**UNIX filter:**

1. **cut -f3** **/data/biocs/b/bio425/data/codon.table.T | sort**

\*

\*

\*

A

A

A

A

C

C

D

D

E

E

F

F

G

G

G

G

H

H

I

I

I

K

K

L

L

L

L

L

L

M

N

N

P

P

P

P

Q

Q

R

R

R

R

R

R

S

S

S

S

S

S

T

T

T

T

V

V

V

V

W

Y

Y

1. **cut -f1 /data/biocs/b/bio425/data/codon.table.T | grep -o "A[A-Z][A-Z]"**

AAA

AAC

AAG

AAT

ACA

ACC

ACG

ACT

AGA

AGC

AGG

AGT

ATA

ATC

ATG

ATT

1. **grep "Arg"** **/data/biocs/b/bio425/data/codon.table.T | awk -F" " '{print $1}'**

AGA

AGG

CGA

CGC

CGG

CGT

1. **read-codon.bash**

**#!/bin/bash**

**file="/data/biocs/b/bio425/data/codon.table.T " ;**

**grep -v "\*" $file | while read line ; do**

**three\_letters\_codon=$(echo $line | cut -f1 -d' ') ;**

**amino\_acids=$(echo $line | cut -f2 -d' ' ) ;**

**symbols=$(echo $line | cut -f3 -d' ' ) ;**

**echo "$three\_letters\_codon codes for $amino\_acids ( $symbols)." ;**

**done ;**

**exit ;**

AAA codes for Lys ( K).

AAC codes for Asn ( N).

AAG codes for Lys ( K).

AAT codes for Asn ( N).

ACA codes for Thr ( T).

ACC codes for Thr ( T).

ACG codes for Thr ( T).

ACT codes for Thr ( T).

AGA codes for Arg ( R).

AGC codes for Ser ( S).

AGG codes for Arg ( R).

AGT codes for Ser ( S).

ATA codes for Ile ( I).

ATC codes for Ile ( I).

ATG codes for Met ( M).

ATT codes for Ile ( I).

CAA codes for Gln ( Q).

CAC codes for His ( H).

CAG codes for Gln ( Q).

CAT codes for His ( H).

CCA codes for Pro ( P).

CCC codes for Pro ( P).

CCG codes for Pro ( P).

CCT codes for Pro ( P).

CGA codes for Arg ( R).

CGC codes for Arg ( R).

CGG codes for Arg ( R).

CGT codes for Arg ( R).

CTA codes for Leu ( L).

CTC codes for Leu ( L).

CTG codes for Leu ( L).

CTT codes for Leu ( L).

GAA codes for Glu ( E).

GAC codes for Asp ( D).

GAG codes for Glu ( E).

GAT codes for Asp ( D).

GCA codes for Ala ( A).

GCC codes for Ala ( A).

GCG codes for Ala ( A).

GCT codes for Ala ( A).

GGA codes for Gly ( G).

GGC codes for Gly ( G).

GGG codes for Gly ( G).

GGT codes for Gly ( G).

GTA codes for Val ( V).

GTC codes for Val ( V).

GTG codes for Val ( V).

GTT codes for Val ( V).

TAC codes for Tyr ( Y).

TAT codes for Tyr ( Y).

TCA codes for Ser ( S).

TCC codes for Ser ( S).

TCG codes for Ser ( S).

TCT codes for Ser ( S).

TGC codes for Cys ( C).

TGG codes for Trp ( W).

TGT codes for Cys ( C).

TTA codes for Leu ( L).

TTC codes for Phe ( F).

TTG codes for Leu ( L).

TTT codes for Phe ( F).

* **Exercise for the use of wildcard & BASH "for" loop:**

1. **Bioseq -l /data/biocs/b/bio425/data/GBB.1con-splitted /Borrelia\_burgdorferi\_3615\_main.fas**

Borrelia\_burgdorferi\_3615\_main 910724

1. **ls -l /data/biocs/b/bio425/data/GBB.1con-splitted/\*.fas**

-rw-r--r-- 1 weigangq student 925943 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_3615\_main.fas

-rw-r--r-- 1 weigangq student 29131 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4004\_lp28-3\_plasmid\_H.fas

-rw-r--r-- 1 weigangq student 24631 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4005\_lp25\_plasmid\_E.fas

-rw-r--r-- 1 weigangq student 30316 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4006\_lp28-2\_plasmid\_G.fas

-rw-r--r-- 1 weigangq student 39528 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4007\_lp38\_plasmid\_J.fas

-rw-r--r-- 1 weigangq student 37515 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4008\_lp36\_plasmid\_K.fas

-rw-r--r-- 1 weigangq student 27832 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4009\_lp28-4\_plasmid\_I.fas

-rw-r--r-- 1 weigangq student 27423 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4013\_lp28-1\_plasmid\_F.fas

-rw-r--r-- 1 weigangq student 9593 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4041\_cp9\_plasmid\_C.fas

-rw-r--r-- 1 weigangq student 54505 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4075\_lp54\_plasmid\_A.fas

-rw-r--r-- 1 weigangq student 17155 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4076\_lp17\_plasmid\_D.fas

-rw-r--r-- 1 weigangq student 5366 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4077\_lp5\_plasmid\_T.fas

-rw-r--r-- 1 weigangq student 31453 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4080\_cp32-8\_plasmid\_L.fas

-rw-r--r-- 1 weigangq student 53905 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4082\_lp56\_plasmid\_Q.fas

-rw-r--r-- 1 weigangq student 30857 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4083\_cp32-4\_plasmid\_R.fas

-rw-r--r-- 1 weigangq student 30389 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4084\_cp32-6\_plasmid\_M.fas

-rw-r--r-- 1 weigangq student 31367 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4085\_cp32-7\_plasmid\_O.fas

-rw-r--r-- 1 weigangq student 30780 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4087\_cp32-3\_plasmid\_S.fas

-rw-r--r-- 1 weigangq student 19117 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4089\_lp21\_plasmid\_U.fas

-rw-r--r-- 1 weigangq student 26991 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4091\_cp26\_plasmid\_B.fas

-rw-r--r-- 1 weigangq student 31215 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4092\_cp32-9\_plasmid\_N.fas

-rw-r--r-- 1 weigangq student 31316 Nov 11 2005 /data/biocs/b/bio425/data/GBB.1con-splitted/Borrelia\_burgdorferi\_4093\_cp32-1\_plasmid\_P.fas

1. **for fasfile in /data/biocs/b/bio425/data/GBB.1con-splitted/\*\*\*.fas**

**-> do**

**-> bioseq -l $fasfile**

**-> done**

Borrelia\_burgdorferi\_3615\_main 910724

Borrelia\_burgdorferi\_4004\_lp28-3(plasmid\_H) 28601

Borrelia\_burgdorferi\_4005\_lp25(plasmid\_E) 24177

Borrelia\_burgdorferi\_4006\_lp28-2(plasmid\_G) 29766

Borrelia\_burgdorferi\_4007\_lp38(plasmid\_J) 38829

Borrelia\_burgdorferi\_4008\_lp36(plasmid\_K) 36849

Borrelia\_burgdorferi\_4009\_lp28-4(plasmid\_I) 27323

Borrelia\_burgdorferi\_4013\_lp28-1(plasmid\_F) 26921

Borrelia\_burgdorferi\_4041\_cp9(plasmid\_C) 9386

Borrelia\_burgdorferi\_4075\_lp54(plasmid\_A) 53561

Borrelia\_burgdorferi\_4076\_lp17(plasmid\_D) 16823

Borrelia\_burgdorferi\_4077\_lp5(plasmid\_T) 5228

Borrelia\_burgdorferi\_4080\_cp32-8(plasmid\_L) 30885

Borrelia\_burgdorferi\_4082\_lp56(plasmid\_Q) 52971

Borrelia\_burgdorferi\_4083\_cp32-4(plasmid\_R) 30299

Borrelia\_burgdorferi\_4084\_cp32-6(plasmid\_M) 29838

Borrelia\_burgdorferi\_4085\_cp32-7(plasmid\_O) 30800

Borrelia\_burgdorferi\_4087\_cp32-3(plasmid\_S) 30223

Borrelia\_burgdorferi\_4089\_lp21(plasmid\_U) 18753

Borrelia\_burgdorferi\_4091\_cp26(plasmid\_B) 26498

Borrelia\_burgdorferi\_4092\_cp32-9(plasmid\_N) 30651

Borrelia\_burgdorferi\_4093\_cp32-1(plasmid\_P) 30750

* **BLAST exercise :**

1. **blastp -query ../../bio425/data/unknown.pep -db ref -evalue 1e-10**

BLASTP 2.2.31+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.

Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.

Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of

protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference for composition-based statistics: Alejandro A. Schaffer,

L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri

I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001),

"Improving the accuracy of PSI-BLAST protein database searches with

composition-based statistics and other refinements", Nucleic Acids

Res. 29:2994-3005.

Database: GBB.pep

1,639 sequences; 432,633 total letters

Query= lcl|N40:1719:ORFJ00057 unnamed protein product

Length=274

Score E

Sequences producing significant alignments: (Bits) Value

BBA15 unnamed protein product 513 0.0

BBA16 unnamed protein product 227 9e-74

>BBA15 unnamed protein product

Length=273

Score = 513 bits (1322), Expect = 0.0, Method: Compositional matrix adjust.

Identities = 271/273 (99%), Positives = 271/273 (99%), Gaps = 0/273 (0%)

Query 1 MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEMNVLVSKEKNKDGKYDLIATVDK 60

MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEM VLVSKEKNKDGKYDLIATVDK

Sbjct 1 MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEMKVLVSKEKNKDGKYDLIATVDK 60

Query 61 LELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQTTLEVFKEDGKTLVSKKVTSKDKS 120

LELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQTTLEVFKEDGKTLVSKKVTSKDKS

Sbjct 61 LELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQTTLEVFKEDGKTLVSKKVTSKDKS 120

Query 121 STEEKFNEKGEVSEKIITRADGTRLEYTEIKSDGSGKAKEVLKGYVLEGTLTAEKTTLVV 180

STEEKFNEKGEVSEKIITRADGTRLEYT IKSDGSGKAKEVLKGYVLEGTLTAEKTTLVV

Sbjct 121 STEEKFNEKGEVSEKIITRADGTRLEYTGIKSDGSGKAKEVLKGYVLEGTLTAEKTTLVV 180

Query 181 KEGTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKE 240

KEGTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKE

Sbjct 181 KEGTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKE 240

Query 241 NTITVQQYDSNGTKLEGSAVEITKLDEIKNALK 273

NTITVQQYDSNGTKLEGSAVEITKLDEIKNALK

Sbjct 241 NTITVQQYDSNGTKLEGSAVEITKLDEIKNALK 273

>BBA16 unnamed protein product

Length=296

Score = 227 bits (578), Expect = 9e-74, Method: Compositional matrix adjust.

Identities = 149/299 (50%), Positives = 184/299 (62%), Gaps = 32/299 (11%)

Query 3 KYLLGIGLILALIACKQ--------------NV--SSLDEKNSVSVDLPGEMN------- 39

+ L+G L LALI C Q N+ SS + DLP

Sbjct 2 RLLIGFALALALIGCAQKGAESIGSQKENDLNLEDSSKKSHQNAKQDLPAVTEDSVSLFN 61

Query 40 ---VLVSKEKNKDGKYDLIATVDKLELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQ 96

+ VSKEKN GKYDL AT+D++ELKGTSDKNNGSG LEG K DKSKVKLT+S DL

Sbjct 62 GNKIFVSKEKNSSGKYDLRATIDQVELKGTSDKNNGSGTLEGSKPDKSKVKLTVSADLNT 121

Query 97 TTLEVFKEDGKTLVSKKVTSKDKSSTEEKFNEKGEVSEKIITRADGTRLEYTEI-KSDGS 155

TLE F + +S KVT K S T E + ++ K +TR++GT LEY++I +D +

Sbjct 122 VTLEAFDASNQK-ISSKVTKKQGSITXETL-KANKLDSKKLTRSNGTTLEYSQITDADNA 179

Query 156 GKAKEVLKGYV-LEGTLTAEKTTLVVKEGTVTLSKNISKSGEVSVELNDTDSSAATKKTA 214

KA E LK + LEG+L KTT+ +KEGTVTL + I K G+V V LNDT + + KKT

Sbjct 180 TKAVETLKNSIKLEGSLVGGKTTVEIKEGTVTLKREIEKDGKVKVFLNDT--AGSNKKTG 237

Query 215 AWNSGTSTLTITVNSKKTKDLVFTKENTITVQQYDSNGTKLEGSAVEITKLDEIKNALK 273

W TSTLTI+ +SKKTKDLVF + TITVQQY++ GT LEGSA EI L E+KNALK

Sbjct 238 KWEDSTSTLTISADSKKTKDLVFLTDGTITVQQYNTAGTSLEGSASEIKNLSELKNALK 296

Lambda K H a alpha

0.305 0.125 0.320 0.792 4.96

Gapped

Lambda K H a alpha sigma

0.267 0.0410 0.140 1.90 42.6 43.6

Effective search space used: 62269400

Database: GBB.pep

Posted date: Feb 27, 2019 12:04 PM

Number of letters in database: 432,633

Number of sequences in database: 1,639

Matrix: BLOSUM62

Gap Penalties: Existence: 11, Extension: 1

Neighboring words threshold: 11

Window for multiple hits: 40

1. >BBA15 unnamed protein product

Length=273

Score = 513 bits (1322), Expect = 0.0, Method: Compositional matrix adjust.

Identities = 271/273 (99%), Positives = 271/273 (99%), Gaps = 0/273 (0%)

Query 1 MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEMNVLVSKEKNKDGKYDLIATVDK 60

MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEM VLVSKEKNKDGKYDLIATVDK

Sbjct 1 MKKYLLGIGLILALIACKQNVSSLDEKNSVSVDLPGEMKVLVSKEKNKDGKYDLIATVDK 60

Query 61 LELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQTTLEVFKEDGKTLVSKKVTSKDKS 120

LELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQTTLEVFKEDGKTLVSKKVTSKDKS

Sbjct 61 LELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQTTLEVFKEDGKTLVSKKVTSKDKS 120

Query 121 STEEKFNEKGEVSEKIITRADGTRLEYTEIKSDGSGKAKEVLKGYVLEGTLTAEKTTLVV 180

STEEKFNEKGEVSEKIITRADGTRLEYT IKSDGSGKAKEVLKGYVLEGTLTAEKTTLVV

Sbjct 121 STEEKFNEKGEVSEKIITRADGTRLEYTGIKSDGSGKAKEVLKGYVLEGTLTAEKTTLVV 180

Query 181 KEGTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKE 240

KEGTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKE

Sbjct 181 KEGTVTLSKNISKSGEVSVELNDTDSSAATKKTAAWNSGTSTLTITVNSKKTKDLVFTKE 240

Query 241 NTITVQQYDSNGTKLEGSAVEITKLDEIKNALK 273

NTITVQQYDSNGTKLEGSAVEITKLDEIKNALK

Sbjct 241 NTITVQQYDSNGTKLEGSAVEITKLDEIKNALK 273

>BBA16 unnamed protein product

Length=296

Score = 227 bits (578), Expect = 9e-74, Method: Compositional matrix adjust.

Identities = 149/299 (50%), Positives = 184/299 (62%), Gaps = 32/299 (11%)

Query 3 KYLLGIGLILALIACKQ--------------NV--SSLDEKNSVSVDLPGEMN------- 39

+ L+G L LALI C Q N+ SS + DLP

Sbjct 2 RLLIGFALALALIGCAQKGAESIGSQKENDLNLEDSSKKSHQNAKQDLPAVTEDSVSLFN 61

Query 40 ---VLVSKEKNKDGKYDLIATVDKLELKGTSDKNNGSGVLEGVKADKSKVKLTISDDLGQ 96

+ VSKEKN GKYDL AT+D++ELKGTSDKNNGSG LEG K DKSKVKLT+S DL

Sbjct 62 GNKIFVSKEKNSSGKYDLRATIDQVELKGTSDKNNGSGTLEGSKPDKSKVKLTVSADLNT 121

Query 97 TTLEVFKEDGKTLVSKKVTSKDKSSTEEKFNEKGEVSEKIITRADGTRLEYTEI-KSDGS 155

TLE F + +S KVT K S T E + ++ K +TR++GT LEY++I +D +

Sbjct 122 VTLEAFDASNQK-ISSKVTKKQGSITXETL-KANKLDSKKLTRSNGTTLEYSQITDADNA 179

Query 156 GKAKEVLKGYV-LEGTLTAEKTTLVVKEGTVTLSKNISKSGEVSVELNDTDSSAATKKTA 214

KA E LK + LEG+L KTT+ +KEGTVTL + I K G+V V LNDT + + KKT

Sbjct 180 TKAVETLKNSIKLEGSLVGGKTTVEIKEGTVTLKREIEKDGKVKVFLNDT--AGSNKKTG 237

Query 215 AWNSGTSTLTITVNSKKTKDLVFTKENTITVQQYDSNGTKLEGSAVEITKLDEIKNALK 273

W TSTLTI+ +SKKTKDLVF + TITVQQY++ GT LEGSA EI L E+KNALK

Sbjct 238 KWEDSTSTLTISADSKKTKDLVFLTDGTITVQQYNTAGTSLEGSASEIKNLSELKNALK 296

1. **blastp -query ../../bio425/data/unknown.pep -db ref -evalue 1e-10 -outfmt 6**

lcl|N40:1719:ORFJ00057 BBA15 99.27 273 2 0 1 2731273 0.0 513

lcl|N40:1719:ORFJ00057 BBA16 49.83 299 118 8 3 2732296 9e-74 227

1. **Explain :**

**Identity = number of identical bases between the query and the subject sequence.**

**Positives = which corresponds to the number of amino acids that are either identical**

**between the query and the subject sequence or have similar chemical properties. Number of residues that are identical or have similar chemical properties.**

**Expect value = the statistical significance of the alignment, sequence. It is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.**