Aufgabenblatt 5

Silke Trißl

Wissensmanagement in der Bioinformatik

Zuerst!

FRAGEN ?

Exercise 1 + 2

- Modify program to
  - compare protein sequence
  - read substitution matrix from a file
- Compare results quantitatively
  - identity matrix (InsDelRep = -1, Match = +1)
  - different BLOSUM Matrix
- BUT: let‘s repreat first

Substitutionsmatrix

- Matrizen für Proteine
  - PAM = **P**oint **a**ccepted **m**utations
  - BLOSUM = **Blo**cks **su**bstitution **m**atrix

- Stellen für jedes Paar von Aminosäuren einen Substitutionswert zur Verfügung

- PAM und BLOSUM beruhen auf beobachteten Übergängen von Aminosäuren

PAM

- PAM: Zwei Bedeutungen
  - 1 PAM – Einheit für den Abstand von Proteinsequenzen
  - PAM-X Matrix – Berechnete Substitutionsmatrix für zwei Sequenzen, die X PAM entfernt sind

- Definition
  Seien S1 und S2 zwei Proteinsequenzen mit |S1|=|S2|. S1 und S2 heißen x PAM entfernt, wenn S1 in S2 überführt wurde mit x Punktmutationen pro 100 Aminosäuren

Berechnung von PAM

- Gegeben:
  - n Alignments von je 2 Proteinsequenzen
  - sehr ähnliche Proteinsequenzen (> 85 %)

```
ATGCCGTA
ATG_CCTA

CGCCACGT
_GCCTCGT

TCGCAGTA
TCG_GG_A
```

Berechnung von PAM

- Ziel:
  - Zahlenwert für ein Paar von Aminosäuren
    - 0, wenn beobachteter = erwarteter Übergang
    - > 0, wenn beobachteter > erwarteter Übergang
    - < 0, wenn beobachteter < erwarteter Übergang

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-2	-3	-1	-1	-2	-1	-3	-3	-3	-2	-2	-4	-1	1	-1	-4	-4	-1	-3	-2	-1	-7
R	-2	7	-1	-3	-5	0	-2	-4	-1	-4	-4	2	-2	-4	-3	-2	-2	-4	-3	-4	-2	-1	-2	-7
N	-2	-1	7	1	-4	-1	-1	-2	0	-5	-5	-1	-4	-5	-4	0	-1	-6	-3	-4	4	-1	-2	-7
D	-3	-3	1	7	-5	-2	1	-3	-2	-6	-6	-2	-5	-5	-3	-1	-2	-7	-5	-5	4	0	-3	-7
C	-1	-5	-4	-5	9	-5	-6	-5	-5	-2	-3	-5	-3	-3	-5	-2	-2	-5	-4	-2	-5	-6	-3	-7
Q	-1	0	-1	-2	-5	7	1	-3	0	-4	-3	1	-1	-4	-2	-1	-2	-3	-3	-3	-1	3	-2	-7
E	-2	-2	-1	1	-6	1	6	-4	-1	-5	-5	0	-4	-5	-3	-1	-2	-5	-4	-3	0	5	-2	-7
G	-1	-4	-2	-3	-5	-3	-4	6	-4	-6	-5	-3	-5	-5	-4	-1	-3	-5	-6	-5	-2	-4	-3	-7
H	-3	-1	0	-2	-5	0	-1	-4	9	-5	-4	-2	-3	-2	-3	-2	-3	-3	1	-5	-1	-1	-2	-7
I	-3	-4	-5	-6	-2	-4	-5	-6	-5	5	1	-4	1	-1	-4	-4	-2	-4	-3	2	-5	-4	-2	-7
L	-3	-4	-5	-6	-3	-3	-5	-5	-4	1	5	-4	2	0	-4	-4	-3	-4	-3	0	-5	-4	-2	-7
K	-2	2	-1	-2	-5	1	0	-3	-2	-4	-4	6	-2	-4	-2	-1	-2	-5	-4	-4	-1	0	-2	-7
M	-2	-2	-4	-5	-3	-1	-4	-5	-3	1	2	-2	8	-1	-4	-3	-2	-3	-3	0	-4	-3	-2	-7
F	-4	-4	-5	-5	-3	-4	-5	-5	-2	-1	0	-4	-1	7	-5	-3	-3	0	3	-2	-5	-5	-3	-7
P	-1	-3	-4	-3	-5	-2	-3	-4	-3	-4	-4	-2	-4	-5	8	-2	-3	-6	-5	-4	-3	-3	-3	-7
S	1	-2	0	-1	-2	-1	-1	-1	-2	-4	-4	-1	-3	-3	-2	6	1	-4	-3	-3	-1	-1	-1	-7
T	-1	-2	-1	-2	-2	-2	-2	-3	-3	-2	-3	-2	-2	-3	-3	1	6	-5	-3	-1	-2	-2	-1	-7
W	-4	-4	-6	-7	-5	-3	-5	-5	-3	-4	-4	-5	-3	0	-6	-4	-5	11	1	-4	-6	-4	-4	-7
Y	-4	-3	-3	-5	-4	-3	-4	-6	1	-3	-3	-4	-3	3	-5	-3	-3	1	8	-3	-4	-4	-3	-7
V	-1	-4	-4	-5	-2	-3	-3	-5	-5	2	0	-4	0	-2	-4	-3	-1	-4	-3	5	-5	-3	-2	-7
B	-3	-2	4	4	-5	-1	0	-2	-1	-5	-5	-1	-4	-5	-3	-1	-2	-6	-4	-5	4	0	-2	-7
Z	-2	-1	-1	0	-6	3	5	-4	-1	-4	-4	0	-3	-5	-3	-1	-2	-4	-4	-3	0	4	-2	-7
X	-1	-2	-2	-3	-3	-2	-2	-3	-2	-2	-2	-2	-2	-3	-3	-1	-1	-4	-3	-2	-2	-2	-2	-7
*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	1

PAM Matrizen

- Definition
  Seien ($S_{1,1}$, $S_{2,1}$), ..., ($S_{1,n}$, $S_{2,n}$) Paare von Sequenzen, die jeweils x PAM entfernt sind. Dann ist die PAM-x Matrix M_x wie folgt definiert:
  - Sei f(i) die relative Häufigkeit der Aminosäure A_i
  - Seien alle Paare optimal aligniert
  - Sei f(i,j) die relative Übergangshäufigkeit von A_i zu A_j in allen alignierten Paaren
    - Übergang ist „richtungslos“: f(i,j) = f(j,i)
    - Paare (A_x, _) werden ignoriert
    - f(i,j) wird auf die Gesamtzahl aller Paare ohne INSDEL normiert
  - Damit:

$$M_x(i,j) = \log\left(\frac{f(i,j)}{f(i) * f(j)}\right)$$

Erläuterung

- Log-Odds Ratio
- Bruch
  - Verhältnis der beobachteten Ersetzung zu den erwarteten Übersetzungen bei statistischer Unabhängigkeit

$$M_x(i,j) = \log\left(\frac{f(i,j)}{f(i) * f(j)}\right)$$

  - M(i,j) = 0 (Bruch = 1)
    - Keine Selektion - Anzahl Übergänge entspricht statistischer Erwartung
  - M(i,j) < 0 (Bruch < 1)
    - Negative Selektion – Übergang wird unterdrückt
  - M(i,j) > 0 (Bruch > 1)
    - Positive Selektion – Übergang wird bevorzugt

Beispiel - PAM

$$M_x(i,j) = \log\left(\frac{f(i,j)}{f(i) * f(j)}\right)$$

```
ATGCCGTA
ATG_CCTA

CGCCACGT
_GCCTCGT

TCGCAGTA
TCG_GG_A
```

Gesucht:
M(G,G)
M(G,A)
M(G,C)

```
#N = 44
#G = 12
#A = 8
#C = 14

#Paare = 20
#GG = 5
#GA = 1
#GC = 1
```

$$M(G,G) = \log\left(\frac{\frac{5}{20}}{\frac{12}{44} * \frac{12}{44}}\right) = 0{,}53$$

$$M(G,A) = \log\left(\frac{\frac{1}{20}}{\frac{12}{44} * \frac{8}{44}}\right) = 0{,}00$$

$$M(G,C) = \log\left(\frac{\frac{1}{20}}{\frac{12}{44} * \frac{14}{44}}\right) = -0{,}24$$

Substitutionsmatrix fürs Alignment

- Gleiches Prinzip wie zur Berechnung des Editabstands
- Nur kleine Veränderungen

$$d(i,0) = \sum_{k=1}^{i} s(A[k], \_) \qquad d(0,j) = \sum_{k=1}^{j} s(\_, B[k])$$

$$d(i,j) = \max \begin{cases} d(i, j-1) + s(\_, B[j]) \\ d(i-1, j) + s(A[i], \_) \\ d(i-1, j-1) + s(A[i], B[j]) \end{cases}$$

Und jetzt in Java

- Einlesen der Matrix
  - Positionen der einzelnen Zeichen merken
    - Eindimensionales Array *letters*
  - Zahlenwerte für Zeichenpaar
    - Zweidimensionales Array *substValues*

- Schreiben der Funktion s(char1, char2)

Exercise 2

- Ergebnis:

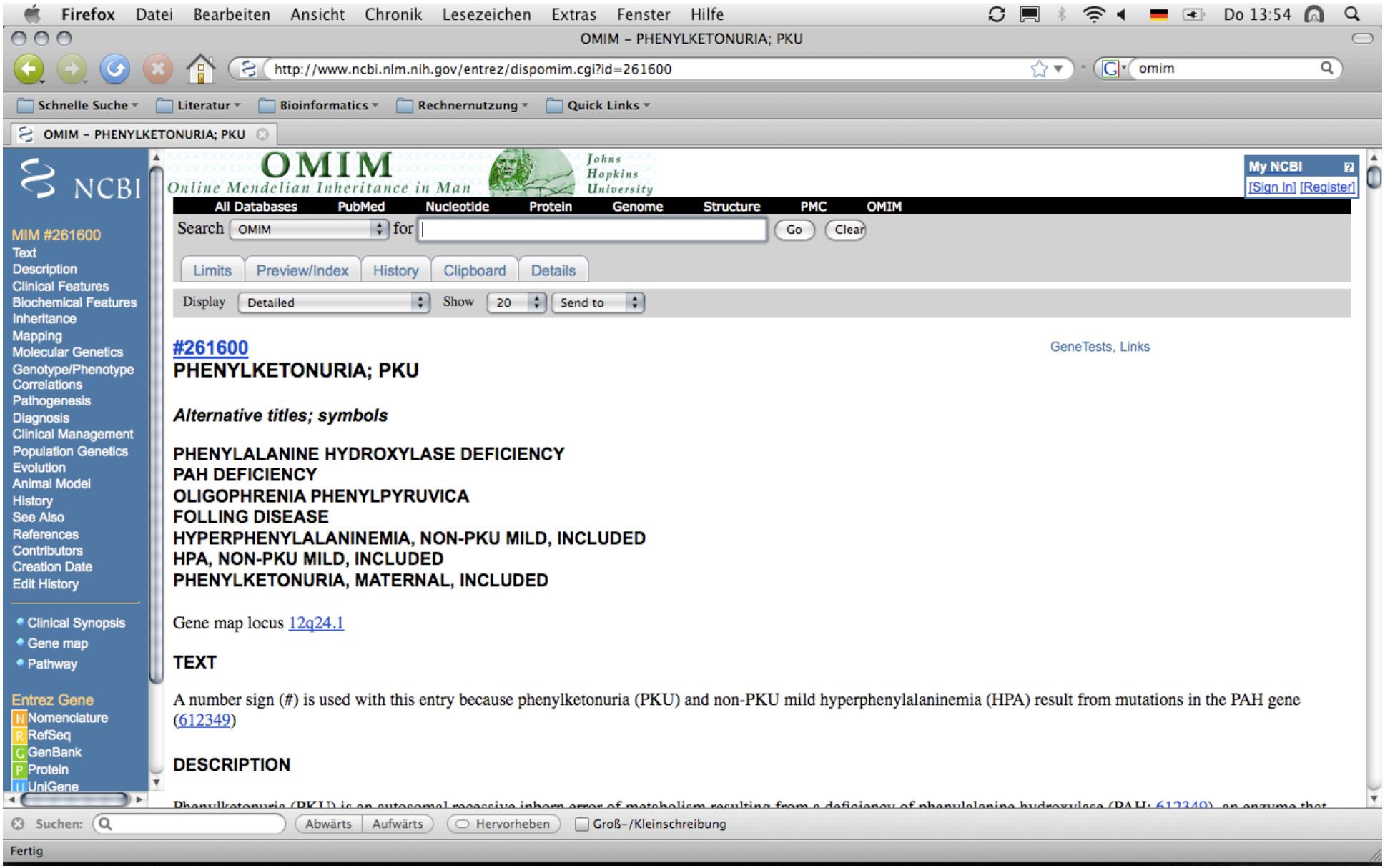
```
MAGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGS
M GQAFRKFL + DRVLVER+AAE V KGGIMLPEKSQGKVLQ TVVAVGS
MTGQAFRKFLLIADRVLVERNAAEIVAKGGIMLPEKSQGKVLQGTVVAVGS

GSKGKGGEIQPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDGDILGKYVD
G KGK GEI+PVSVKVGDKVLLPEYGG +VVLDDK         D
GRKGKSGEIEPVSVKVGDKVLLPEYGGNEVVLDDK_________D________
```

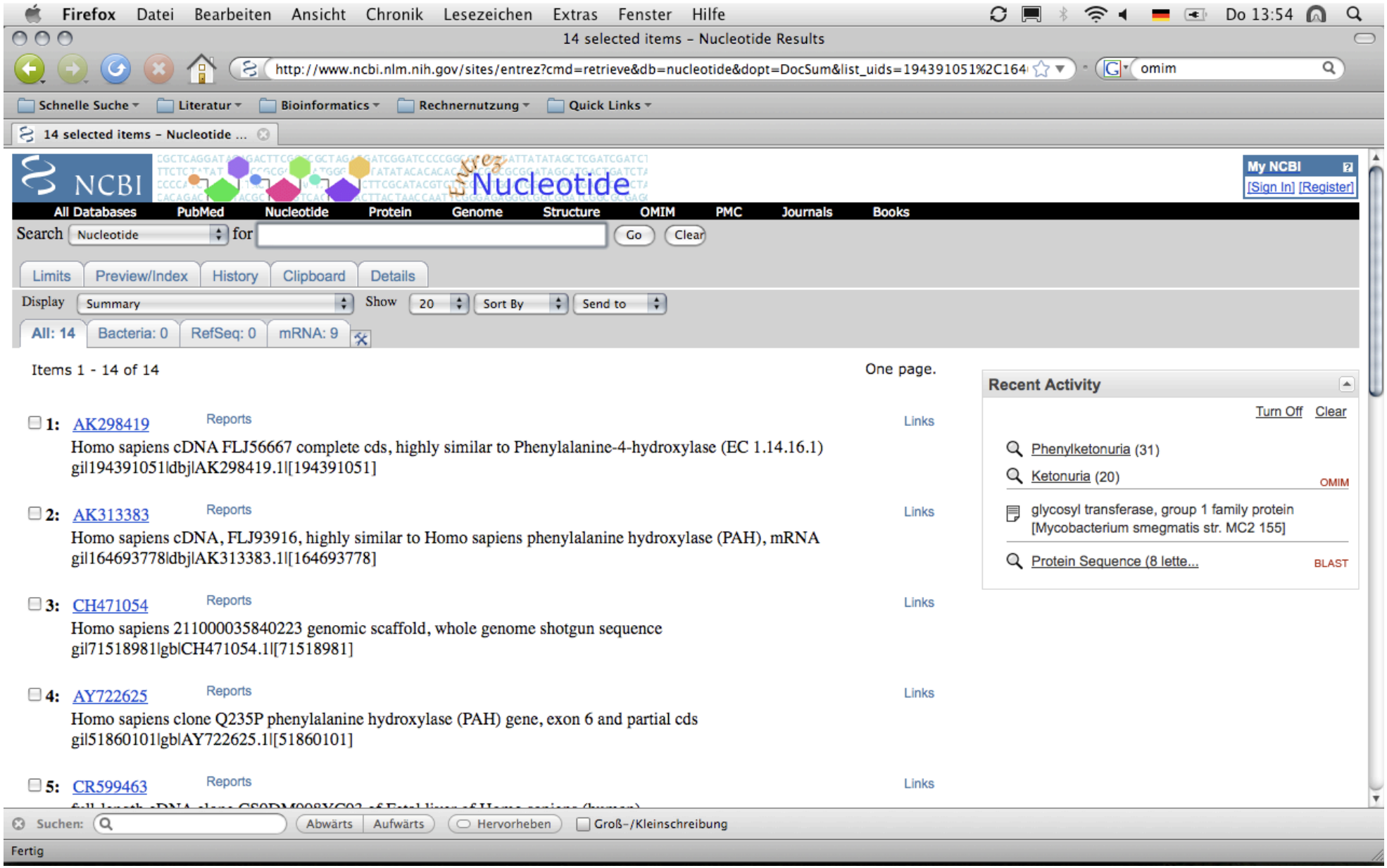
Matrix	Alignmentscore	# Alignments
Einheitsmatrix	47	30
Blosum50	394	30
Blosum62	306	15
Blosum80	297	15
Blosum100	323	15

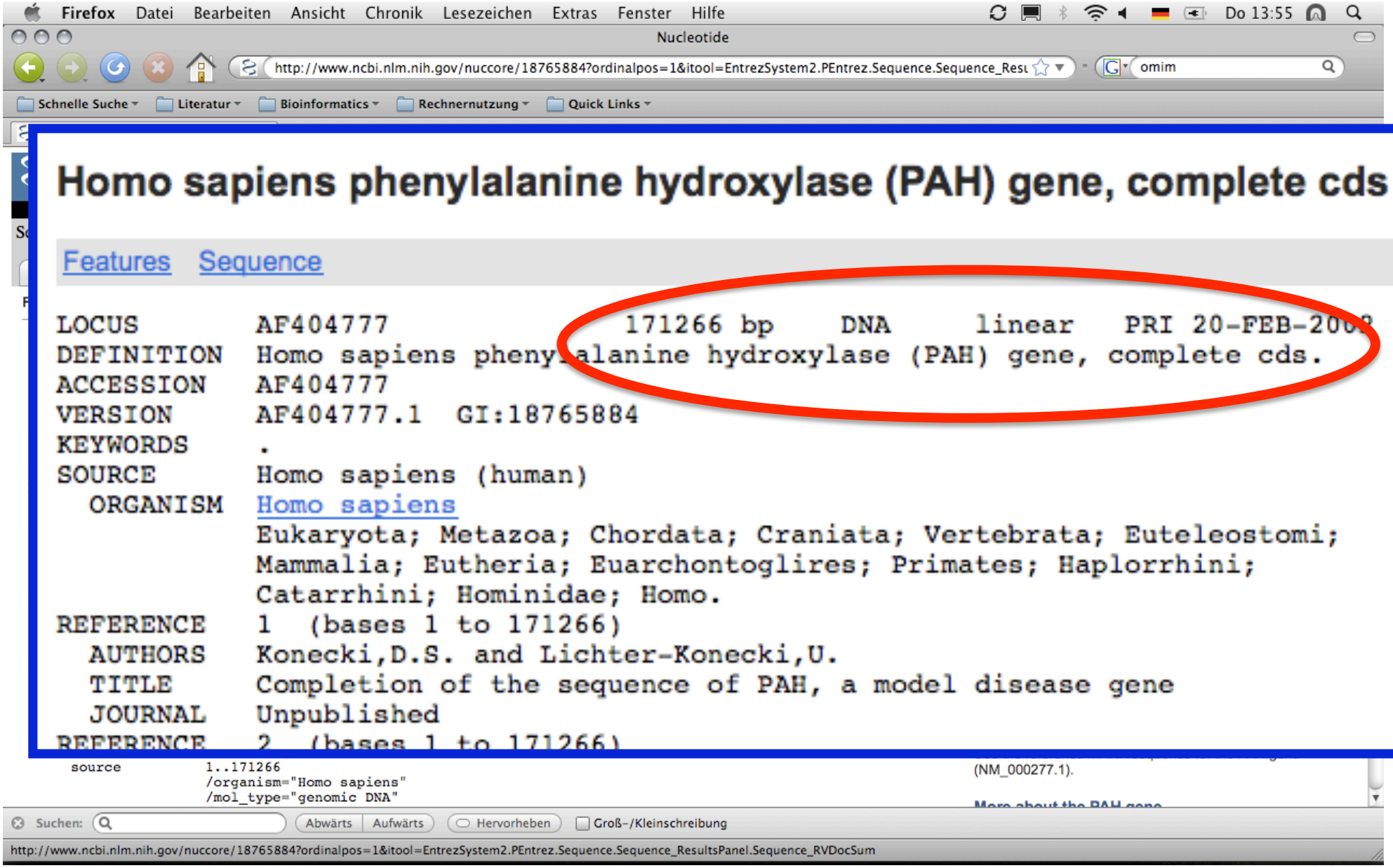
Exercise 3

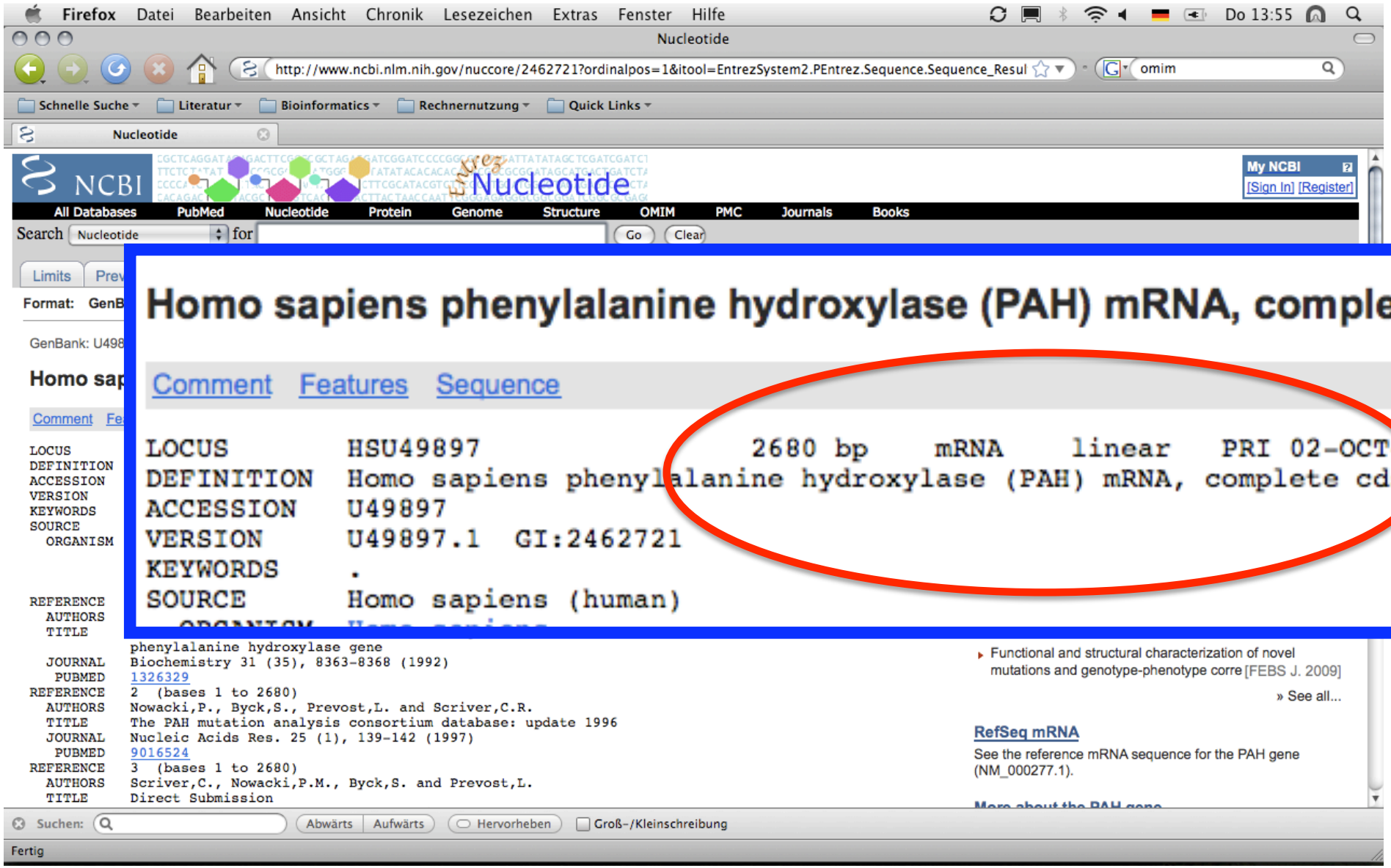
- Phenylketonuria
  - Erbkrankheit

- OMIM – **O**nline **M**endelian **I**nheritance in **M**an
  - Datenbank über Erbkrankheiten beim Menschen
    - Kuriertes Wissen zusammengetragen
    - Gen, Erscheinungsformen, Therapien, ...


Firefox Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Fenster Hilfe
Do 13:54
OMIM – PHENYLKETONURIA; PKU
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=261600
omim
Schnelle Suche
Literatur
Bioinformatics
Rechnernutzung
Quick Links
OMIM – PHENYLKETONURIA; PKU
NCBI
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC OMIM
Search OMIM for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display Detailed
Show 20
Send to
MIM #261600
Text
Description
Clinical Features
Biochemical Features
Inheritance
Mapping
Molecular Genetics
Genotype/Phenotype Correlations
Pathogenesis
Diagnosis
Clinical Management
Population Genetics
Evolution
Animal Model
History
See Also
References
Contributors
Creation Date
Edit History
Clinical Synopsis
Gene map
Pathway
Entrez Gene
Nomenclature
RefSeq
GenBank
Protein
UniGene
#261600
PHENYLKETONURIA; PKU
GeneTests, Links
Alternative titles; symbols
PHENYLALANINE HYDROXYLASE DEFICIENCY
PAH DEFICIENCY
OLIGOPHRENIA PHENYLPYRUVICA
FOLLING DISEASE
HYPERPHENYLALANINEMIA, NON-PKU MILD, INCLUDED
HPA, NON-PKU MILD, INCLUDED
PHENYLKETONURIA, MATERNAL, INCLUDED
Gene map locus 12q24.1
TEXT
A number sign (#) is used with this entry because phenylketonuria (PKU) and non-PKU mild hyperphenylalaninemia (HPA) result from mutations in the PAH gene (612349)
DESCRIPTION
Suchen:
Abwärts
Aufwärts
Hervorheben
Groß-/Kleinschreibung
Fertig


Firefox Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Fenster Hilfe
Do 13:54
14 selected items - Nucleotide Results
http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=nucleotide&dopt=DocSum&list_uids=194391051%2C164
omim
Schnelle Suche
Literatur
Bioinformatics
Rechnernutzung
Quick Links
14 selected items - Nucleotide ...
NCBI
Entrez Nucleotide
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search Nucleotide for Go Clear
Limits Preview/Index History Clipboard Details
Display Summary Show 20 Sort By Send to
All: 14 Bacteria: 0 RefSeq: 0 mRNA: 9
Items 1 - 14 of 14
One page.
1: AK298419 Reports Links
Homo sapiens cDNA FLJ56667 complete cds, highly similar to Phenylalanine-4-hydroxylase (EC 1.14.16.1)
gil194391051|dbj|AK298419.1|[194391051]
2: AK313383 Reports Links
Homo sapiens cDNA, FLJ93916, highly similar to Homo sapiens phenylalanine hydroxylase (PAH), mRNA
gil164693778|dbj|AK313383.1|[164693778]
3: CH471054 Reports Links
Homo sapiens 211000035840223 genomic scaffold, whole genome shotgun sequence
gil71518981|gb|CH471054.1|[71518981]
4: AY722625 Reports Links
Homo sapiens clone Q235P phenylalanine hydroxylase (PAH) gene, exon 6 and partial cds
gil51860101|gb|AY722625.1|[51860101]
5: CR599463 Reports Links
Recent Activity
Turn Off Clear
Phenylketonuria (31)
Ketonuria (20) OMIM
glycosyl transferase, group 1 family protein
[Mycobacterium smegmatis str. MC2 155]
Protein Sequence (8 lette... BLAST
Suchen: Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung
Fertig


Firefox Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Fenster Hilfe
Do 13:55
Nucleotide
http://www.ncbi.nlm.nih.gov/nuccore/18765884?ordinalpos=1&itool=EntrezSystem2.PEntrez.Sequence.Sequence_Resu
omim
Schnelle Suche
Literatur
Bioinformatics
Rechnernutzung
Quick Links
Homo sapiens phenylalanine hydroxylase (PAH) gene, complete cds
Features
Sequence
LOCUS AF404777 171266 bp DNA linear PRI 20-FEB-2003
DEFINITION Homo sapiens phenylalanine hydroxylase (PAH) gene, complete cds.
ACCESSION AF404777
VERSION AF404777.1 GI:18765884
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 171266)
AUTHORS Konecki,D.S. and Lichter-Konecki,U.
TITLE Completion of the sequence of PAH, a model disease gene
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 171266)
source 1..171266
/organism="Homo sapiens"
/mol_type="genomic DNA"
(NM_000277.1).
Suchen:
Abwärts
Aufwärts
Hervorheben
Groß-/Kleinschreibung
http://www.ncbi.nlm.nih.gov/nuccore/18765884?ordinalpos=1&itool=EntrezSystem2.PEntrez.Sequence.Sequence_ResultsPanel.Sequence_RVDocSum


Firefox Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Fenster Hilfe
Do 13:55
Nucleotide
http://www.ncbi.nlm.nih.gov/nuccore/2462721?ordinalpos=1&itool=EntrezSystem2.PEntrez.Sequence.Sequence_Resul
omim
Schnelle Suche
Literatur
Bioinformatics
Rechnernutzung
Quick Links
NCBI
Nucleotide
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search Nucleotide for
Go
Clear
Limits
Format:
Homo sapiens phenylalanine hydroxylase (PAH) mRNA, complete cds
Comment Features Sequence
LOCUS HSU49897 2680 bp mRNA linear PRI 02-OCT-1997
DEFINITION Homo sapiens phenylalanine hydroxylase (PAH) mRNA, complete cds.
ACCESSION U49897
VERSION U49897.1 GI:2462721
KEYWORDS .
SOURCE Homo sapiens (human)
REFERENCE
AUTHORS
TITLE
phenylalanine hydroxylase gene
JOURNAL Biochemistry 31 (35), 8363-8368 (1992)
PUBMED 1326329
REFERENCE 2 (bases 1 to 2680)
AUTHORS Nowacki,P., Byck,S., Prevost,L. and Scriver,C.R.
TITLE The PAH mutation analysis consortium database: update 1996
JOURNAL Nucleic Acids Res. 25 (1), 139-142 (1997)
PUBMED 9016524
REFERENCE 3 (bases 1 to 2680)
AUTHORS Scriver,C., Nowacki,P.M., Byck,S. and Prevost,L.
TITLE Direct Submission
Functional and structural characterization of novel mutations and genotype-phenotype corre [FEBS J. 2009]
» See all...
RefSeq mRNA
See the reference mRNA sequence for the PAH gene (NM_000277.1).
Suchen:
Abwärts
Aufwärts
Hervorheben
Groß-/Kleinschreibung
Fertig

BLAST Parameter

- Zunächst
  - Suche nach guten lokalen Alignments in DNA Sequenzen
- Gegeben
  - Suchsequenz P, Datenbank DB={S_1,...,S_n}
  - Substitutionsmatrix M
  - Parameter w: Länge der „Seeds“
  - Parameter t: Initialer Schwellwert
  - Parameter c: Gesamtschwellwert
    - Wird berechnet in Abhängigkeit von t, M, |DB|, |P|
  - Parameter v: Erwünschte Anzahl Treffer
    - Blast berechnet die v ähnlichsten Subsequenzen

BLAST Schritt 1 und 2

- Schritt 1
  - Bestimme alle Teilwörter $P_1,...,P_m$ der Länge w in P
    - Mit Überlappung – keine Partitionierung
    - Wie viele gibt es?
- Schritt 2
  - Suche nach Hits von $P_1,...,P_m$ in DB mit Score über t
    - Hits müssen nicht exakt sein
    - Vergleiche alle P_i mit allen Teilwörtern in DB der Länge w in T
      - Keine INSDELs, Verwendung von M
  - Durch den „unscharfen" Hit mit Schwellwert t
    - werden auch Hits gefunden, die keine perfekten Matches sind
    - werden wenig aussichtsreiche Positionen in DB ausgeschlossen
      - Annahme: Ein gutes lokales Alignment A in Sequenz S muss mindestens einen Hit enthalten
      - Es ist nicht notwendig, dass jedes Teilwort in A mit seinem Teilwortpartner in P einen Hit hat.

BLAST Schritt 3

- Schritt 3
  - Für jeden Hit H zwischen DB-Sequenz S und P_i
  - Verlängere Bereich um H sowohl in P als auch in S
    - Gapfree Alignment nach links und rechts wachsen
    - Solange, bis
      - Sequenz P oder S zu Ende ist, oder
      - Alignmentscore fällt unter geschätzten Schwellwert c, oder
      - Alignmentscore fällt „signifikant“ unter bisherige v beste Treffer
        - „Signifikant“ heuristisch bestimmt
    - Ergibt „Maximal Segment Pairs (MSP)“
    - Die besten v MSP sind das Ergebnis

Beispiel

```
W=5, t=5, Kosten: M=+1, R=-3
P=ACGTGATA
S=GATTGACGTGACTGCAAGTGATACTATAT
```

Schritt 1
Teilwörter

P_1=ACGTG
P_2=CGTGA
P_3=GTGAT
P_4=TGATA

Schritt 2
Hitsuche

```
GATTGACGTGACTGCAAGTGATACTATAT
GATTGACGTGACTGCAAGTGATACTATAT
GATTGACGTGACTGCAAGTGATACTATAT
```

Schritt 3
Verlängerung

```
GATTGACGTGACTGCAAGTGATACTATAT
      ACGTGATA                  5
      ACGTGATA                  5+1=6
      ACGTGATA                  6-3=3
      ...                       ...
```

Sensitivität und Spezifität

		Reality +	Reality -
Prediction	+	TruePositive (TP)	FalsePositive (FP)
	-	FalseNegative (FN)	TrueNegative (TN)

- Spezifität = TP/(TP+FP) (Precision)
  - Wie viele der Treffen des Verfahrens sind wirklich welche?
- Sensitivität = TP/(TP+FN) (Recall)
  - Wie viele der echten Treffer findet das Verfahren?
- Oftmals eine Balance
  - Algorithmen berechnen einen Score pro Sequenz
  - Hoher Score – Positiv; Niedriger Score – Negativ
  - Wenn Score mit Wahrscheinlichkeit für korrekte Klassifikation korreliert, folgt daraus
    - Ergebnismenge klein: SP=hoch, SE=klein
    - Ergebnismenge groß: SP=niedrig, SE=hoch

Exercise 3

- Alle Sequenzen mit allen vergleichen
  - ähnlichste: Mus – Rattus Score: 2830
  - unähnlichsten: Homo – Drosophila Score: 1529

- Ähnlichkeitsmatrix notwendig für ClustalW
  - ähnlichsten zuerst zum multiplen Sequenzalignment zusammenfassen

Klausur

- 30.07.2009
- 14 – 16 Uhr
- Hörsaal 12

Zuletzt!

Weitere

Fragen ?