0. Programy pakietu LINKAGE/FASTLINK i ich zastosowanie

Programy główne/analityczne

- **mlink** – analiza dwupunktowa, podaje wartość LOD score dla kolejnych wartości θ. Od niego zwykle zaczyna się analizę, by stwierdzić, czy dany marker może w ogóle być sprzężony z locus, może też służyć do szacowania ryzyka
- **linkmap** – analiza wielopunktowa, podaje najbardziej prawdopodobne położenie badanego *locus* względem ustalonej mapy markerów
- **ilink** – analiza dwu- i wielopunktowa, służy do podania najbardziej prawdopodobnej wartości θ wychodząc od przybliżonej wartości początkowej (maksimum lokalne), w analizie wielopunktowej podaje też najbardziej prawdopodobna kolejność *loci*. Jako wartość poczatkową dobrze podać wielkość oszacowaną przez MLINK/LINKMAP
- **lodscore** – program podobny do ILINK, ograniczony do analizy dwupunktowej
- **unknown** – program pomocniczy – oblicza możliwe genotypy w rodowodzie, zmniejszając liczbę kombinacji do przeanalizowania przez programy główne

Programy pomocnicze:

- **makeped** – tworzy plik rodowodu w formacie programów głównych
- **preplink** – tworzy plik opisu *loci* dla programów głównych
- **lcp** – tworzy skrypt przeprowadzający wybraną analizę. Służy do wybrania konkrenych *loci* do analizy dwupunktowej, ustalenia parametrów analizy wielopunktowej itp.
- **lrp** – przekształca wyniki analizy zaprogramowanej przez **lcp** w postać czytelnej tabeli

Typowy przebieg analizy to:

1. Dla szeregu markerów oddalonych o 8-20 cM zbadać możliwość sprzężenia z badanym *locus* (MLINK) – wstępne przeszukanie
2. Po zgrubnym zidentyfikowaniu obszaru sprzężenia odnaleźć dokładniejszy obszar, korzystając z markerów rozmieszczonych co 1- 4 cM. (MLINK)
3. Ustalić położenie badanego *locus* względem najbliższych markerów, których mapa jest znana (LINKMAP)
4. Dokładnie zmapować *locus* względem markerów, wykorzystując oszacowane wartosci z poprzedniego etapu (ILINK).

Markery do etapu 1 mogą pochodzić np. z CHLC (www.chlc.org), do etapu drugiego z CHLS, MIT lub Généthon.

1. Kodowanie cech

W programach pakietu LINKAGE/FASTLINK allele badanych *loci* koduje się numerycznie. Istnieją następujące typy *loci*

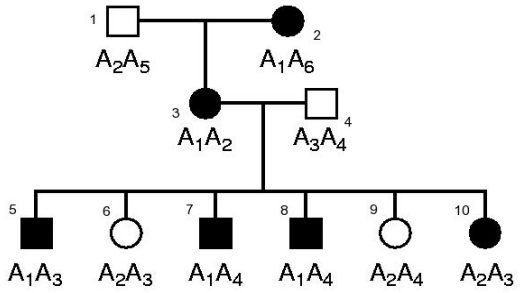
- *locus* choroby (*affection status*), służy do kodowania chorych i zdrowych członków rodziny. Uwaga: w przypadku cech recesywnych zdrowy może też być nosicielem. 1 oznacza zdrowego, 2 oznacza chorego, 0 oznacza brak danych. Możliwe jest zakodowanie niepełnej penetracji, sprzężenia z płcią, różnych klas ryzyka itd. (patrz opis programu **preplink**).
- *loci* numerowanych alleli (markerowe) (*numbered alleles*), służy do kodowania kodominujących markerów, np. RFLP (nadaje się też do kodowania niektórych układów grup krwi). Kolejne allele oznaczane są liczbami od 1 do *n*, 0 zawsze oznacza brak danych. Dla każdego osobnika w rodowodzie podaje się zawsze dwie liczby odpowiadające obu allelom, w przypadku cech sprzeżonych z płcią u mężczyzn drugą wartością może być 0 (lub powtórzony ten sam numer allelu)
- cechy ilościowe (*quantitative traits*), do kodowania cech ciągłych opisywanych liczbą rzeczywistą, np. poziomu enzymu, wzrostu itp.
- czynniki binarne (*binary factors*), każdemu allelowi w genotypie przypisuje się 1 (występuje) lub 0 nie występuje, potrzeba zatem tyle pozycji, ile alleli występuje w *locus*. Stosowany jako alternatywny zapis *loci* markerowych lub do opisywania złożonych systemów takich, jak układ grup krwi AB0, obecnie rzadziej stosowany.

Znajomość typów analizowanych *loci* potrzebna jest do prawidłowego zakodowania rodowodu.

2. Kodowanie rodowodu

Rodowody koduje się w plikach tekstowych, w zapisie uproszczonym (format .ped, niekiedy oznaczany też rozszerzeniem .pre), który następnie można przekształcić na format ostateczny programem **makeped**. W jednym pliku tekstowym można zapisać dowolną ilość rodowodów dla tych samych *loci* (LOD score są addytywne!). W kolejnych wierszach pliku koduje się kolejne osoby. Dla uproszczenia warto na rodowodzie kolejno ponumerować osoby.

Format pliku .pre objaśnia prosty przykład poniżej:



Plik .pre dla tego rodowodu wygląda następująco

1 2 3 4 5 6 7 8 numery kolumn **nie występują w pliku**

```
001   1  0  0  1  1  2  5
001   2  0  0  2  2  1  6
001   3  1  2  2  2  1  2
001   4  0  0  1  1  3  4
001   5  4  3  1  2  1  3
001   6  4  3  2  1  2  3
001   7  4  3  1  2  1  4
001   8  4  3  1  2  1  4
001   9  4  3  2  1  2  4
001  10  4  3  2  2  2  3
```

Kolejne kolumny oznaczają

1 – identyfikator rodziny, w tym pliku jest tylko jedna rodzina, ale identyfikator musi wystepować

2 – identyfikator kolejnej osoby

3 – identyfikator ojca, 0 oznacza, że nie występuje w tym rodowodzie

4 – identyfikator matki, 0 oznacza, że nie występuje w tym rodowodzie

5 – płeć: 1 – mężczyzna, 2 – kobieta

Dalsze kolumny kodują allele kolejnych *loci*, w tym rodowodzie mamy :

jeden *locus* choroby (kolumna 6, 1 – zdrowy, 2 – chory)

jeden locus markerowy (kolumny 7 i 8 dla obu alleli u każdego osobnika)

W pliku rodowodu można zapisać dowolną ilość *loci*, na etapie przygotowania analizy można wybrać dowolne dwa dla analizy dwupunkowej. Wygodnie jest zapisywać *locus* choroby jako pierwszy.

Plik .ped należy teraz przekształcić na format czytany przez pozostałe programy. W tym celu wykorzystujemy program **makeped** wydając następującą komendę

`makeped pedfile.ped pedfile.dat n`

gdzie *pedfile.ped* to przygotowany przez nad plik tekstowy a *pedfile.dat* to nazwa pliku wynikowego. Opcja **n** oznacza, że w rodowodzie nie występują pętle (pętle pojawiają się np. przy

kojarzeniu krewniaczym), jeżeli mamy pętle to wpisujemy tu **y**.

Istnieje też program pozwalający na graficzną edycję i modyfikację rodowodów zapisanych w plikach .ped – patrz Dodatek II.

3. Przygotowanie opisu *loci* – program preplink

Kolejny plik tekstowy opisuje *loci* i sposób analizy. Jego format jest dosyć zagmatwany, ale istnieje program **preplink** pozwalający na jego przygotowanie. W programie **preplink** poszczególne opcje zmienia się wybierając pozycje z menu tekstowego.

Ekran startowy preplink wygląda następujaco:

```
********** PRESENT STATUS **************
(a) Number of loci               :    2
(b) Sexlinked                    :    N
(c) Calculate Risk               :    N
(d) Mutation                     :    N
(e) Haplotype frequencies        :    N
(f) Locus Order                  :    1 2
(g) Interference                 :    N
(h) Recombination sex difference :    N
(i) Program used                 : MLINK
(j) Recombination values         :
       0.100
*********** OTHER OPTIONS **************
(k) See or modify loci description
(l) See or modify recombination to vary
(m) Read datafile
(n) Write datafile
(o) Exit
*************************************
Press letter to modify or see values
```

Najważniejsze opcje to liczba *loci* (a), sprzężenie z płcią (b) i typy *loci* (k).

Początkową wartość θ (j), krok jej zmiany (l), kolejność *loci* (f) i program używany do analizy (i) będziemy ustalać później przy pomocy programu **lcp,** więc na razie ich modyfikacja jest zbędna.

Jak widać program pozwala na uwzględnienie w analizie wielu zaawansowanych możliwości (mutacje *de novo* z zadaną częstością, różne częstości rekombinacji zależnie od płci, interferencja itp. Na potrzeby większości analiz należy pozostawić te opcje niezmienione.

W naszym przykładzie mamy *locus* choroby i *locus* markera A z 6 allelami. Po naciśnięciu *k* pojawia się następujący ekran

```
****************************************
(1) allele numbers   GENE FREQS :  0.50000  0.50000
(2) allele numbers   GENE FREQS :  0.50000  0.50000
****************************************
(a) SEE OR MODIFY A LOCUS
(b) DELETE LOCUS
(c) ADD LOCUS
(d) CHANGE ORDER TO CORRESPOND TO PEDIGREE FILE (NOT CHROMOSOME ORDER)
(e) CHANGE LOCUS TYPE
(f) RETURN TO MAIN MENU
****************************************
Press letter to modify values
```

Domyślnym typem jest locus numerowanych alleli z dwoma allelami. Najpierw więc musimy zmienić typ pierwszego locus na *locus* choroby. Po wciśnięciu *e* i 1 (zmieniamy *locus* 1) pojawiają się 4 opcje odpowiadające czterem typom *loci*). Wybieramy *locus* choroby (*affection status*), czyli *c*.

Teraz musimy zmienić parametry obu naszych *loci*. Najpierw *locus* choroby (1), wybieramy *a* i 1. Uzyskujemy następujący widok opcji:

```
******************************

LOCUS NUMBER :            1
******************************
(a) NUMBER OF ALLELES           :    2
(b) NUMBER OF LIABILITY CLASSES :    1
(c) PENETRANCES :
GENOTYPE  1 1  0.00000000
GENOTYPE  1 2  0.00000000
GENOTYPE  2 2  1.00000000
(d) GENE FREQUENCIES :
  0.50000  0.50000
(e) EXIT
******************************
Press letter to modify values
```

Najważniejsza jest opcja (c) pozwalająca na ustalenie typu dziedziczenia. Domyślne wartości odpowiadają chorobie recesywnej (0% penetracji dla heterozygoty), musimy zmienić te wartości by zakodować dziedziczenie dominujące. Po wciśnięciu *c* pojawi się pytanie o penetrację dla genotypu 1 1 (wpisujemy 0), genotypu 1 2 (wpisujemy 1 – choroba dominująca z pełną penetracją) i dla genotypu 2 2 (wpisujemy 1). Dla cech sprzeżonych z płcią program zapyta o penetracje dla genotypów mężczyzn i kobiet osobno.

Musimy jeszcze wprowadzić przybliżone częstości alleli, Allel zmutowany jest rzadki, więc po wybraniu *d* możemy wpisać 0.99 0.01. Znajomość dokładnej częstości alleli jest potrzebna przy bardziej złożonych analizach, nie wpływa znacząco na wyniki w prostych przykładach.

Po powrocie do poprzedniego menu (e), w podobny sposób zmieniamy liczbę alleli dla drugiego *locus*, po zmianie liczby alleli musimy wprowadzić nowe częstości alleli.

Uwaga: ważna jest kolejność wprowadzania zmian: najpierw ustala się typ *locus*, następnie liczbę alleli, a na końcu penetracje i/lub częstości alleli. Zmiana typu *locus* powoduje powrót do domyślnych ustawień tych parametrów.

Po ustaleniu parametrów *loci* zapisujemy plik opcją (n) menu głównego. Moża wczytać zapisany wcześniej plik (opcja (m)) w celu jego modyfikacji. Uwaga: program bez ostrzeżenia nadpisuje istniejące pliki o tej samej nazwie, co podana przy zapisie. Przyjęto nadawać plikom z parametrami rozszerzenie .dat. Trzeba uważać, żeby nie nadpisać pliku z rodowodem, który też ma rozszerzenie .dat.

Uwaga: programy LINKAGE posługują się plikami o domyślnych nazwach, które mogą zostać nadpisane przez programy. Należy w związku z tym unikać nadawania własnym plikom następujących nazw: datafile.dat, final.dat, ipedfile.dat, lsp.log, lsp.stm, lsp.tmp, outfile.dat, pedfile.dat, recfile.dat, speedfile.dat, stream.dat, tempdat.dat, tempped.dat. Należy też unikać zbyt długich nazw zawierających znaki specjalne.

4. Przygotowanie i wykonanie analizy – program lcp

W przypadku bardzo prostej analizy (tylko dwa *loci* opisane w rodowodzie, prosta analiza mlink) można bezpośrednio wywołac program mlink. Plik z rodowodem musi nazywać się pedfile.dat, a plik z parametrami *loci* datafile.dat. Praktycznie wszystkie poważniejsze analizy nalezy jednak przygotować posługując się programem **lcp.** Program lcp wykorzystuje plik z rodowodem (przygotowany przez **makeped**) i opisem *loci* (przygotowany przez **preplink**) i przygotowuje plik komend (skrypt shell w systemie Unix albo plik .bat w DOS), który wywoła właściwe programy. W tym przypadku musimy pamiętac, by nie nadawać naszym plikom danych zastrzeżonych nazw pedfile.dat i datafile.dat, gdyż **lcp** stworzy je automatycznie. Program **lcp** pozwala na wybór dowolnych dwóch *loci* do analizy dwupunktowej ze zbiorów, w których zakodowano więcej *loci* oraz na wybór typu analizy i użytego programu. Po wpisaniu **lcp** pojawia się pierwszy ekran, na którym podajemy nazwy wykorzystywanych zbiorów z danymi i nazwy zbiorów wyjściowych. Wygląda on następująco:

```
          COMMAND file name [pedin] : pedin
          LOG file name [final.out] : final.out
     STREAM file name [stream.out] : stream.out
   PEDIGREE file name [pedin.dat] : pedin.dat
PARAMETER file name [datain.dat] : datain.dat
 Secondary PEDIGREE file name [] :
Secondary PARAMETER file name [] :
```

Pomiędzy liniami przemieszczamy się klawiszami kursora, działa też klawisz backspace. W pierwszej linii podajemy nazwę, jaką ma nosić stworzony przez **lcp** skrypt do analizy. Dwie kolejne linie definiują zbiory wynikowe. Możemy pozostawić wartości domyślne, nalezy tylko pamiętać, że kolejne analizy w tym samym katalogu nadpiszą te zbiory. W kolejnych polach podajemy nazwę pliku z rodowodem (stworzonego przez **makeped**) i pliku z opisem *loci* (przygotowanego przez **preplink**). Po wypełnieniu wszystkich wartości przechodzimy do następnego ekranu wciskając kombinację klawiszy Cntrl-N. Kolejny ekran wygląda następujaco:

```
                  General pedigrees : <-
            Three-generation pedigrees :
          Experimental cross pedigrees :
```

Praktycznie wszystkie omawiane tu analizy dotyczą ogólnych rodowodów, wybieramy więc pierwszą opcję wciskając ENTER. Kolejny ekran pozwala na wybór programu wykorzystanego do przeprowadzenia analizy:

```
                               LODSCORE :
                                  ILINK :
                                 LINKMAP :
                                   MLINK : <-
```

Wybieramy MLINK, następny ekran służy do wyboru typu analizy:

```
                    Specific evaluation :
                        Lod score table : <-
             Multiple pairwise Lod table :
```

Pierwsze dwie opcje dają w sumie podobne możliwości: w pierwszej ustalamy początkową wartość θ, czy ma sie zmieniać, a jeżeli tak, to z jakim krokiem. W drugiej możemy podać wprost ciąg wartosci θ, dla których program ma policzyć Lod score. Ostatnia opcja służy do wyliczenia dwóch tabel dla dwóch par *loci*.

Po wyborze którejkolwiek opcji musimy potwierdzić, że nie uwzględniamy różnicy w częstości rekombinacji między płciami (w obecnej wersji program **mlink** na to nie pozwala, różnicę tę można uwzględnić w programie **ilink**) i uzyskujemy następujacy ekran:

```
                       Locus order [] :
          Recombination fractions [.0] : .0
              Recombination varied [1] : 1
    Other recomb. [.01 .05 .1 .2 .3 .4] : .01 .05 .1 .2 .3 .4
```

W pierwszym polu podajemy, które *loci* wybieramy do analizy (mlink analizuje *loci* parami), w drugim poczatkową wartość θ, w trzecim 1 oznacza, ze ta wartość ma się zmieniać, ostatnie pole podaje kolejne wartości θ, dla których program obliczy Lod score. W pierwszym polu wpisujemy `1 2`, gdyż w zbiorze przykładowym mamy tylko dwa *loci*. Gdy mamy zdefiniowanych więcej *loci*, tu wybieramy, którą parę bierzemy do analizy. Następnie naciskamy Cntrl-N. W tym momencie program zapisuje plik skryptu, nie wyświetla jednak żadnego komunikatu. Uwaga: nie należy w tym momencie naciskać Cntrl-N więcej, niż raz, gdyż za każdym razem do skryptu dopisana zostanie kolejna analiza. Wychodzimy z programu naciskając Cntrl-Z. W katalogu powinien pojawić się plik o nazwie, którą wybraliśmy na pierwszym ekranie w pierwszym polu. Jest to skrypt wykonywalny, który uruchamiamy wpisując

sh *nazwa*

Skrypt najpierw uruchamia program **unknown**, który analizuje rodowód i ustala możliwe genotypy wszystkich osób (nie jest to konieczne, ale przyspiesza dalsze obliczenia). Następnie przygotowuje odpowiednie pliki wejściowe (korzystając z programu **lsp**) i wywołuje właściwe programy. Wyniki należy teraz przekształcić na czytelny format.

5. Formatowanie wyników – program lrp

Wyniki analizy zapisane są w pliku tekstowym o nazwie, którą podaliśmy w trzeciej linijce pierwszego ekranu **lcp** (domyślnie stream.out). Program **lrp** przekształca ten plik na czytelną postać. Obsługa podobna jest do programu **lcp**. Po uruchomieniu **lrp** pierwszy ekran pozwala na podanie nazwy obrabianego pliku wynikowego oraz, opcjonalnie, tytułu raportu. Do następnego ekranu przechodzimy przez Cntrl-N, tam wybieramy opcję “General pedigree reports”, w następnym ekranie wybieramy nazwę programu, którego użylismy do analizy (w tym przykładzie MLINK). Na następnym ekranie wybieramy “Table format” (druga opcja wyświetli cały plik wejściowy). Na kolejnym ekranie wybieramy elementy, które mają znaleźć się w raporcie. Domyślnie jest to tabela Lod score, można jeszcze dodatkowo wybrać wyniki w postaci logarytmu naturalnego (Log e) oraz zdecydować, czy oprócz wyników sumarycznych przedstawiać też wyniki dla pojedynczych rodzin (jeżeli w pliku z rodowodami zakodowaliśmy więcej niż jedną rodzinę – opcja “Include pedigrees”). Cntrl-N przenosi do ostatniego ekranu, gdzie decydujemy, zcy wynik ma być wyświetlony na ekranie, czy zapisany do pliku tekstowego (domyślna nazwa report.txt, może być zmieniona). Ostateczny wynik wygląda następująco:

```
                          L O D   T A B L E   R E P O R T

File: stream.out                                                    Screen:   1 of 1
_____________________________________________________________________________________

Order      0.0    0.01    0.05     0.1    0.15     0.2     0.3     0.4    0.45
-----  ------- ------- ------- ------- ------- ------- ------- ------- -------
1=2    -infini   -0.22    0.39    0.58    0.63    0.62    0.51    0.30    0.16  a


= = Test Interval
a = LOD Scores                  (LOG 10)
b = LOG 10 Likelihoods          (LOG 10)
```

Dodatek I – instalacja programów

Główne programy FASTLINK (MLINK, ILINK, LINKMAP, LODSCORE, UNKNOWN) dostępne są w postaci kodu źródłowego w języku C ze strony:

http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html

Po rozpakowaniu archiwum przechodzimy do katalogu `fastlink/4.1P/src` i komendą `make` kompilujemy źródła.

Przed kompilacją można zmienić domyślne parametry (ograniczenia) programów, np. maksymalną liczbę analizowanych *loci* (domyślnie 8), maksymalną liczbę alleli w *locus* (domyślnie 25), maksymalną liczbę osobników w analizie (domyślnie 1000) itp. W tym celu edytujemy plik tekstowy `commondefs.h`

Najważniejsze parametry to:

maxlocus – maks. liczba *loci*

maxall - maks. liczba alleli w *locus*

maxped - maks. liczba rodowodów w jednej analizie

maxind – maks. liczba osobników

maxchild - maks. liczba rodzeństwa jednej pary

Edytując plik `Makefile` można dobrać parametry kompilacji (np. flagi optymalizacji kompilatora, można też skompilować programy do wykonywania równoległego na maszynach wieloprocesorowych).

Programy pomocnicze (makeped, lcp, lsp, lrp) dostępne są pod adresem:

ftp://fastlink.nih.gov/pub/staff/schaffer/linux-aux

w postaci kodu źródłowego w języku C oraz gotowych plików binarnych dla systemu Linux i386.

Oryginalny kod programów LINKAGE (w języku Pascal) i dużo przydatnych informacji można znaleźć na stronie ich autora (http://www.jurgott.org/linkage/LinkagePC.html).

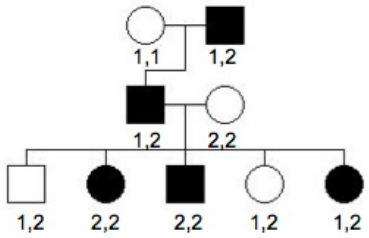
Dodatek II – program pelican – graficzny edytor rodowodów

Program pelican (napisany w języku Java i działający na wszystkch platformach, na których działą Java) dostępny jest ze strony: http://www.mrc-bsu.cam.ac.uk/personal/frank/software/pelican/

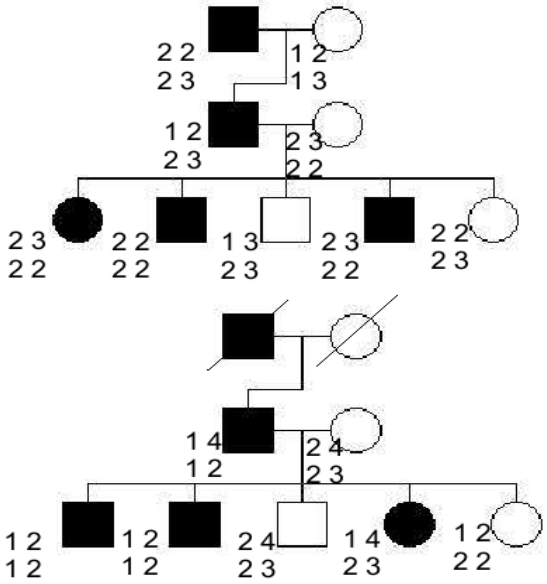
Program pozwala na rysowanie rodowodów i zapisywanie ich w pliku w formacie .ped. Można zakodować więzy pokrewieństwa i chorobę, genotypy markerów trzeba dopisać do utworzonego pliku tekstowego. Program jest intuicyjny w obsłudze – ogólnie: lewym klawiszem myszy selekcjonujemy osobnika, prawy klawisz myszy otwiera menu opcji, z którego możemy dodać partnera i/lub potomków, zmienić stan choroby, płeć itd.

Ćwiczenia

1. Przeprowadzić analizę MLINK dla przykładu analizowanego na poprzednich zajęciach i porównać wyniki. Podobnie, jak poprzednio powtórzyć analizę przy założeniu braku znajomości genotypów dziadków (1,2). Rodowód załączono poniżej:



2. *(opcjonalne) Poniżej przedstawiono dwa rodowody, w których analizowane jest sprzężenie pewnej choroby dominującej autosomalnej z dwoma markerami RFLP.*
   - *W analizach dwupunktowych (MLINK) oszacuj, czy dwa użyte w badaniu markery są sprzężone z genem choroby*
   - *Ustal przybliżoną mapę obszaru w analizie trzypunktowej (kolejność genów, przybliżone odległości – w* ***lcp*** *wybierz program LINKMAP, opcja "All intervals", locus choroby jako locus testowany, loci markerów jako ustalone (kolejność 2 3, w tym przypadku nieistotna)*
   - *Na podstawie wyników analizy LINKMAP ustal dokładną mapę – użyj programu ILINK, wybierz opcję "Specific order" i wprowadź kolejność i przybliżone odległości na podstawie wyników ana*lizy LINKMAP



4. Skopiuj (nie przenieś!) dwa pliki z katalogu IMD_cwiczenia do nowego katalogu. Obejrzyj rodowód (FamD.pre) za pomocą edytora pelican (pamiętaj o genotypach markerów), a plik z parametrami analizy (FamD.dat) za pomocą preplink. Zmień nazwę pliku FamD.pre na taką z rozszerzeniem .ped, która nie będzie powtarzać się z nazwą FamD.dat! Przedstawiają one wyniki badań (z Instytutu Matki i Dziecka w Warszawie) nad dziedziczną niepełnosprawnością intelektualną. Na którym chromosomie jest *locus* choroby, z iloma markerami badano sprzężenie? Jak dziedziczy się cecha? Przekształć rodowód na format .dat i za pomocą lcp zaprogramuj analizę MLINK *locus* choroby kolejno z każdym z markerów (za pomocą parametru „locus order", najpierw „1 2", potem „1 3", itd.). W pobliżu którego markera należy szukać genu z mutacją sprawczą?