Wprowadzenie do Informatyki Biomedycznej

Wykład 4:

Mikromacierze, analiza ekspresji genów

Marek Kretowski
pokój 206
e-mail: m.kretowski@pb.edu.pl
http://aragorn.pb.bialystok.pl/~mkret

Wprowadzenie

- Mikromacierze są jednym z najnowszych osiągnięć (prawdziwy przełom) w dziedzinie eksperymentalnej biologii molekularnej;
  - pozwalają na monitorowanie ekspresji tysięcy genów równolegle i produkują niewyobrażalne ilości danych
  - Aktualnie celem badań genomicznych jest przestawienie się z sekwencjonowania na wykorzystanie sekwencji genomu do zrozumienia sposobu jego funkcjonowania
- Dane zawarte w macierzach ekspresji genów mogą być analizowane na 3 poziomach (analiza jest w tej chwili wąskim gardłem tej technologii)
  - pojedyncze geny - próbuje się ustalić czy rozpatrywany gen (w izolacji od pozostałych) zachowuje się odmiennie w warunkach szczególnych i kontrolnych
  - wiele genów - grupy genów są analizowane pod względem wspólnej funkcjonalności, wzajemnego oddziaływania i regulacji, ...
  - na trzecim poziomie podejmowane są próby wnioskowania o (nieznanej) sieci powiązań genów i białek, która leży u podstaw obserwowanych wzorców

Mikromacierze (ang. microarrays)

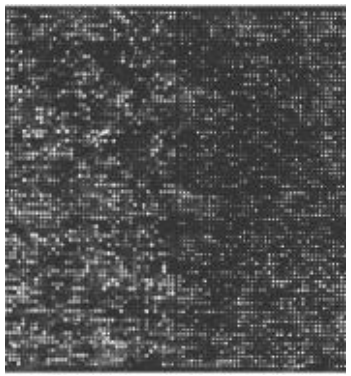
- Technologia wykorzystująca zdobycze analizy sekwencji stworzone w ramach projektów analizy genomów w celu odpowiedzenia na pytanie:

Które geny są aktywne (ang. expressed) w konkretnym typie komórek organizmu, w danym momencie, w określonych warunkach?

  - np. porównanie profili ekspresji genów pomiędzy zdrową i chorą komórką
- Różne nazwy: DNA microarrays, DNA arrays, DNA chips, gene chips, ...
- Mikromacierz jest zwykle szklaną płytką (ang. slide) na której cząsteczki DNA są przytwierdzane w ustalonych miejscach (ang. spots)
  - nawet rzędu dziesiątek tysięcy punktów, z których każdy zawiera ogromną liczbę identycznych cząsteczek DNA (łańcuchów o długości od 20 do kilkuset nukleotydów)
  - w celu wykorzystania do analizy ekspresji genów, każde położenie powinno zawierać jeden gen lub ekson genomu (w praktyce nie zawsze łatwe do osiągnięcia)

Przykład podświetlonej mikromacierzy

- Typowy rozmiar macierzy to około 1 cala lub mniej; promień punktu rzędu 0.1mm (w niektórych nawet mniej)



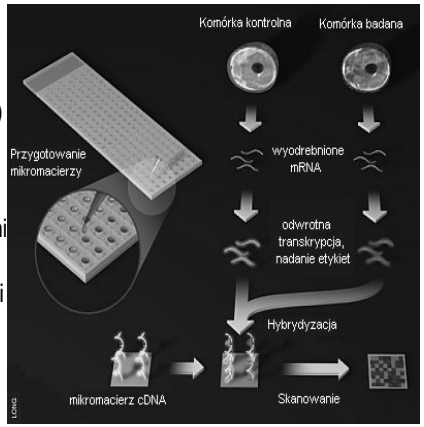
- Punkty na macierzy są drukowane przez robota, syntetyzowane przez fotolitografię (podobnie jak podczas produkcji układów scalonych) lub drukowane jak w drukarce atramentowej
- Mikromacierze zawierające wszystkie około 6000 genów z genomu drożdży były dostępne od 1997
- Istnieją różne sposoby wykorzystania technologii mikromacierzy do pomiarów ekspresji genów
- Jedno z najpopularniejszych zastosowań pozwala na porównywanie poziomów ekspresji genów w dwóch różnych próbkach np. tego samego typu komórki w stanie normalnym i patologicznym

Rodzaje mikromacierzy

- Macierze z komplementarnym DNA (cDNA)
  - użyte w nich kwasy nukleinowe (łańcuchy nukleotydów) są produktami PCR (Polymerase Chain Reaction) uzyskiwanymi w procesie replikacji ze zbiorów cDNA
  - kwas nukleinowy na powierzchni płytki umieszczany jest poprzez maszynowe nakrapianie (ang. robotically spotted)
  - są to macierze różnicowe
  -
- Macierze oligonukleotydowe (oligomacierze)
  - synteza *in situ* (łac. in situ - na miejscu) przez fotolitografię
  - na macierzach umieszczane są krótsze łańcuchy oligonukleotydów (zwykle około 25 par baz, ale może być nawet 50-70), a każdy gen jest reprezentowany przez wiele oligonukleotydów
  - fragmenty DNA są tak dobierane, aby była jak najmniejsza reaktywność krzyżowa z innymi genami a przez to nieswoista hybrydyzacja będzie zminimalizowana (ale i tak występuje => druga sekwencja ze zmienioną środkową bazą)

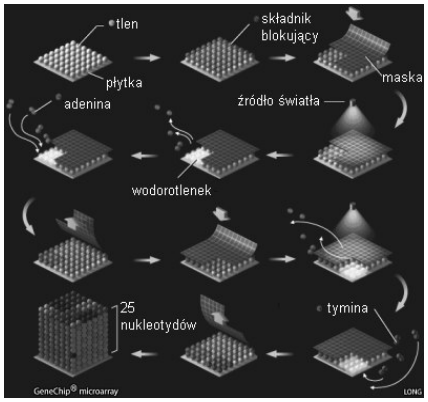
Macierze różnicowe

- Pełne mRNA z komórek w dwóch różnych warunkach (stanach) jest wydobywane i etykietowane (w rzeczywistości syntetyzowane są pojedyncze komplementarne łańcuchy DNA) dwoma różnymi fluorescencyjnymi barwnikami Oba ekstrakty są następnie rozprowadzane na powierzchni mikromacierzy; zabarwione produkty genowe łączą się z komplementarnymi sekwencjami w punktach
- Barwnik pozwala następnie na zmierzenie ilości dowiązanych próbek w poszczególnych punktach (poziom fluorescencji po wzbudzeniu laserem); przykładowo jeżeli RNA z próbki stanu normalnego jest liczniejsze wówczas punkt będzie zielony, jeżeli jest odwrotnie czerwony; jeżeli RNA z obu próbek jest równoliczne wówczas punkt będzie żółty a jeżeli żadne nie występuje - czarny

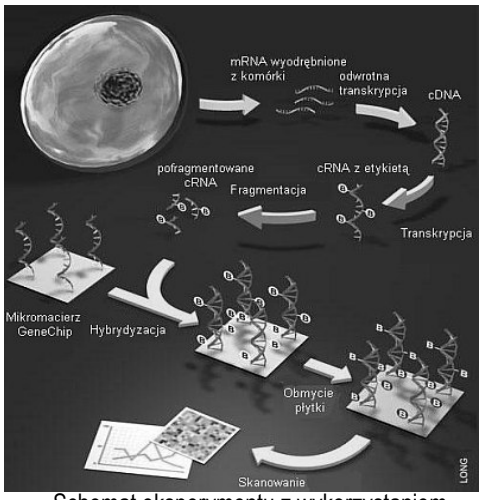


Bazując na poziomie fluorescencji i kolorze punktów można określić relatywny poziom ekspresji genów w obu próbkach

Macierze oligonukleotydowe



Fotolitografia *GeneChip* – kombinacja fotolitografii z chemia kombnatoryczną



Schemat eksperymentu z wykorzystaniem mikromacierzy *GeneChip*

Urządzenia mikromacierzowe firmy Affymetrix

Oligonucleotide microarray GeneChip®

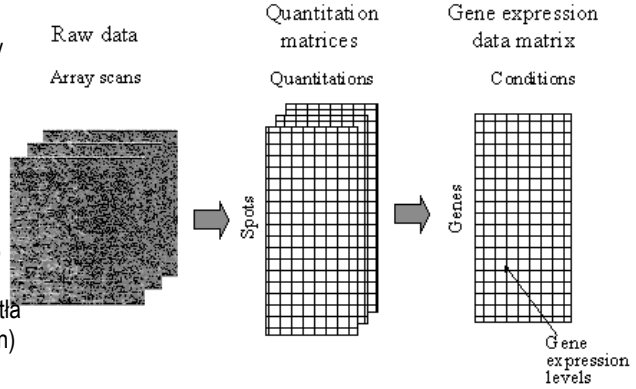
Hybridization oven

Fluid station

Image scanner and computer

Transformacja obrazów mikromacierzy w profile ekspresji genów

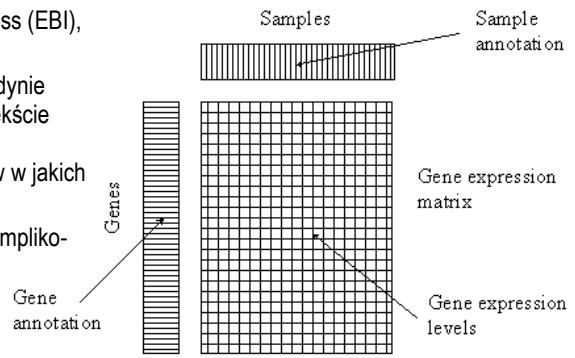
- Surowe dane w postaci obrazów mikromacierzy są analizowane przy użyciu wyspecjalizowanego oprogramowania, które pozwala precyzyjnie określić położenie punktów, zmierzyć jego intensywność i porównać z poziomem tła (ang. image quantitation)



- Aby uzyskać wynikową macierz ekspresji genów, wszystkie wartości związane z danym genem (na tej samej macierzy lub na serii macierzy odnoszących się do tych samych warunków w powtarzanych eksperymentach) są łączone (uśredniane) a następnie cała macierz jest skalowana aby umożliwić porównywanie różnych macierzy

Bazy danych uzyskanych z mikromacierzy

- Mikromacierze już w tej chwili produkują ogromne ilość informacji; dane te podobnie jak sekwencje genomu muszą być starannie gromadzone a następnie udostępniane
- Przykładowe bazy: ArrayExpress (EBI), GEO (NCBI)
- Dane ekspresji genów mają jedynie rzeczywiste znaczenie w kontekście konkretnej biologicznej próbki i dokładnie opisanych warunków w jakich próbkę pobrano
- Opis genów jest dodatkowo kompliko-wany przez wiele relacji wiele-do-wielu pomiędzy genami, brak standardów nazewnictwa
- Koncepcyjnie baza ekspresji genów może być traktowana jako składająca się z trzech części: macierzy ekspresji genów, opisu genów i opisu próbek



Analiza macierzy ekspresji genów

- Dwa podstawowe typy analiz:
  - porównywanie profili ekspresji poszczególnych genów poprzez porównywanie rzędów w macierzach ekspresji genów (np. przewidywanie klas funkcjonalnych genów)
  - porównywanie profili ekspresji próbek poprzez porównywanie kolumn w macierzy (np. przewidywanie typu tkanki, jej stanu, ...)
- Możliwe są również kombinacje powyższych metod przy założeniu, że przeprowadzona została odpowiednia normalizacja danych
- Rodzaje analiz:
  - nadzorowana - klasyfikacja, rozpoznawanie
  - nienadzorowana - analiza skupień (ang. cluster analysis), grupowanie

Sposoby pomiaru podobieństwa (odległości) pomiędzy profilami

- Każda z miar ma swoje zalety i wady; wybór może być związany z typem analizy; aktualnie nie ma teorii wyboru najlepszej miary
- Najprostszym przykładem określania podobieństwa jest wykorzystanie odległości euklidesowej (lub bardziej ogólnie odległości L^p)
- Korelacja pomiędzy wektorami ekspresji (współczynnik korelacji Pearsona - iloczyn skalarny znormalizowanych wektorów lub cosinus kąta pomiędzy nimi), wychwytuje dobrze podobieństwo kształtu, ale nie kładzie nacisku na wielkość dwóch serii pomiarów, jest czuła na wartości odstające (ang. outliers)
- Inne miary odległości: współczynnik korelacji rang (ang. rank correlation coefficient), oparte na mutual-information
- Przykładowo rozpatrzmy dwa niezwiązane geny, których ekspresja waha się blisko poziomu zerowego
  - duże podobieństwo - odległość euklidesową bliska 0
  - małe podobieństwo - korelacja bliska 0

Analiza nienadzorowana
(ang. unsupervised analysis)

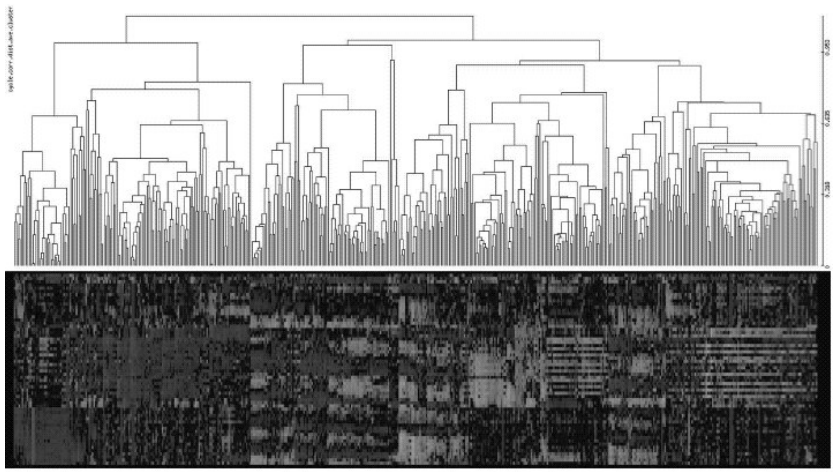
- Celem analizy skupień (ang. clustering) jest pogrupowanie obiektów (genów lub próbek), które charakteryzują się zbliżonym zachowaniem
  - grupowanie nie jest nową techniką, istnieje wiele algorytmów, które mogą być stosowane, przykładowo: hierarchiczne grupowanie, K-średnie, mapy Kohonena (ang. self-organizing maps)
  - może być w pewnym sensie widziane jako redukcja wymiarowości systemu
- Problem wyboru liczby grup jest delikatną sprawą
  - zależy m.in. od skali obserwacji danych
  - najczęściej stosowaną i efektywną techniką jest ciągle jeszcze, w dużej mierze manualna, metoda prób i błędów
  - ponieważ zwykle liczba grup (K) jest relatywnie mała, wszystkie dopuszczalne wartości K są sprawdzane
  - kluczową kwestią jest ocena poszczególnych podziałów, która opiera się na dążeniu do maksymalnego podobieństwa wewnątrz grup i maksymalnego odróżniania od pozostałych grup

Hierarchiczne grupowanie

Algorytm zachłanny tworzenia dendrogramu:

- Liście(N) są przypisywane do poszczególnych elementów (genów) i wyliczana jest macierz podobieństwa (odległości)
- W kolejnych N-1 krokach :
  - dwa najbardziej zbliżone elementy (grupy genów) są wyznaczane na podstawie m. podobieństw i tworzony jest węzeł je łączący;
  - wyliczany jest uśredniony profil w nowej grupie i modyfikowana macierz podobieństw
- Na wyjściu: drzewo binarne a nie zbiór grup, aby wyznaczyć grupy należy ustalić "poprzeczkę"
- Po utworzeniu dendrogramu pozostaje jeszcze kwestia wizualizacji
- Liście są wyświetlane w liniowym porządku i interpretacja biologiczna często jest powiązana z tym porządkiem
  - zakłada się, że kolejne geny są powiązane w pewien sposób
- W każdym węźle dwa elementy mogą być umieszczone w różnej kolejności
  - przy N-1 połączeń liczba liniowych porządków zgodnych ze strukturą drzewa wynosi 2^{N-1}
  - opracowano algorytm pozwalający znaleźć optymalne uporządkowanie przy wykorzystaniu programowania dynamicznego - $O(N^4)$
- Optymalne uporządkowanie liści pomaga w ustaleniu granic grup i związków pomiędzy grupami

Przykład hierarchicznego grupowania macierzy ekspresji genów



Grupowanie 505 genów drożdży w 3 różnych cyklach komórkowych (60 próbek); kolor czerwony - pozytywna wartość, zielona - negatywna, niebieska - brak

Algorytm K-średnich
(ang. K-means)

- Prosta i bardzo często stosowana metoda; zwykle ustalona liczba K grup
  - odpowiadająca np. spodziewanej liczbie wzorców regulatorów
- Początkowe położenia K centrów grup (prototypy, centroidy) inicjowane są zwykle losowo
- W kolejnych krokach:
  - każdy punkt jest przypisywany do grupy, do której centroidu ma najbliżej
  - po przypisaniu do grup, położenie centroidów jest przeliczane na podstawie punktów w grupach (np. poprzez uśrednianie lub środek ciężkości)
  - powyższe kroki są powtarzane, aż do uzyskania zbieżności lub akceptowalnych fluktuacji
- Wynikowe grupy zależą od wyboru startowego położenia prototypów:
  - zalecane jest wielokrotne powtórzenie algorytmu z różnymi startowymi położeniami
- Istnieje wiele odmian powyższego algorytmu (sposób startu, wyliczania odległości i centroidów, ścisła lub rozmyta przynależność do klas, ...)

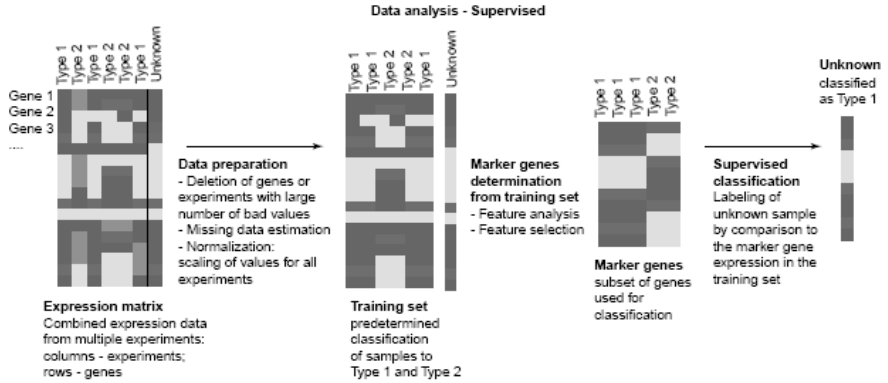
Analiza nadzorowana
(ang. supervised analysis)

- Celem jest tworzenie klasyfikatorów (np. liniowej funkcji dyskryminacyjnej, drzewa decyzyjnego czy *support vector machines*), które pozwalają przypisywać predefiniowane klasy do profili ekspresji genu lub próbek
- Przykładowo, klasyfikator umożliwiający rozpoznanie klasy funkcjonalnej genu drożdży na podstawie profilu ekspresji próbek
  - 6 klas, 79 próbek, kilkaset (?) genów
  - geny z niektórych klas (np. białka rybosomowe) rozpoznawane bez trudności, inne (np. kinaza) nie posiadały wyraźnego profilu ekspresji
  - w tym zastosowaniu najwyższą jakość klasyfikacji uzyskano przy użyciu SVM
- Możliwy jest odmienny scenariusz wykorzystania informacji zawartej w macierzy ekspresji genów (analiza kolumn) :
  - np. klasyfikator pozwalający odróżniać na bazie profilu ekspresji genów zbliżone morfologicznie nowotwory wykorzystywany może być w diagnostyce
  - przed wykorzystaniem w praktyce klinicznej niezbędna jest ocena jakości klasyfik.

Analiza nadzorowana (2)

  - ponadto, jeżeli klasyfikator jest oparty na relatywnie prostym zestawie reguł, wówczas może być wykorzystany do poznania mechanizmów związanych z poszczególnymi typami nowotworów
- W przypadku klasyfikacji próbek pojawić się może problem znacznie większej liczby atrybutów (genów) niż obiektów (próbek), na podstawie których tworzony jest klasyfikator - problem tzw. underfittingu
  - prowadzi to do sytuacji, w której bez trudu tworzony jest pozornie idealny klasyfikator (bezbłędny jedynie na zbiorze uczącym) o ile nie istnieją ograniczenia na jego złożoność
  - w celu uniknięcia problemu należy poszukiwać możliwie najprostszych klasyfikatorów, w których osiąga się kompromis pomiędzy prostotą i jakością klasyf.
- Zagadnieniem związanym bezpośrednio z powyższym problemem jest selekcja najbardziej znaczących cech (wybór istotnych genów)
  - najczęściej stosowane metody krokowe (ang. step-wise), ale również np. algorytmy ewolucyjne

Schemat procesu klasyfikacji



Rola selekcji cech w klasyfikacji - przykład

102 próbka (50 zdrowych, 52 rak)
po wstępnych przetwarzaniu
i filtracji uzyskano 6034 genów

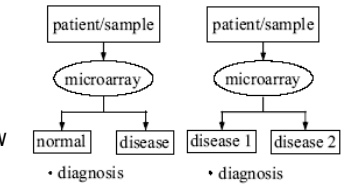
podzbiór38 różnicujących genów

Sposoby medycznego wykorzystania mikromacierzy

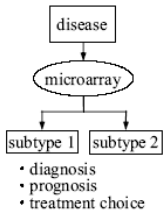
- Czy istnieją jakieś specyficzne wzorce profili ekspresji związane z konkretnym schorzeniem? (*diagnoza molekularna*)

lub

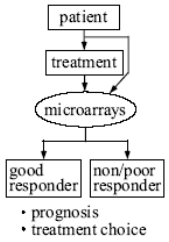
- Czy istnieją rozpoznawalne różnice w ekspresji genów pomiędzy schorzeniami? (*klasyfikacja schorzeń*)



- Czy można wyróżnić podkategorie choroby na podstawie profili ekspresji? (*odkrywanie podtypów*)



- Ekspresja których genów ulega zmianie pod wpływem konkretnego leczenia? (*odpowiedź na leczenie*)



Identyfikacja potencjalnych sygnałów regulacyjnych

- Istnieje duża zgodność co do hipotezy, że geny posiadające zbliżone profile ekspresji mają zapewne cechy wspólne mechanizmu regulacji:

  co-expressed genes ==> co-regulated genes

- Grupując geny z podobnymi profilami ekspresji możliwe jest odnalezienie potencjalnie analogicznie regulowanych genów oraz poszukiwanie domniemanych środków regulacji (ang. regulatory signals)
- Podstawowe kroki metody:
  - grupowanie genów w oparciu o wybór pomiarów ekspresji,
  - wydobycie domniemanych sekwencji promotorów genów z danej grupy,
  - poszukiwanie wzorców sekwencji, którą są nadreprezentowane w sekwencjach,
  - ocena jakości odkrytych wzorców przy użyciu kryteriów statystycznego znaczenia.

Przykład: identyfikcja regulacji genów u drożdzy (1)

- Zbiór zawierający ekspresję 6221 genów w 80 różnych warunkach
- Krok 1: grupowanie profili ekspresji
  - metoda K-średnich z odległością euklidesową, K od 2 do 1000, 10 różnych konfiguracji startowych
  - w rezultacie z 900 grupowań wybrano 52100 grupy (rozmiar od 20 do 100 genów)
- Krok 2: poszukiwanie sekwencji wzorców
  - w każdej grupie sekwencje 600 baz poprzedzające geny zostały przeanalizowane
  - wzorce (o dowolnej długości) występujące w co najmniej 10 sekwencjach z grupy podlegają ocenie (w oparciu o p-wo ich pojawienia się)
- Krok 3: ustalenie progu istotności na podstawie eksperymentu
  - w celu ustalenia istotności statystycznej wzorców krok 2 powtórzono na danych z losowo wybranych grup
  - próg 10^{-8} pozwala uniknąć wyboru wzorców występujących również w sytuacji losowej

Przykład: identyfikcja regulacji genów u drożdzy (2)

- Krok 4: wybór wzorców
  - z ponad 6000 wzorców wiele zaobserwowano w grupach genów homologicznych, których grupy łatwo zidentyfikowano; usunięto te grupy (łącznie 169 genów)
  - pozostałe grupy z niehomologicznymi sekwencjami poprzedzającymi zawierały 3727 ORF i razem przedstawiały 1498 wzorców
- Krok 5: grupowanie wzorców
  - ponad 1000 wzorców jest zbyt dużą liczbą do analizy przez człowieka
  - wykorzystano miarę podobieństwa bazująca na *common information content*
  - uzyskano 62 grupy podobnych wzorców i dla każdej grupy wyznaczono przybliżone dopasowanie i reprezentatywne wzorce (ang. consensus patterns)
- Krok 6: skonfrontowanie odkrytych wzorców wobec znanych miejsc dowiązań podczas transkrypcji zapisanych w SCPD
  - 48 wzorców zostało dopasowanych, pozostałych 14 nie (potencjalnie najbardziej interesujące jako cel przyszłych badań)