Rearanżacje genomu, a sortowanie przez inwersję.

Treść seminarium

- Podstawy biologiczne
- Trochę o inwersjach
- Model matematyczny
- Algorytm 4-przybliżony dla sortowania bez znaku
- Permutacje ze znakiem:
  - Wstęp
  - Graf breakpointowy
  - Cykliczna dekompozycja
  - Lepsza granica
  - Transformacja permutacji
  - Poprawna inwersja
  - Krawędzie i cykle zorientowane

Treść seminarium – cd…

- Krawędzie i cykle zorientowane
- Grafy przekładane
- Składowe spójności
- Częściowy porządek
- Przeszkody
- (g,b)-split
- Permutacja uogólniona
- (Safe) (g,b)-padding
- Bezpieczna inwersja
- Graf pokrycia
- Super przeszkody i proste przeszkody
- Odcinanie i bezpieczne odcinanie przeszkód
- Łączenie i bezpieczne przeszkód
- Forteca i 3-Forteca
- Dokładna odległość inwersyjna

- Algorytmy
- Pytania

Historycznie

- Początkowo uwaga skupiała się na porównywaniu łańcuchów reprezentujących pojedyncze geny, czy białka.
  - Mutacje punktowe
  - Podmiany aminokwasów
  - Pojedyncze insercje i delecje nukleotydów
- Obecnie nacisk kładzie się na badanie bardziej rozległych mutacji, co daje bardziej rozległe spojrzenie na ewolucję genomu.

Rearanżacje genomu

- Rozległe mutacje, takie jak:
  - Delecje i insercje dużych fragmentów nici
  - Translokacja
  - Transpozycja
  - Inwersje
- Zwykle znacznie wolniejsze niż mutacje w obrębie pojedynczych genów.
- Dobre narzędzie do dedukowania procesu ewolucji w historii.
- Często występują u roślin (np. kapustowate)

Inwersje, odległość inwersyjna

- Inwersje są dominującą formą mutacji w chromosomie, choć w ogólności są stosunkowo rzadkie i mają charakter pozornie losowy.
- Do badania mamy dwa łańcuchy reprezentujące kolejność genów w chromosomie i próbujemy poznać jak mógł wyniknąć (lub jak faktycznie wynikł) drugi łańcuch z pierwszego.
- Niestety pełen model inwersji w DNA nie jest znany i nikt nie wie jak można by takiej rekonstrukcji dokonać.

Odległość inwersyjna

- Odległość inwersyjną będziemy określać jako:

  Def. Odległość inwersyjna, to minimalna ilość możliwych transformacji łańcucha.

- Jest to problem NP-trudny.
- Zaprezentujemy dwie heurystyki…

Geny w chromosomie

- Występują zwykle tylko w jednej kopii.
- Można więc uznać, że są rozróżnialne i numerowane.
- Tym samym możemy numerować geny w chromosomie, co będziemy nazywać **permutacją**.

Definicje i fakty

- Permutacja liczb 1..n

  $\Pi = \pi_1, \pi_2, \pi_3, \ldots, \pi_n$

  np. $\Pi = 3, 2, 1, 4, 5, 6, 7$

  $\pi_1 = 3, \pi_2 = 2, \ldots$

- Permutacja tożsamościowa

  $\pi_1 = 1, \pi_2 = 2, \pi_3 = 3, \ldots, \pi_n = n$

Definicje i fakty

- Breakpoint

  występuje między π_i, π_{i+1} dla $1 \leq i \leq n-1$ wtedy i tylko wtedy, gdy $|\pi_i - \pi_{i+1}| \neq 1$, na początku Π gdy $\pi_i \neq 1$ oraz na końcu Π gdy $\pi_n \neq n$.

- $\mathbf{\Phi(\Pi)}$ - liczba breakpointów w permutacji $\mathbf{\Pi}$

- Przykład:

  Permutacja $\Pi = 3, 2, 4, 5, 1$ zawiera breakpointy na początku i na końcu, między 2 a 4 oraz między 5 a 1.

Definicje i fakty

- Strip – maksymalna sekwencja w Π nie zawierająca breakpointa.
  - np: $\Pi = \underline{1}, \underline{5, 6, 7}, \underline{4, 3, 2}$
- Rosnący strip – taki, w którym liczby sukcesywnie rosną.
- Malejący strip – taki, w którym liczby sukcesywnie maleją, lub stip pojedynczy.

Lemat 1

- Odległość inwersyjna dowolnej permutacji Π, wynosi co najmniej $\Phi(\Pi)/2$.

- Dowód:

  Jedna operacja inwersji może zredukować liczbę breakpointów o co najwyżej dwa (na początku i na końcu inwersji).

  Musimy więc wykonać przynajmniej połowę jedną inwersję dla każdych dwóch breakpointów.

Lemat 2

- Niech Π nie będzie permutacją tożsamościową, oraz nie będzie zawierała stripów malejących.
  Istnieje wówczas inwersja w Π, taka że nie zwiększy ona liczby breakpointów, a jednocześnie wynikowa permutacja zawierała będzie strip malejący.

- Dowód:
  - Jeżeli nie ma stripów malejących, to każdy strip jest długości co najmniej dwa.
  - Jeżeli nie jest to permutacja tożsamościowa, to istnieje strip mający breakpoint na oku końcach.
  - Inwersja tego stripa tworzy strip malejący bez zwiększania liczby breakpointów.

Lemat 3

- Jeżeli Π zawiera strip malejący, to istnieje inwersja, która zmniejsza ilość breakpointów o co najmniej jeden.

- Dowód...

Algorytm 4-przybliżony

while (istnieje breakpoint w permutacji)

begin

if (istnieje malejący strip)

znajdź i odwróć taki malejący strip, który zmniejsza liczbę breakpointów (z lematu 3.)

else

znajdź i odwróć taki rosnący strip, który nie zwiększy liczby breakpointów (z lematu 2.)

end

Sortowanie inwersyjne – wersja ze znakiem

- Znaczenie problemu

Znakowane permutacje lepiej opisują ewolucję DNA. W podwójnej helisie DNA „działające” geny znajdują się tylko na jednej z dwóch nici.

Sortowanie inwersyjne – wersja ze znakiem

- w wersji ze znakiem każda liczba w permutacji posiada znak (+ lub –), który zmienia się gdy liczba należy do odwracanego fragmentu permutacji.
- Definicja problemu

Przekształcić daną znakowaną permutację w permutację tożsamościową, w której wszystkie znaki są dodatnie.

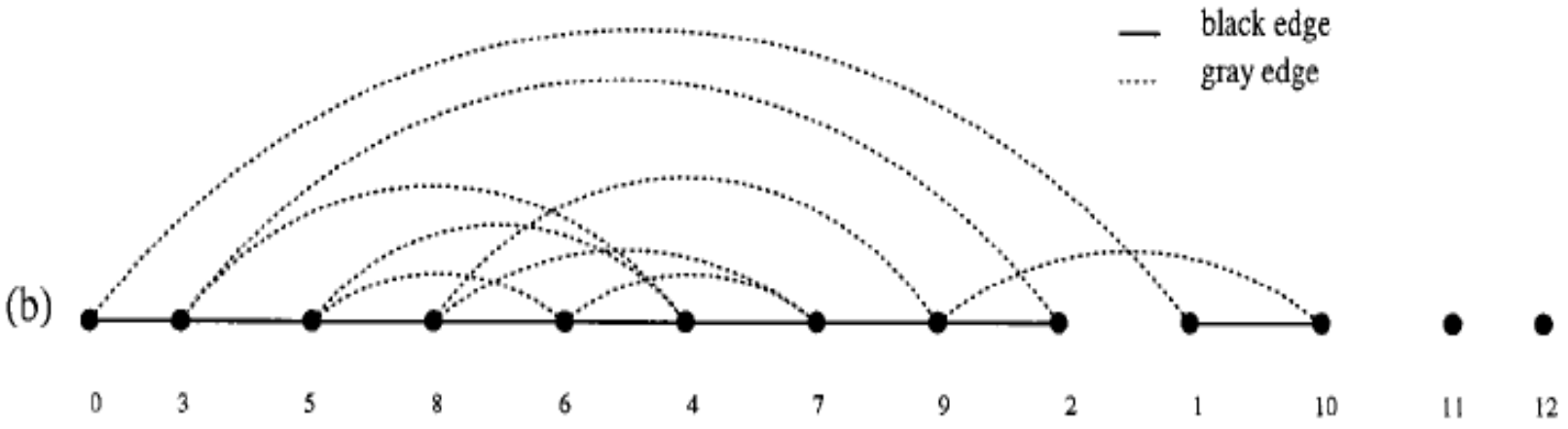
Permutacja

- Tworzymy permutację nie znakowaną według redukcji.
- Znaki nadajemy przypisując jakiejś wartości z Π znak +, a każdej innej znak – wtedy i tylko wtedy, gdy reprezentuje gen znajdujący się na tej samej nici w dokładnie jednym z dwóch porządków DNA.

Graf breakpointowy - konstrukcja

- Rozszerzamy permutację Π o $\pi_0 = 0$ i $\pi_{n+1} = n+1$
- Graf G(Π) kolorowany krawędziowo o n+2 wierzchołkach $\{\pi_0, \pi_1, \ldots, \pi_n, \pi_{n+1}\} = \{0, 1, \ldots, n, n+1\}$ i krawędziach zdefiniowanych następująco:
  - Czarna krawędź – łączy wierzchołki πi i πj gdy (πi, πj) jest breakpointem w Π.
  - Szara krawędź – łączy wierzchołki π_i i π_j gdy (i, j) jest breakpointem w Π^{-1}(zbiór indeksów Π).

Graf breakpointowy - przykład



Cykliczna dekompozycja

- Cykl zmienny – cykl, w którym każde dwie kolejne krawędzie mają różne kolory.
- $l(C)$ – długość cyklu, liczba czarnych (lub szarych) krawędzi.
- $l(C) = 2$ – cykl krótki
- $l(C) > 2$ – cykl długi
- $c(\Pi)$ – maksymalna liczba krawędziowo rozłącznych cykli zmiennych należących do cyklicznej dekompozycji grafu G.

Cykliczna dekompozycja

- Permutacja prosta
- Każda inwersja zmienia parametr $b(\pi) - c(\pi)$ o co najwyżej 1.
- $d(\pi) >= b(\pi) - c(\pi)$

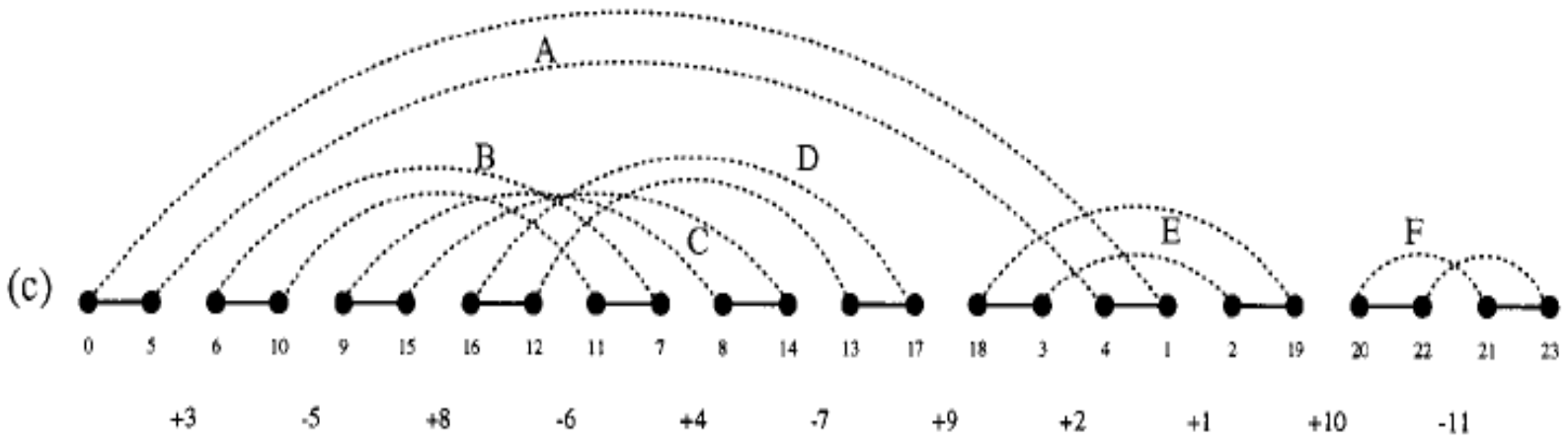
Dokładniejsza granica

- $d(\pi) >= b(\pi) - c(\pi) + h(\pi)$

gdzie $h(\pi)$ – ilość przeszkód

Transformacja permutacji

- Transformacja permutacji ze znakiem (Π o rozmiarze n) w permutację bez znaku (Π' o rozmiarze 2n)
  - Ujemne wartości x zastępuje sekwencja 2x, 2x-1
  - Dodatnie wart. x zastępuje sekwencja 2x-1, 2x
- Permutacja Π' nazywa się obrazem permutacji Π


(c)
A
B
C
D
E
F
0 5 6 10 9 15 16 12 11 7 8 14 13 17 18 3 4 1 2 19 20 22 21 23
+3 -5 +8 -6 +4 -7 +9 +2 +1 +10 -11

Poprawna inwersja

- $\Delta b \equiv \Delta b(\Pi, \rho) = b(\Pi\rho) - b(\Pi)$

$\Delta c \equiv \Delta c(\Pi, \rho) = c(\Pi\rho) - c(\Pi)$

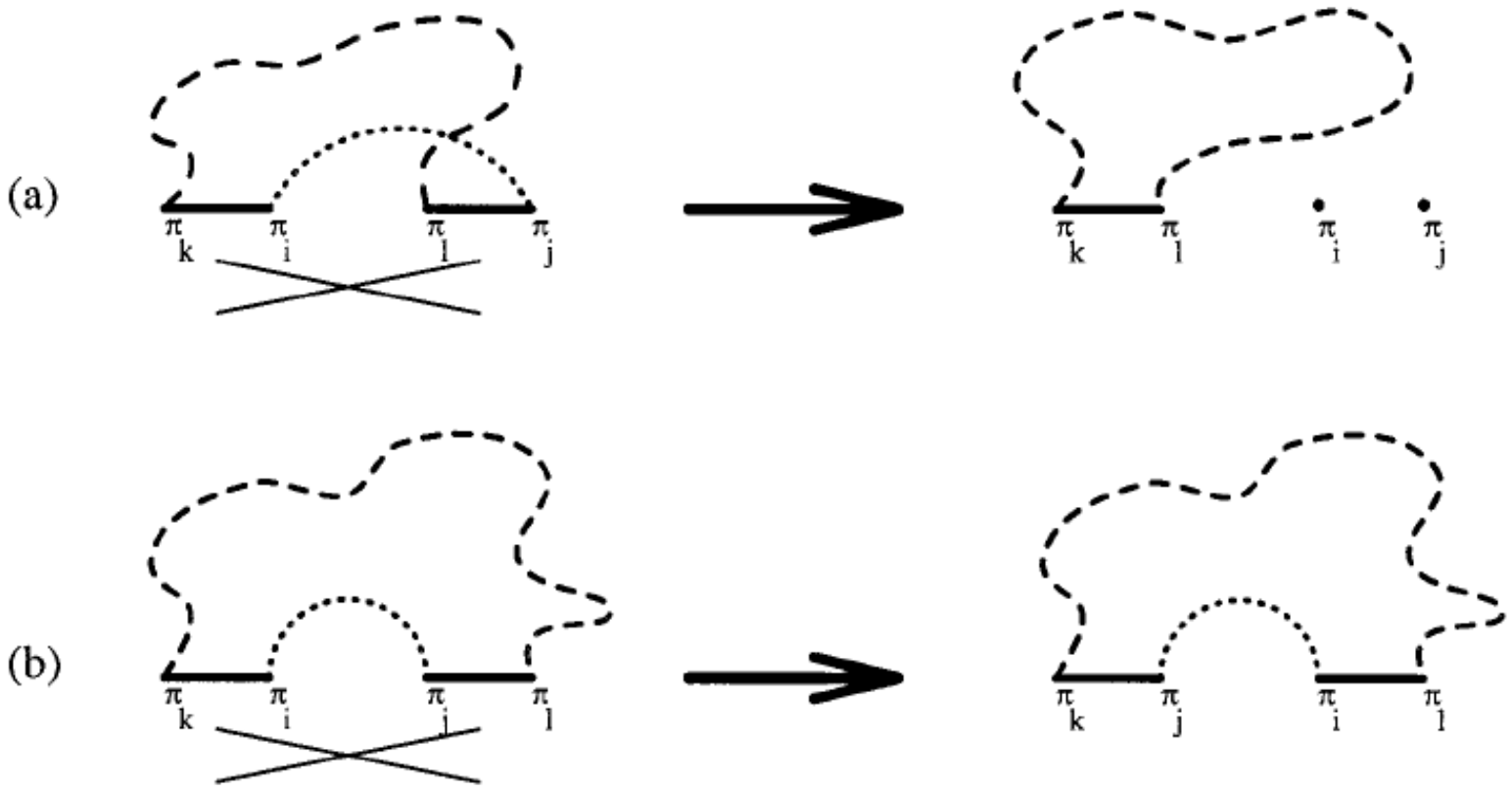
$\Delta(b - c) \equiv \Delta b(\Pi, \rho) - \Delta c(\Pi, \rho) \geq -1$

- inwersja jest poprawna gdy $\mathbf{\Delta(b - c) = -1}$.

Krawędzie zorientowane

- Szara krawędź g jest zorientowana gdy inwersja dwóch czarnych krawędzi incydentnych do g jest poprawne. W innym przypadku krawędź g jest niezorientowana.
- Lemat
  - Niech (π_i, π_j) będzie szarą krawędzią incydentną do czarnych krawędzi (πk, πi) i (π_j, π_l). Wtedy (π_i, π_j) jest zorientowana wtedy i tylko wtedy, gdy i – k = j – l.

Dowód..



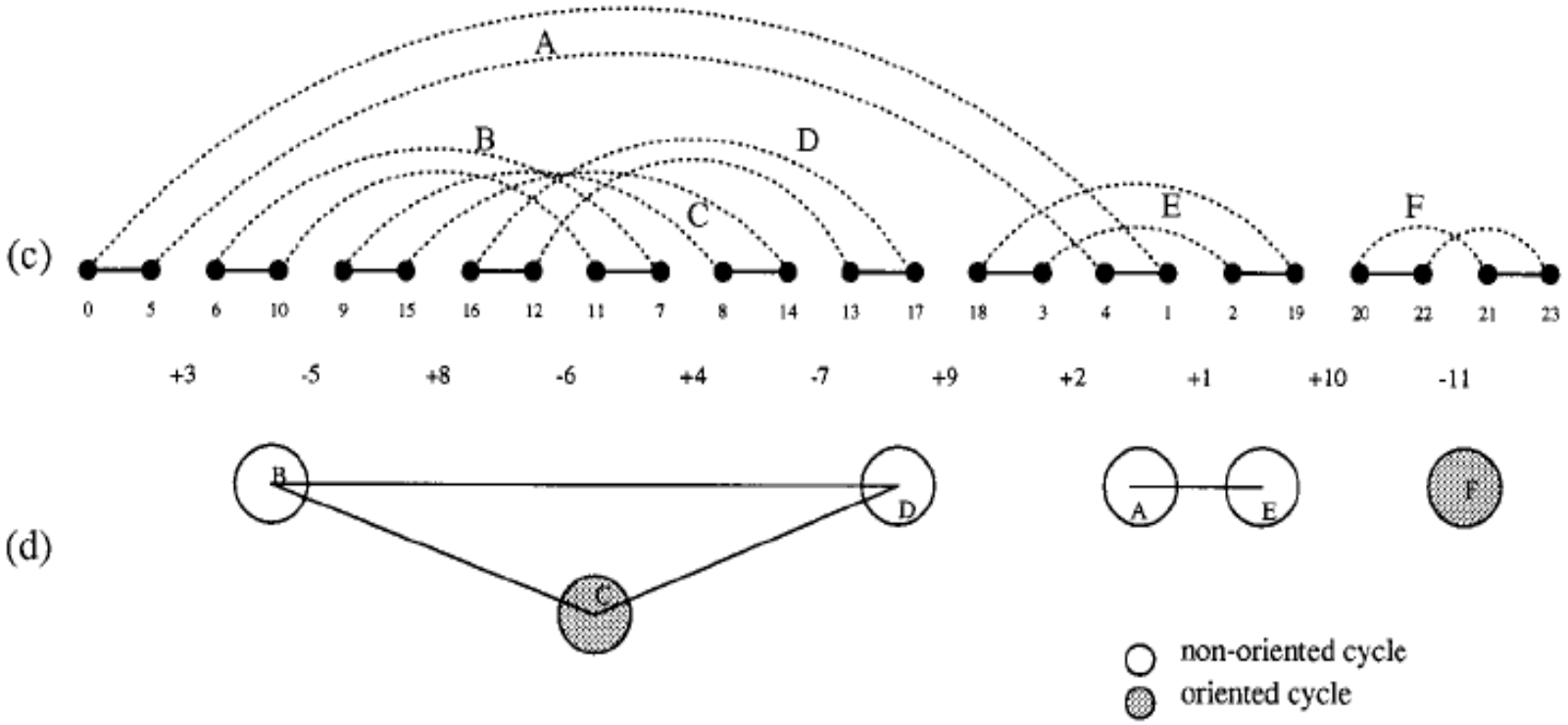
Cykl zorientowany

- Cykl nazywamy zorientowanym gdy zawiera co najmniej jedną zorientowana krawędź
- W niezorientowanych cyklach nie istnieje poprawna inwersja.

Grafy przekładane $H_{\Pi}(C_{\Pi}, E_{\Pi})$

- Krawędzie przekładane – szare krawędzie (π_i, π_j) i (π_k, π_t) są przekładane gdy przedziały [i, j] i [k, t] nakładają się, ale nie zawierają.
- Cykle przekładane – cykle C1 i C2 są przekładane jeżeli istnieją szare krawędzie g1 należy do C1 i g2 należy do C2

Składowe spójności



Składowe spójności

- Składowa spójności grafu H_{Π} jest zorientowana jeżeli zawiera przynajmniej jeden zorientowany wierzchołek (zorientowany cykl z cyklicznej dekompozycji grafu breakpointowego).
- Extent(U) - przedział [Umin, Umax];

$$U_{\min} = \min_{C \in U} \min_{\pi_i \in C} i \qquad \text{and} \qquad U_{\max} = \max_{C \in U} \max_{\pi_i \in C} i$$

Składowe spójności

- Mówimy, że składowa spójności U separuje składowe U’ i U” w Π jeżeli istnieje szara krawędź (π_i, π_j) taka, że Extent(U’) zawiera się w przedziale [i, j], ale część wspólna Extent(U”) i [i, j] jest zbiorem pustym.

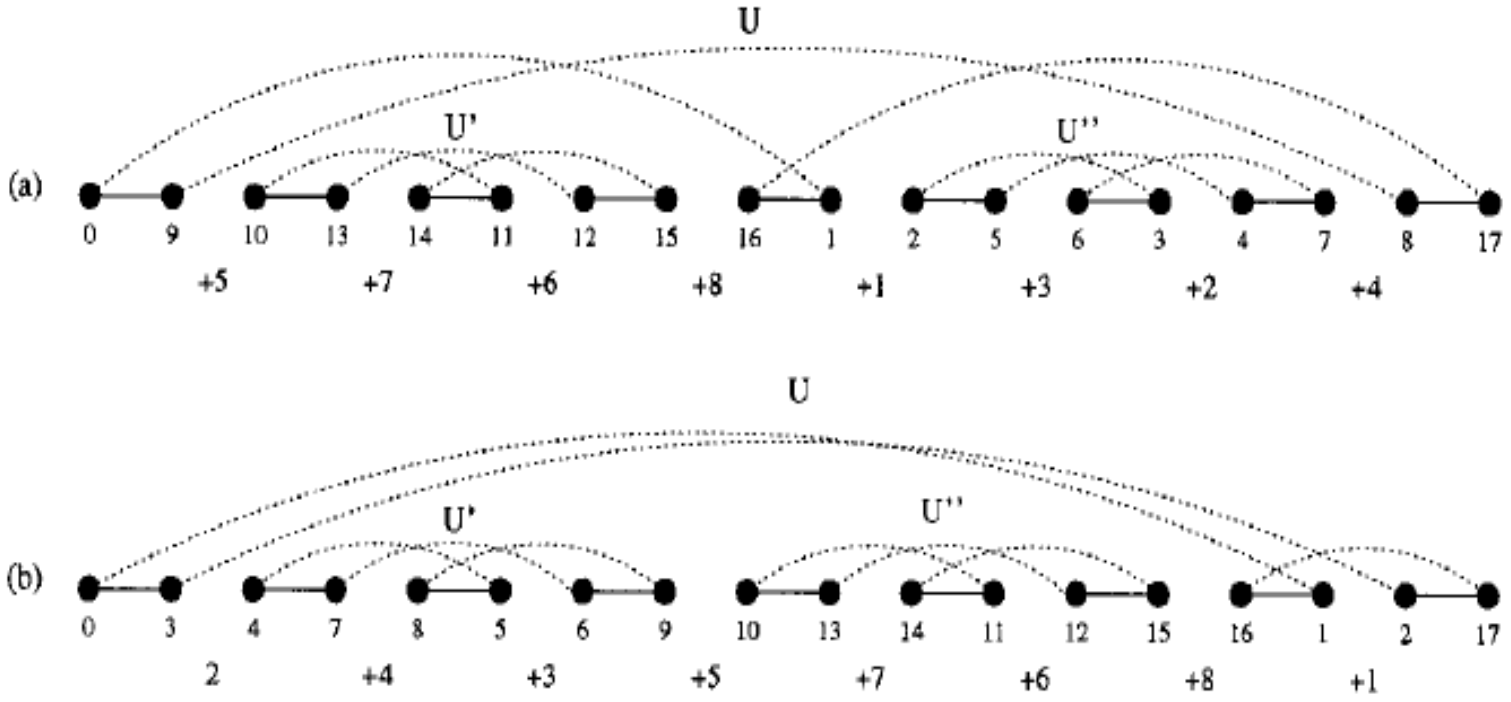
Częściowy porządek

- Częściowy porządek r zawierania na zbiorze niezorientowanych składowych spójności U_{Π} w H_{Π} jest zdefiniowany następująco:
- U r W $\leftrightarrow$ Extent(U) zawiera się w Extent(W) dla U, W należących do U_{Π}.

Przeszkody (hurdles)

- Minimalny hurdle (przeszkoda) to niezorientowana składowa spójności minimalna w relacji r w zbiorze U_{Π}.
- Największy hurdle (przeszkoda) to niezorientowana składowa spójności największa w relacji r w zbiorze U_{Π} i nie separująca żadnych dwóch minimalnych przeszkód.
- $h(\Pi)$ – liczba przeszkód w grafie H_{Π}.

Przeszkody (hurdles)

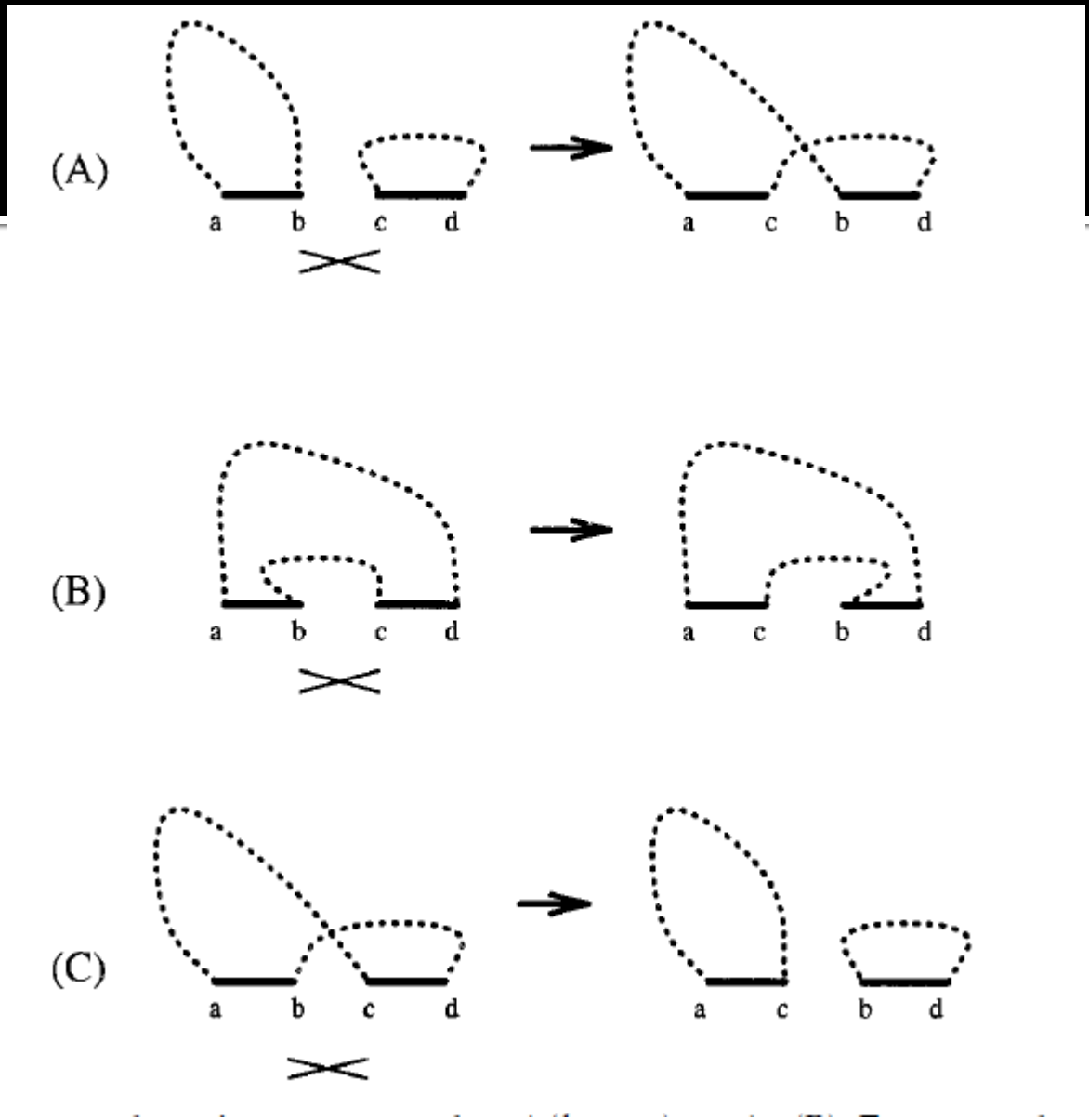


Twierdzenie

- Dla znakowanych permutacji

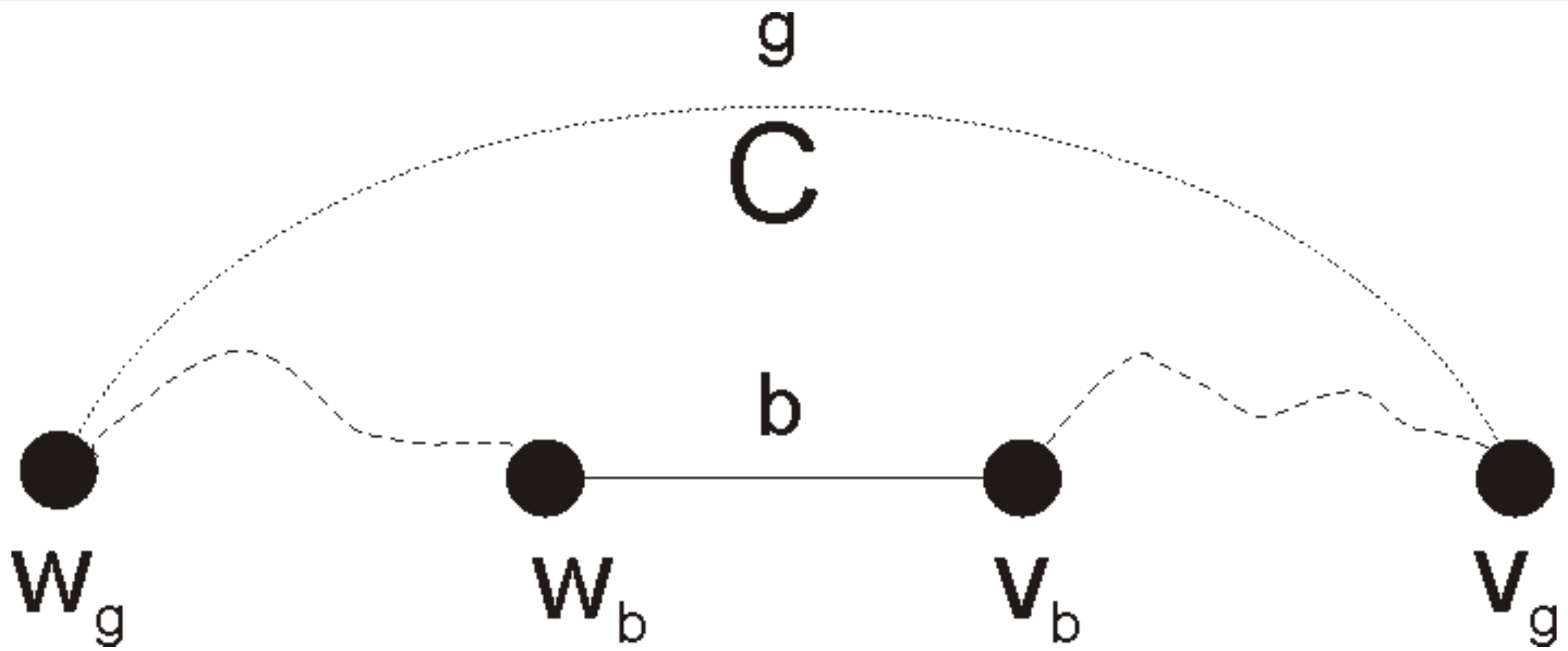
$$d(\Pi) \geq b(\Pi) - c(\Pi) + h(\Pi).$$

Dowód



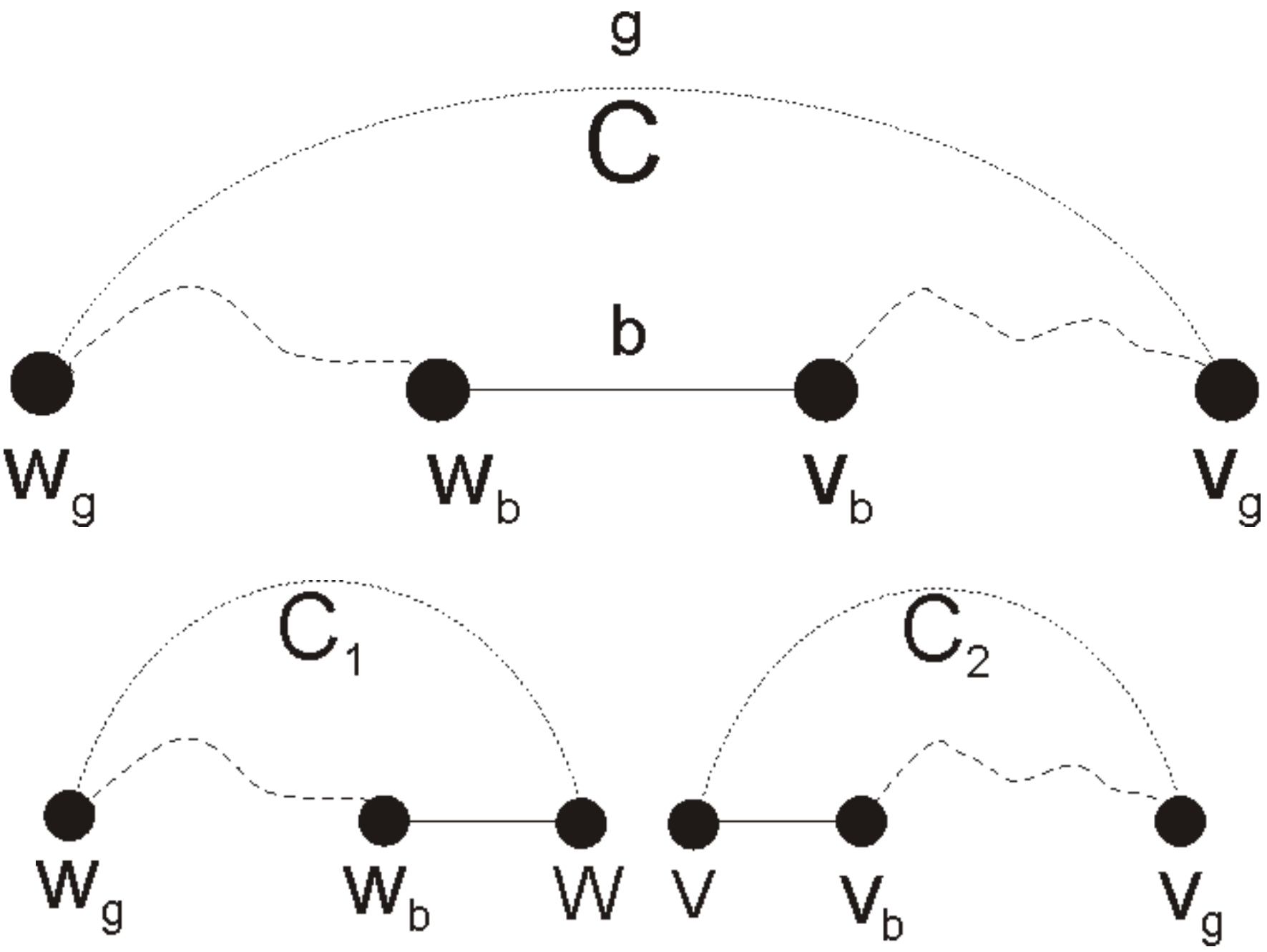
(g, b)-split

- Niech b = (v_b,w_b) – czarna krawędź
- Niech g = (w_g,v_g) – szara krawędź
- b i g należą do cyklu C=…,v_b,w_b, …, w_g,v_g, … w grafie breakpointowym G(Π) permutacji Π.


g
C
b
w_g
w_b
v_b
v_g

(g,b)-split

- Przez przycięcie (g,b) rozumiemy nowy graf G'(Π) powstały z grafu G(Π) poprzez:
- Usunięcie krawędzi g i b
- Dodanie dwóch nowych wierzchołków v i w
- Dodanie dwóch nowych czarnych krawędzi (wb, v) i (w, wb)
- Dodanie dwóch nowych szarych krawędzi (wg, w) i (v, vg)


g
C
b
w_g
w_b
v_b
v_g
C_1
C_2
w_g
w_b
W
V
v_b
v_g

Permutacja uogólniona

- Przez permutację uogólnioną π" rozumiemy permutację π rzutowaną na liczby rzeczywiste

(g, b)-padding

- Niech $b = (\pi_{i+1}, \pi_i)$ – czarna krawędź
- $g = (\pi_j, \pi_k)$ – szara krawędź
- Krawędzie b i g należą do cyklu $C = \ldots, \pi_{i+1}, \pi_i, \ldots, \pi_j, \pi_k, \ldots$ w grafie breakpointowym $G(\Pi)$.

(g, b)-padding

Niech:

- $\Delta = \pi_k - \pi_j$,
- $v = \pi_j + (\Delta /3)$,
- $w = = \pi_k + (\Delta /3)$.

(g, b)-padding na Π jest permutacją n+2 elementów otrzymaną z Π przez wpisanie v i w za i-tym elementem.

- $\pi'' = (\pi_1\ \pi_2 \ldots \pi_i\ vw \pi_{i+1} \ldots \pi_n)$

Safe (g, b)-padding

Rozbija długie cykle na mniejsze.

Bezpieczna inwersja i Graf pokrycia

- **Bezpieczna inwersja**

  inwersja jest bezpieczna jeżeli $\Delta(b - c + h) = -1$
- **Graf pokrycia**

  Niech rel będzie częściowym porządkiem na zbiorze P. x **jest pokryte przez** y w P gdy x rel y oraz nie istnieje z należący do P taki, że x rel z rel y. Graf Ω relacji rel o zbiorze wierzchołków P i krawędzi {(x, y): x, y należy do P i x jest pokryte przez y}.

Super przeszkody i proste przeszkody

- **Super przeszkody**
  Przeszkoda v **chroni** niezorientowaną składową U, która nie jest przeszkodą, jeżeli usunięcie v zmieni U w przeszkodę.
  Przeszkoda w Π jest super przeszkodą jeżeli chroni niezorientowaną składową U.
- **Prosta przeszkoda**
  Przeszkoda w Π jest prostą przeszkodą, jeżeli nie chroni żadnych niezorientowanych składowych.

(Bezpieczne) odcinanie przeszkód

- **Odcinanie przeszkód**

  Każda inwersja ρ na cyklu należącym do przeszkody K odcina liść K z grafu pokrycia permutacji Π.

  $\Omega_{\Pi\rho} = \Omega_{\Pi} \setminus K$.

- **Bezpieczne odcinanie przeszkód**

  Inwersja działająca na cyklu należącym do prostej przeszkody jest bezpieczna.

Łączenie przeszkód

Inwersja ρ, działająca na czarnych krawędziach należących do przeszkód L i M, łączy zbiór niezorientowanych składowych na ścieżce łączącej liść L z liściem M w grafie pokrycia z ostatnim wspólnym przodkiem L i M, który ich nie separuje. $\Omega_{\Pi\rho} = \Omega_{\Pi}(L, M)$.

Bezpieczne łączenie przeszkód

- Piszemy U < W jeżeli $U_{max} < W_{max}$

  Ustawiamy przeszkody w kolejności:

  U(1)<…<U(l) ≡ L<…<U(m) ≡ M<…<U(h(Π))
- definiujemy zbiory przeszkód:
  - BETWEEN(L, M) = {U(i): l<i<m}
  - OUTSIDE(L, M) = {U(i): i nie należy do [l, m]}
- Jeżeli ρ jest inwersją łączącą przeszkody L i M i oba zbiory BETWEEN(L, M) i OUTSIDE(L, M) są puste to inwersja jest bezpieczna.

Forteca i 3-forteca

- **3-forteca**

  Jeżeli graf pokrycia Ω_{Π} jest grafem homeomorficznym z 3-gwiazdą z trzema super przeszkodami, to te super przeszkody nazywamy 3-fortecą.

- **Forteca**

  Permutacja Π jest fortecą jeżeli liczba przeszkód jest nieparzysta i wszystkie są super przeszkodami.

Wzór dokładny na odległość inwersyjną

- Dla każdej permutacji Π, d(Π) jest równe:
  - b(Π) - c(Π) + h(Π) + 1, gdy Π jest fortecą
  - b(Π) - c(Π) + h(Π), gdy Π nie jest fortecą

Algorytm sortowania przez inwersje (ze znakiem)

while (Π nie jest posortowane)
    if (Π zawiera długi cykl)
        zaznacz inwersję safe (g, b)-padding ρ w Π
    else if (Π zawiera zorientowaną składową)
        zaznacz bezpieczną inwersję ρ w tej składowej
    else if (Π zawiera parzystą liczbę hurdle'i)
        zaznacz bezpieczną inwersję ρ łączącą dwa hurdle
    else if (Π zawiera co najmniej jeden simple hurdle)
        zaznacz bezpieczną inwersję ρ obcinającą hurdle'a
    else if (Π jest fortecą z więcej niż trzema superhurdle)
        zaznacz bezpieczną inwersję ρ łączącą dwa (super)hurdle
    else /* Π jest 3-fortecą */
        zaznacz (nie)bezpieczną inwersję ρ łączącą dwa (super)hurdle
    $\Pi \leftarrow \Pi \bullet \rho$
endwhile

Uproszczony algorytm sortowania przez inwersje (ze znakiem)

- f(Π) jest równe
  - 1, jeżeli Π jest fortecą
  - 0, jeżeli Π nie jest fortecą
- Inwersja jest legalna jeżeli

  $\Delta(b - c + h + f) = -1$

Uproszczony algorytm sortowania przez inwersje (ze znakiem)

- Dla każdej permutacji Π istnieje legalna inwersja. Każda sekwencja legalnych inwersji jest optymalna.
- while (Π nie jest posortowane)

  zaznacz legalną inwersję ρ

  $\Pi \leftarrow \Pi \bullet \rho$

endwhile

Przykład działania algorytmu

PYTANIA?