SCORING AMINO ACID SUBSTITUTIONS IN ΦHAGE GENOMES

Promita Bose,Robert Edwards,Barbara Bailey,Peter Salamon
Computational Science Research Center, San Diego State University
Contact pbose@sciences.sdsu.edu

Abstract

Substitution matrices are among the most widely used scoring techniques: BLAST, Phylip and other alignment packages, all use them. However these matrices are general; they ignore organism specific properties and do not provide customized scoring schemes. We present a phage- specific scoring matrix based on the abundances of aligned substitutions. These matrices use information from approximately five and a half million significantly similar protein alignments from over five hundred phage genomes. Our scoring matrix is different from the existing PAM and BLOSUM matrices. This indicates the need for similar treatments for other groups of organisms.

Methods

MATRIX OF REPLACEMENTS

	A	R	N	D	C	Q	E	G	H	V
A	2540121	777058	861656	998387	170987	812685	1247322	1625804	226812	.....
R	777058	1334675	571128	523766	53809	662445	750317	570505	208075	.....
N	861656	571128	1328653	1219297	54530	584579	908350	911480	25020	.....
D	998387	523766	1219297							
C	170987	53809								
Q	812685	662445								
E	1247322									
G	1625804									
H	226812	208075								
V										

Preliminary Conclusions

CORRELATION COEFFICIENT MATRIX

	ΦHAGE	Blosum45	Blosum50	Blosum62	Blosum80	Blosum90
ΦHAGE	1.000	0.8794	0.8956	0.8973	0.8898	0.8768
Blosum45	0.8794	1.0000	0.9862	0.9726	0.9677	0.9646
Blosum50	0.8956	0.9862	1.0000	0.9826	0.9781	0.9753
Blosum62	0.8973	0.9726	0.9826	1.0000	0.9847	0.9841
Blosum80	0.8898	0.9677	0.9781	0.9847	1.0000	0.9884
Blosum90	0.8768	0.9646	0.9753	0.9841	0.9884	1.0000

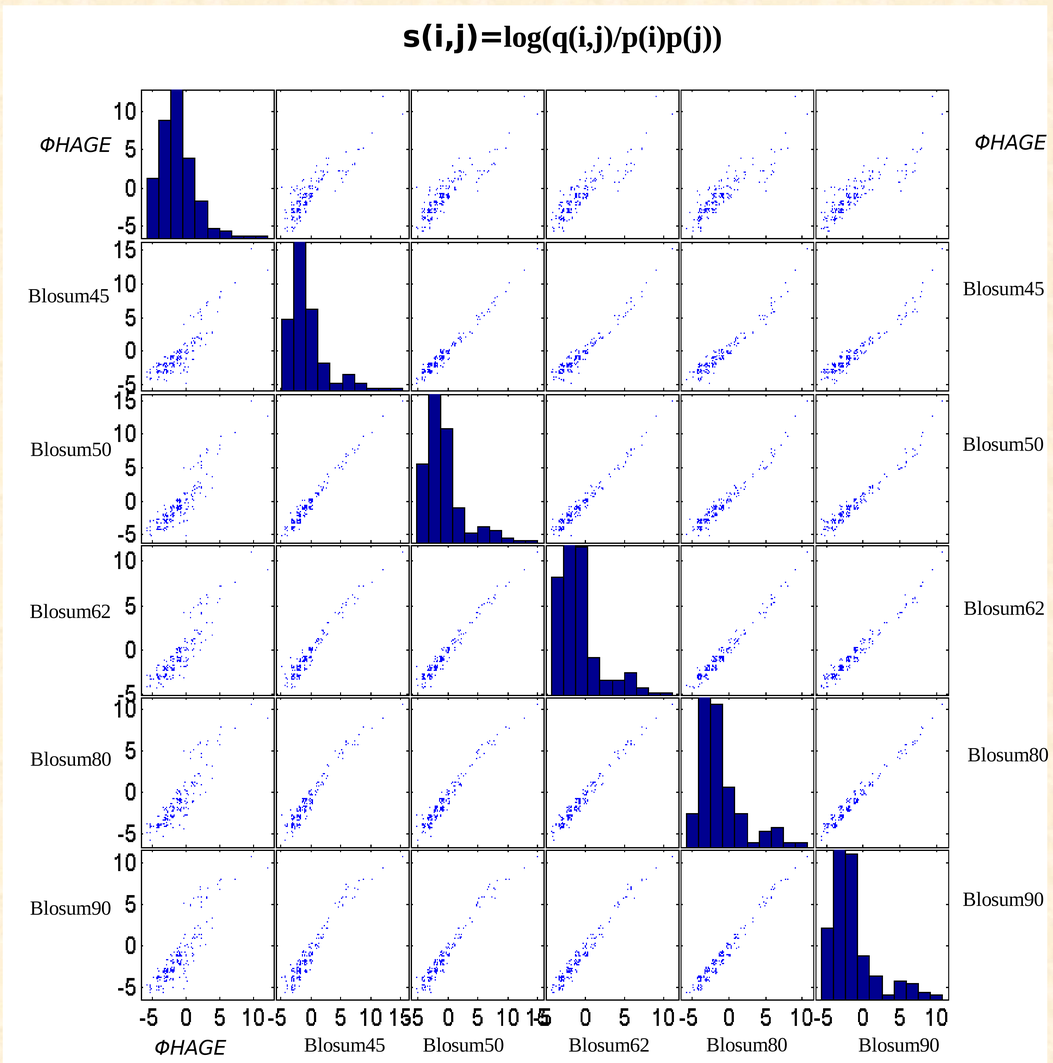
→ Suggests a significant level of difference between the phage scoring matrix and the standard matrices

Sample entry from Database

Putative platelet-binding protein from a phage-associated region of Streptococcus pyogenes MGAS315
Rz endopeptidase from Stx2 converting phage II

```
     2    657
9729       MEVQGLISIK DDNTPLVAAF EDEITQEANS DYKLSFKYPA KHEYRPLIKK
12933      ---------- ---------- ---------- ---------- ----------

           GIILEADDLH GTQLFRVFEI TKRHGYINVY GNQVADDLNG YAIDSISVDR
           ---------- ---------- ---------- ---------- ----------

           VQGMTVMSEL AGNIKREHPF SFFSDIDGRH TFNQSDVSVM DALANGKHSI
           ---------- ---------- ---------- ---------- ----------

           MGQWGGELVR DKYQVNLLKK AGKDTETLFM YKKNLKSYEE TDTIKGLISI
           ---------- ---------- ---------- ---------- ---MNRVLCV

           LHLVAEVEEQ REEESREVSD ADIGHNEVSK KKTIRVTVES KLKDTHPIIV
           VIIVLAVG-- ---------- ---------- ---------- ----------

           EKTIKVQDQD VKTEEDLLAY GKKYFEKTLC DIPGNSLKID VTNNYEGAVR
           ---------- ---------Y GALWLA---- ---------- -TNHYR----

           LFDTAIVFHE LYDRDLRMQI TGYRFAPMAN RLKSIIFGEI KTNLAKQISN
           --DNALTYKA QRDKKAR--- ---------- ---------- --ELEQANAT

           QIDNKVAEST AQHDAAFEAK LQKQIDNANR IFDTKEAKLR EEIEDGIKKA
           ITDMQVRQ-- -RDVAALDAK YSRELADAR- ---AENETLR ADVAAGRKRL

           EANAEVKVAE VNAKVLEAEE LAKAVDERLK KFLSDADTKE QDFDKKLEEF
           RINATCSGTV R--EATGTSG VDNATGPRL- -----ADTAE RDYFILRERL

           RTSLKDLEVD EKQIDDALAK AGFSKDSLAD IKAKLEDTSE TATVTANIVG
           ITMQKQLEGT QKYINEQCR- ---------- ---------- ----------

           STGGTFYNRN RLDGDTDKVI TFEQGYIDIA HNGEGFEEGK TYTISFEATC
           ---------- ---------- ---------- ---------- ----------

           ELLRKVGITV TQANMKGARL VLTPKNPKLV VESFGLTKDT ETINVYPFSY
           ---------- ---------- ---------- ---------- ----------

           TVLVTSDWYK SKQIDLNASE VQELALEMAY KDVVDGNNAT IAGQWSDSPQ
           ---------- ---------- ---------- ---------- ----------

           MILDGGS
           -------
```

COMPARISON OF AMINO ACID SUBSTITUTION SCORES



q(i,j) are the expected probability to observe residues i and j aligned in homologous sequence alignments. p(i) and p(j) are the background frequencies of observing amino acids i and j

Future Work

Is the difference statistically significant ?

- Capture the significance

Are all factors pertinent to amino acid scoring matrices balanced among the two datasets ?

References

A Protein Alignment Scoring System Sensitive at All Evolutionary Distances
J Mol Evol (1993) 36:290-300
Stephen F. Altschul

Amino acid substitution matrices from protein blocks
Steven Henikoff and Jorja G. Henikoff
Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 10915-10919,
November 1992

Where did the Blosum62 alignment score matrix come from ?
Sean R Eddy
Nature Biotechnology Vol22 Number 8
August 2004