTGAC i ACGTCAC, z kolei dla trybu *Interaktywnego* porównywano sekwencje CAGTAAG i GTAAG. W obu przypadkach dopasowywano sekwencje z wykorzystaniem algorytmu Needlemana-Wunscha.

5.1. Tryb Pokazowo-szkoleniowy

W trybie *Pokazowo-szkoleniowym* serwis OPAL wizualizuje poszczególne kroki działania wybranego algorytmu dopasowania sekwencji. Zanim rozpocznie się wizualizacja, użytkownikowi przedstawiona zostaje pusta macierz z wypełnionymi warunkami granicznymi. Kolejne wartości w warunkach granicznych zależne są od wybranego rodzaju kary. Zakładając, iż przyjęto karę za przerwę w postaci liniowej równej -5, wartości kolejnych komórek (rozpoczynając liczenie od prawego dolnego rogu macierzy) będą następujące: -5, -10, -15 i tak dalej, aż do ostatniej komórki. Wartości pozostałych komórek macierzy przyjmują wartość 0 (rys. 3a).

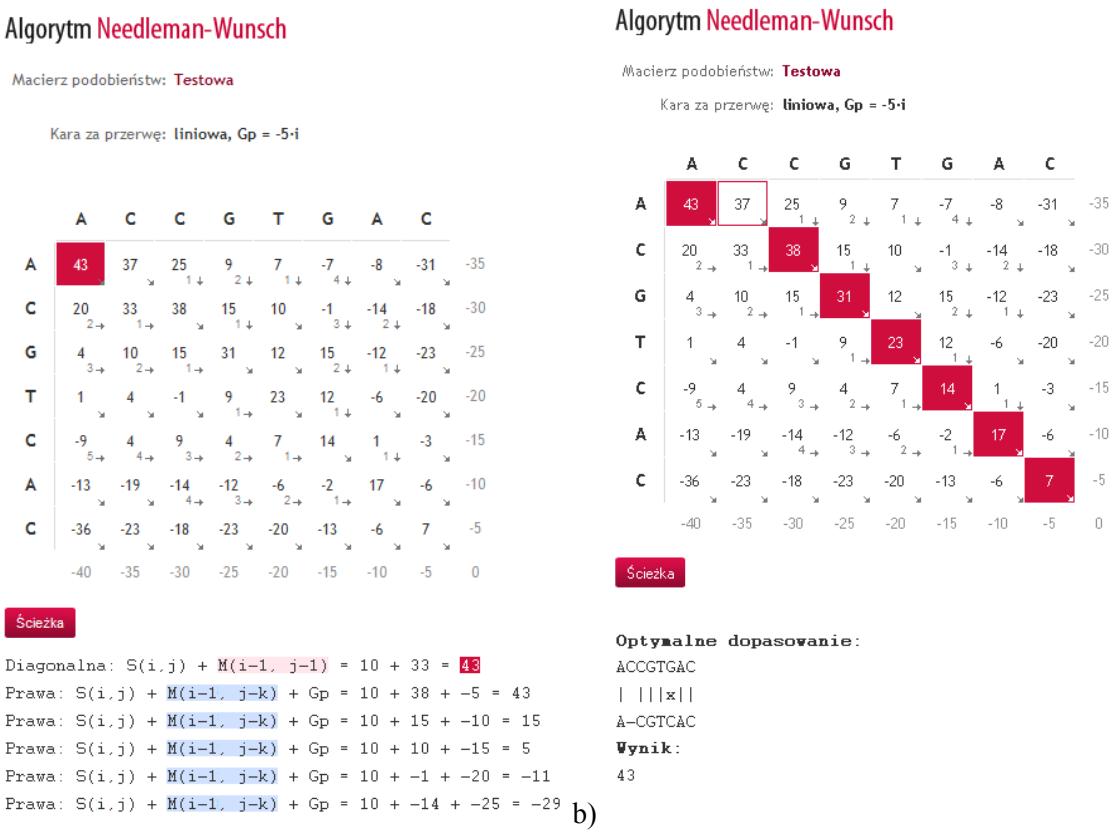
Naciśnięcie przycisku *Start* powoduje rozpoczęcie etapu uzupełniana macierzy dopasowania (rys. 3b). Dla oryginalnego algorytmu Needlemana-Wunscha pierwszą sprawdzaną komórką jest komórka leżąca w prawym dolnym rogu (dla pozostałych algorytmów w lewym górnym), następnie poruszając się w lewo i do góry (dla pozostałych algorytmów w prawo i w dół) uzupełniana jest cała macierz. Wraz ze wzrostem liczby wyznaczonych wartości, wzrasta liczba elementów, które należy uwzględnić do określania wartości kolejnych komórek. Wszystkie brane pod uwagę komórki oznaczane są odpowiednimi kolorami.



Rys. 3. Macierz podobieństwa: a) przed rozpoczęciem dopasowania, b) częściowo uzupełniona
Fig. 3. Similarity matrix: a) before alignment, b) partially completed

Ostatnim krokiem w każdej iteracji jest dodanie odpowiedniej strzałki kierunkowej i opcjonalnie cyfry przy niej, w prawym dolnym rogu komórki. Strzałka określa, z jakiego warunku (kierunku) uzyskana została aktualna wartość, cyfra z kolei wskazuje odległość do pola, które pozwoliło uzyskać tę wartość (rys. 3b).

W każdym momencie etapu uzupełniania macierzy dopasowania użytkownik może przerwać oraz następnie wznowić działanie algorytmu (przyciski *Start* i *Stop*), a także obserwować działania matematyczne, które doprowadziły do wyprowadzenia wartości danej komórki (rys. 4a). W chwili, gdy uzupełniona zostaje ostatnia komórka macierzy, można wyznaczyć ścieżkę dopasowania (przycisk *Ścieżka*) W przypadku algorytmu Needlemana-Wunscha wizualizacja ścieżki rozpoczyna się od komórki znajdującej się w lewym górnym rogu. Następnie przechodzi po komórkach mających najwyższą wartość w najbliższym sąsiedztwie, aż finalnie osiągnięty zostaje prawy dolny róg macierzy. Koniec etapu i zarazem algorytmu następuje w momencie dotarcia do ostatniej komórki (rys.k 4b).



a) b)

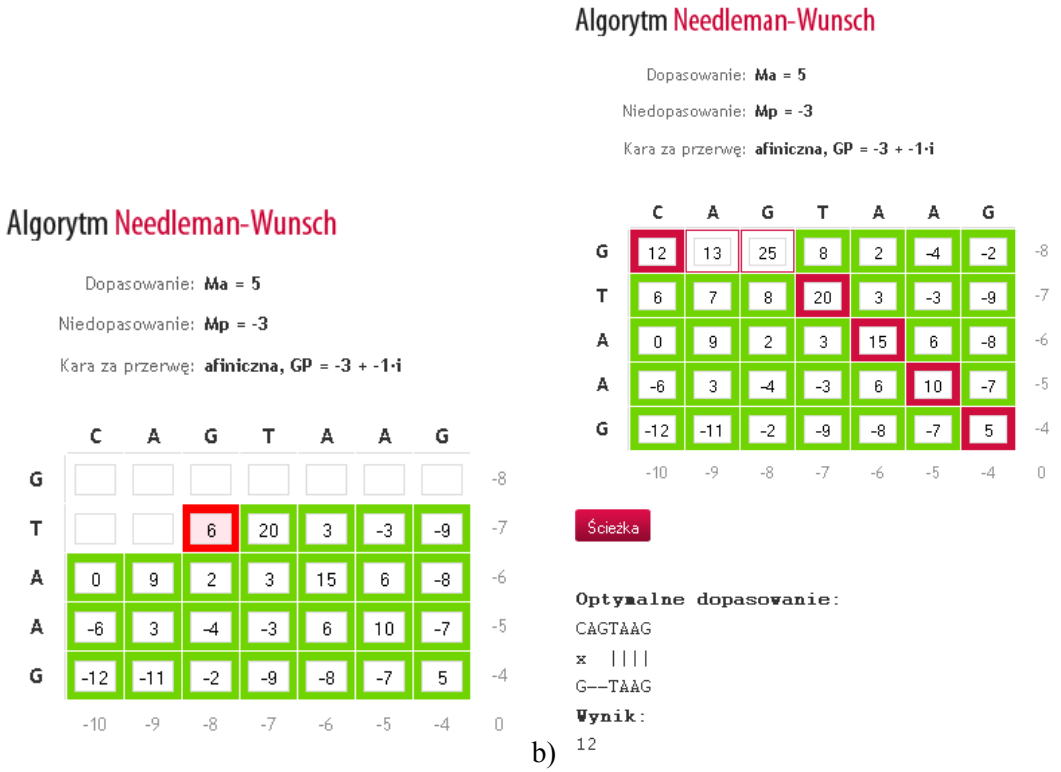
Rys. 4. Macierz podobieństwa: a) całkowicie wypełniona, b) z oznaczoną ścieżką dopasowania
Fig. 4. Similarity matrix: a) fully completed, b) with optimal alignment path

Po przejściu przez macierz na ekranie zostaje wyświetlone optymalne dopasowanie sekwencji względem zadanych parametrów. Pola macierzy, które zaznaczone zostały czerwonym tłem i białą czcionką oznaczają, iż elementy sekwencji są wzajemnie sparowane. Natomiast pola, które mają jedynie obramowanie w różowym kolorze określają przesunięcie w jednej z sekwencji (przerwę w dopasowaniu).

W optymalnym dopasowaniu wyświetlonym pod macierzą powyższe zależności zaznaczone zostały pomiędzy sekwencjami z wykorzystaniem znaków specjalnych takich, jak: dla dopasowania, x dla niedopasowania, spacja dla przesunięcia oraz kropka dla pozycji konserwatywnej, w przypadku porównywania aminokwasów. Pod dopasowaniem zamieszczono wartość numeryczną wyświetlonego dopasowania (rys. 4b).

5.2. Tryb Interaktywny

Tryb *Interaktywny* istotnie zmienia etap inicjalizacji macierzy dopasowania, ponieważ użytkownik ma w nim możliwość praktycznego sprawdzenia znajomości danego algorytmu, poprzez samodzielne wypełnienie macierzy podobieństwa. Zamiast wartości liczbowych w macierzy dopasowania wyświetlone zostają pola tekstowe, z jedną zaznaczoną kolorem komórką. Wyróżniona komórka określa pole, komórkę macierzy podobieństwa, którą w danym momencie należy wypełnić odpowiednią wartością, zgodnie z działaniem algorytmu. Wpisanie w polu wartości zgodnej z tą, jaką uzyskano by na podstawie wzorów, w wybranym algorytmie, powoduje oznaczenie komórki kolorem zielonym (rys. 5a).



a) b)

Rys. 5. Praca w trybie *Interaktywnym*: a) wypełnianie macierzy podobieństwa z kontrolą poprawności, b) macierz podobieństwa z oznaczoną ścieżką dopasowania

Fig. 5. *Interactive* mode: a) filling similarity matrix with accuracy control, b) similarity matrix with optimal alignment path

Natomiast, gdy wpisana wartość jest niepoprawna, komórka zostaje oznaczona kolorem czerwonym. Dzięki zastosowaniu podświetleń użytkownik na bieżąco informowany jest o postępach w uzupełnianiu macierzy.

Poprawne uzupełnienie ostatniej komórki daje możliwość pokazania ścieżki optymalnego dopasowania (przycisk *Ścieżka*, rys. 5b). Komórki, przez które przechodzi ścieżka, zostają oznaczone kolorem czerwonym, dla sparowanych elementów obu sekwencji lub kolorem białym, w przypadku wprowadzenia przerwy w dopasowaniu. Po wyznaczeniu ścieżki dopasowania pod macierzą wyświetlone zostaje optymalne dopasowanie oraz wartości miary dopasowania *Score*.

6. Podsumowanie

W przedstawionym artykule zaprezentowano serwis OPAL służący do wizualizacji wybranych algorytmów optymalnego dopasowania sekwencji nukleotydów i aminokwasów. Przedstawiono dwa z trzech dostępnych trybów działania aplikacji. Zasada działania trybu *Szybki wynik* jest identyczna jak w przypadku trybu *Pokazowo-szkoleniowego*, zmienia się jedynie forma prezentacji dopasowania sekwencji – następuje szybkie wyświetlenie wyniku dopasowania bez wizualizacji kolejnych etapów całego procesu.

W porównaniu do aplikacji B.A.B.A narzędzie OPAL jest znacznie bardziej rozbudowane, zwraca wyniki zgodne z aplikacjami komercyjnymi i umożliwia wprowadzanie różnych kar za przerwy w dopasowaniu (m.in. kary afinicznej). Serwis OPAL wizualizuje także działanie algorytmu Needlemana-Wunscha-Sellersa, oprócz występujących zwykle w tego typu aplikacjach algorytmów Needlemana-Wunscha i Smitha-Watermana. Dodatkowo, serwis OPAL pozwala na wyświetlenie wartości miary dopasowania oraz daje możliwość wyświetlenia wszystkich wykonywanych obliczeń.

W aplikacjach komercyjnych możliwości prezentacji wyników końcowych oraz wyświetlenia statystyk są znacznie bardziej rozbudowane, natomiast prezentacja pośrednich kroków wykonania algorytmów praktycznie nie ma miejsca. W narzędziu OPAL uwaga skoncentrowana została jednak na aspektach edukacyjnych. Serwis OPAL, w przeciwieństwie do narzędzi CLC Genomics Workbench i Geneious, obrazuje kolejne kroki dopasowania. Ponadto, w odniesieniu do wszystkich dostępnych aplikacji, narzędzie OPAL ma tryb interaktywny, który prowadzi użytkownika przez wszystkie etapy i pozwala na dogłębne zrozumienie działania algorytmów.

BIBLIOGRAFIA

1. Winter P. C., Hickey G. I., Fletcher H. L.: Genetyka. Krótkie wykłady, PWN, Warszawa 2008.
2. Baran A. A., Silverman K. A., Zeskand J., Koratkar R., Palmer A., McCullen K., Curran W. J., Bocker E. T., Siracusa L. D., Buchberg A. M.: The Modifier of Min 2 (Mom2) locus: embryonic lethality of a mutation in the Atp5a1 gene suggests a novel mechanism of polyp suppression, Genome Res, Vol. 17 (5), 2007, s. 566÷76.
3. Cormen T. H., Leiserson C. E., Rivest R. L.: Wprowadzenie do algorytmów. WNT, Warszawa 2001.
4. Internet: http://www.clcbio.com, dostęp: 16.02.2011.
5. Internet: http://www.geneious.com, dostęp: 16.02.2011.
6. Internet: http://zhanglab.ccmb.med.umich.edu/NW-align/, dostęp: 16.02.2011.
7. Internet: http://baba.sourceforge.net/, dostęp: 16.02.2011.
8. Lewiński K.: Porównywanie sekwencji białek i kwasów nukleinowych. Uniwersytet Jagielloński, Materiały dydaktyczne.
9. Rosenberg M. S.: Sequence Alignment. Methods, Models, Concepts, and Strategies. University of California Press, 2009.
10. Agarwal P. K., Vaidya M.: Local Alignment and Substitution Matrices. Algorithms in Computational Biology, Duke University, 2003.
11. Budd A.: European Molecular Biology Laboratory, Courses, Internet*:* http://www.embl.de/~seqanal/courses/tuebingenMpiPhyloMsaFeb2009/usingRaxml.html, dostęp: 21.01.2011.
12. Henikoff S., Henikoff J.: Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA, Vol. 89 (22), 1992, s. 10915÷10919.
13. 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, Vol. 48 (3), 1970, s. 443÷53.
14. Smith T.F., Waterman M.S., Fitch W.M.: Comparative Biosequence Metrics. J Mol Evol, Vol. 18 (1), 1981, s. 38÷46.
15. Smith T. F., Waterman M. S.: Identification of Common Molecular Subsequences. J Mol Biol, Vol. 147 (1), 1981, s. 195÷197.
16. Waterman M. S.: Efficient Sequence Alignment Algorithms, J Theor Biol, Vol. 108 (3), 1984, s. 333÷337.
17. Sellers P. H.: On the theory and computation of evolutionary distances. SIAM J Appl Math, Vol. 26 (4), 1974, s. 787÷793.
18. Vingron M., Waterman M.S.: Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol, Vol. 235 (1), 1994, s. 1÷12.

Recenzent: Dr Ewa Romuk

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

Visualization of optimal alignment algorithms for nucleotide and protein sequences is very helpful in understanding basic rules of how these algorithms work. It also has a great potential to illustrate biological issues that have to be taken into account.

In the paper, we show the OPAL service, which is a web application for visualization of optimal alignment algorithms of nucleotide and protein sequences. The main objective of the OPAL application is to match nucleotide and amino acid sequences in accordance with the following algorithms: Needleman-Wunsch (original), Smith-Waterman and Needleman-Wunsch-Sellers (simplified Needlemana-Wunsch algorithm) [9], [13-17]. The most important elements in determining the actual best alignment are cost function and substitution matrix. Different sets of scoring values may lead to different optimal alignments and the best alignment is not only dependent on the algorithm, but also on chosen parameter [8-12].

In this paper, we used sample sequences to illustrate filling the scoring matrices and finally to show the traceback route giving a globally optimum alignment. This article presents two of three modes of the OPAL application (*Visualization and training* mode and *Interactive* mode). The rules of the third mode (*Fast result*) remain unchanged, since the change concerns only the form of presentation.

Adresy

Adam SKOWRON: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, adam.skowron@polsl.pl.
Dariusz MROZEK: Politechnika Śląska, Instytut Automatyki, ul. Akademicka 16, 44-101 Gliwice, Polska, dariusz.mrozek@polsl.pl.