Scoring Amino Acid Substitutions In Φhage Genomes

Promita Bose, R.Edwards, P.Salamon and Hedley Morris *

Abstract—**Substitution matrices are among the most widely used scoring techniques : BLAST, Muscle 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 Φhage-specific scoring matrix based on the abundances of aligned substitutions. These matrices use information from approximately five and a half million similar protein alignments from over five hundred Φhage 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.**

Keywords: ***ungrouped, Φhage, alignment, score, BLAST***

1 Introduction

Substitution Matrices specify a score for aligning each pair of amino acids and these scores are subsequently used for aligning protein sequences.These matrices are uniquely tailored for amino acid pairs with a specific probability distribution. Given a model of protein evolution from which such distributions may be derived, a substitution matrix adapted to detecting relationships at any chosen evolutionary distance can be constructed [13]. The goal is to select a "substitution matrix" best able to distinguish biologically meaningful from chance similarities.

In most cases, the type and nature of the scoring system influence the construction of the alignment. For example, optimal strategies for detecting similarities between DNA protein coding regions differ from those for non-coding regions [17]. These constraints are problem specific and must sometimes be treated in a custom manner.

Among other things the substitution matrix is affected by the alignment quality of the training data and the substitutions generated by the training set. Does one scoring scheme adequately represent all the substitutions in the training data ? Alternate alignment methodologies are well described in the literature [6].

*P.Bose(bosepromita@gmail.com),R.Edwards(redwards@cs.sdsu.edu), P.Salamon (salamon@math.sdsu.edu) are at the Computational Science Research Center, San Diego State University, San Diego, CA 92182-1245,USA and Hedley Morris (hedley.morris@cgu.edu) is with the School of Mathematical Sciences, Claremont Graduate University, CA 91711,USA.

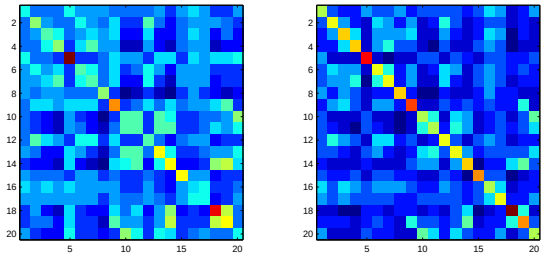

Figure 1: Images of the Φhage and the BLOSUM62 Scoring Matrix

2 Methodology

Many mature scoring systems such as the Blosum and PAM matrices have successfully solved the problem of scoring alignments that only pays attention to the matching database similarities, but none of them need have any biological significance [15].

The BLOSUM approach was introduced by Henikoff and Henikoff [10]. We apply a Blosum approach to a set of Φhage protein sequences from public databases. Unlike most previous work, the protein sequences are not grouped into related families or percent identities. Instead pairwise aligned sequences were generated according to the BLOSUM62 scoring matrix [14]. To capture all possible substitution patterns of the training data, proteins with very low sequence similarity were also included in the training set. These alignments provide the data for building the Φhage Blosum Scoring Matrix.

2.1 What is Φhage Blosum Score

A brief description of the mathematical analysis of the formula is as follows. Assuming a random protein model in which the amino acids occur independently with background probabilities $\vec{p}$, the equation for calculating the score s_{ij} for aligning two residues i and j can be written uniquely in the form

$$s_{ij} = \frac{1}{\lambda}\ln\left(\frac{q_{ij}}{p_i p_j}\right)$$

The numerator q_{ij} is the likelihood of the hypothesis we want to test, that the two residues i and j are correlated

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0.12	-0.42	-0.48	-0.41	0.15	-0.29	-0.25	-0.14	-0.48	-0.43	-0.43	-0.4	-0.32	-0.65	-0.16	0.04	-0.15	-0.78	-0.71	-0.12
R	-0.42	0.76	-0.25	-0.42	-0.37	0.15	-0.12	-0.55	0.08	-0.8	-0.66	0.47	-0.57	-0.93	-0.4	-0.35	-0.34	-0.33	-0.5	-0.69
N	-0.48	-0.25	0.43	0.27	-0.52	-0.14	-0.09	-0.25	0.1	-0.96	-1	-0.05	-0.83	-0.95	-0.54	-0.03	-0.19	-0.98	-0.54	-0.89
D	-0.41	-0.42	0.27	0.59	-0.76	-0.09	0.48	-0.3	-0.12	-1.22	-1.25	-0.16	-1.04	-1.26	-0.38	-0.18	-0.32	-1.24	-0.81	-1
C	0.15	-0.37	-0.52	-0.76	2.86	-0.67	-0.74	-0.52	0	-0.14	-0.15	-0.73	-0.1	0	-0.62	-0.04	-0.19	0.04	0.06	0.15
Q	-0.29	0.15	-0.14	-0.09	-0.67	0.41	0.27	-0.55	0.09	-0.72	-0.54	0.17	-0.33	-0.82	-0.3	-0.24	-0.28	-0.65	-0.61	-0.63
E	-0.25	-0.12	-0.09	0.48	-0.74	0.27	0.48	-0.54	-0.1	-0.85	-0.85	0.15	-0.67	-1.08	-0.3	-0.24	-0.32	-0.99	-0.75	-0.65
G	-0.14	-0.55	-0.25	-0.3	-0.52	-0.55	-0.54	0.69	-0.6	-1.32	-1.26	-0.6	-1.04	-1.3	-0.6	-0.21	-0.61	-0.85	-1.13	-1.02
H	-0.48	0.08	0.1	-0.12	0	0.09	-0.1	-0.6	1.64	-0.66	-0.57	-0.1	-0.45	-0.09	-0.36	-0.32	-0.33	-0.07	0.52	-0.66
I	-0.43	-0.8	-0.96	-1.22	-0.14	-0.72	-0.85	-1.32	-0.66	0.47	0.53	-0.72	0.39	0.15	-0.77	-0.77	-0.37	-0.35	-0.26	0.65
L	-0.43	-0.66	-1	-1.25	-0.15	-0.54	-0.85	-1.26	-0.57	0.53	0.46	-0.69	0.55	0.39	-0.66	-0.8	-0.54	-0.11	-0.12	0.26
K	-0.4	0.47	-0.05	-0.16	-0.73	0.17	0.15	-0.6	-0.1	-0.72	-0.69	0.35	-0.52	-0.95	-0.37	-0.26	-0.24	-0.87	-0.64	-0.67
M	-0.32	-0.57	-0.83	-1.04	-0.1	-0.33	-0.67	-1.04	-0.45	0.39	0.55	-0.52	1.04	0.23	-0.79	-0.59	-0.36	-0.22	-0.25	0.14
F	-0.65	-0.93	-0.95	-1.26	0	-0.82	-1.08	-1.3	-0.09	0.15	0.39	-0.95	0.23	1.07	-0.94	-0.88	-0.69	0.74	0.94	-0.09
P	-0.16	-0.4	-0.54	-0.38	-0.62	-0.3	-0.3	-0.6	-0.36	-0.77	-0.66	-0.37	-0.79	-0.94	1.25	-0.16	-0.22	-1.02	-0.8	-0.59
S	0.04	-0.35	-0.03	-0.18	-0.04	-0.24	-0.24	-0.21	-0.32	-0.77	-0.8	-0.26	-0.59	-0.88	-0.16	0.07	0.17	-0.8	-0.64	-0.57
T	-0.15	-0.34	-0.19	-0.32	-0.19	-0.28	-0.32	-0.61	-0.33	-0.37	-0.54	-0.24	-0.36	-0.69	-0.22	0.17	0.22	-0.84	-0.63	-0.16
W	-0.78	-0.33	-0.98	-1.24	0.04	-0.65	-0.99	-0.85	-0.07	-0.35	-0.11	-0.87	-0.22	0.74	-1.02	-0.8	-0.84	2.26	0.82	-0.52
Y	-0.71	-0.5	-0.54	-0.81	0.06	-0.61	-0.75	-1.13	0.52	-0.26	-0.12	-0.64	-0.25	0.94	-0.8	-0.64	-0.63	0.82	1.23	-0.38
V	-0.12	-0.69	-0.89	-1	0.15	-0.63	-0.65	-1.02	-0.66	0.65	0.26	-0.67	0.14	-0.09	-0.59	-0.57	-0.16	-0.52	-0.38	0.34

Figure 2: Φhage Blosum Scoring Matrix

because they are homologous [9]. Hence q_{ij} are positive numbers that sum to 1 and are known as the target frequencies. $p_i p_j$ is the likelihood of the null hypothesis, that the two residues i and j are uncorrelated and unrelated, occurring independently. λ is a scaling factor that lets us round off all the terms in the Φhage Blosum Scoring Matrix to sensible integers.
The Φhage Blosum Score is positive when $q_{ij} > p_i p_j$ that is the residues i and j aligned together in homologous sequences more often than we expect them to occur by chance. Careful curatorial work yields the training data representing true biological relationships from which the q_{ij} are obtained by counting the frequency at which each residue pair occurs. Rearrangement of the log-odds equation gives

$$q_{ij} = p_i p_j e^{\lambda s_{ij}} \Longrightarrow \sum_{ij} p_i p_j e^{\lambda s_{ij}} = 1$$

The background and the target frequencies are tied by the equality $p_i = \sum_{j=1}^{20} q_{ij}$ so that the background probability distribution is the marginal of the joint target frequencies [18].

Gaps and Observed Patterns

An important consideration regarding our calculation is that it does not formally take gaps into account. This may be acceptable from a biological perspective, since a gap is not a real component of a sequence. Nevertheless the Φhage Blosum Scoring Matrix can be extended, by treating the gap as equivalent to a 21st amino acid. Then pairs of the form $(i, -)$ or $(-, j)$ where the symbol "$-$" represents the gap are also included. Moreover, the correct positioning of a gap in any given alignment is never certain, as it is introduced a posteriori as the product of an alignment algorithm that takes the two sequences X and Y and tries to minimize the deletions, insertions or number of changes that allow to transform X into Y or vice versa. The minimization may involve a heuristic approach or an exact procedure [2]. The consequence of assuming independence is that $q_{-j} = q_{-} q_j$ leads to a null contribution of the corresponding score, so that for

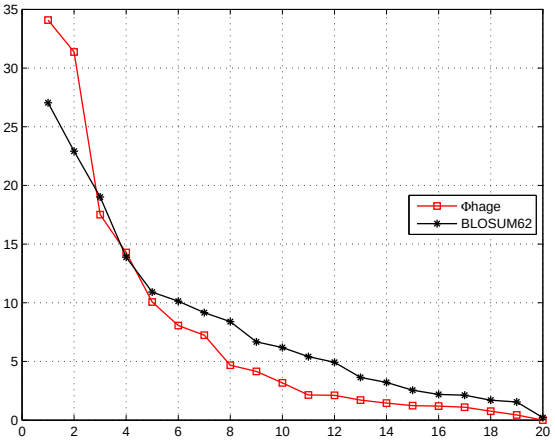

Figure 3: Singular Value Spectra

gapped sequences we simply assign a score equal to zero whenever an amino acid is paired with a gap.

3 Performance of the Φhage Blosum Scoring Matrix

To rate the performance of the custom matrix, an independent set of Φhage protein sequences were aligned [8], referred to as the test set. BLAST [4] was used with the Φhage and BLOSUM62 matrices respectively. This was done to collect all possible types of similar Φhage proteins. Both scoring schemes yielded identical results, thus establishing the ability of the Φhage Blosum Matrix to be at par with published data.

3.1 Comparing Φhage and BLOSUM62

To escape from any existing bias an idea from [7] was adapted to specifically compare the Φhage and BLOSUM62 matrices. Figure 1 indicates that the BLOSUM62 matrix is much more diagonally dominant. The output of the Singular Value Decomposition adheres to this observation which is evinced in Figure 3. This reflects the cease of dominance of amino acids' mutabilities in the respective substitution matrices, and in the mode of sequence conservation in the underlying populations. Although qualitative consistency is preserved, the different modes of sequence conservation imply selective evolutionary pressures are in effect.

3.2 Φhage Blosum and Alignment Tools

The Φhage Blosum Scoring Matrix can be used as a plug-in for sequence alignment tools like ClustalW and Muscle [8]. This was implemented with Muscle to assess the quality of the alignments. The current approach uses linear

```
2 139
Bacillus_O MFRCPACAQA HEMHVVLKNG EPYAVRQNKE MTEVIFHNGI IEINKDHALH
uniprot|Q9 MWMNFESLQI ARAYLF---G EVKYLDLMLV LNIIDIITGV IKAWKFKELR

ESTEKRIKEE TDMEQLIECP ECKEVNTLAN YIEAKVDTLK YHEMNDDQLC HCGGELWMD-
SRSA-WFGYV RKMLSFLVVI VANAIDTIMD LNGVLTFATV LFYIANEGLS ITENLAQIGV

RIPG--TPKY GFVCDKCSWV KPKKVVNGG
KIPAVITDRL HVIESDNDQK TEKDDQAAG
```

Figure 4: Example of a Φhage distinguished protein pair

```
2 148
Bacillus_O ---MFRCPAC AQAHEMHVVL KNGEPYAVRQ NKEMTEVIFH NGIIEINKDH
uniprot|Q9 MWMNFESLQI ARAYLF---- --GEVKYLDL MLVLNIIDII TGVIKAWKFK

ALHESTE--K RIKEETDMEQ LIECPECKEV NTLANYIEAK VDTLKYHEMN D-----DQLC
ELRSRSAWFG YVRKMLSFLV VIVANAIDTI MDLNGVL--T FATVLFYIAN EGLSITENLA

HCGGELWMDR IPG--TPKYG FVCDKCSWVK PKKVVNGG
QIG-----VK IPAVITDRLH VIESDNDQKT EKDDQAAG
```

Figure 5: Same protein pair using BLOSUM

gap penalties so that the alignment with fewer gaps is favored over the alignment with more gaps.

The reduced number of gaps point to the existence of dissimilar substitution patterns between the Φhage proteins and the BLOCKS database [14]. This may be taken into consideration during the process of protein sequence design. Also computation time may be reduced by exploiting data specific attributes.

3.3 Intepreting the Test Data

Alignments of the type displayed in Figure4 and Figure 5 were further processed to corroborate any underlying biological correlation and to test whether the matrices being used were the most appropriate ones for measuring it. This was achieved by applying a standard tool from *Information Theory* [19] known as *Mutual Information*. The *Mutual Information* $I(X,Y)$ between two random variables X and Y

$$I(X,Y) = \sum_{ij} p_{ij} \ln\left(\frac{p_{ij}}{p_i p_j}\right)$$

where p_{ij}, p_i, p_j are, respectively the joint probabilty distribution and the marginals associated to the random variables X and Y [2]. Hence in the context of protein sequences, we interpret p_{ij} as the relative frequency of finding amino acids i and j paired in X and Y. Analogously p_i (p_j) is the relative frequency of finding amino acid i (j) in sequence X (Y). Therefore mutual information may be used to measure the stochastic correlation between two sequences.
$I(X,Y)$ was computed for both groups of alignments. The output is displayed in Figure 6. The disparate results clearly indicates that the Φhage protein model is different from the protein model over which the Blosum matrices was computed. Also it may be safe to state that

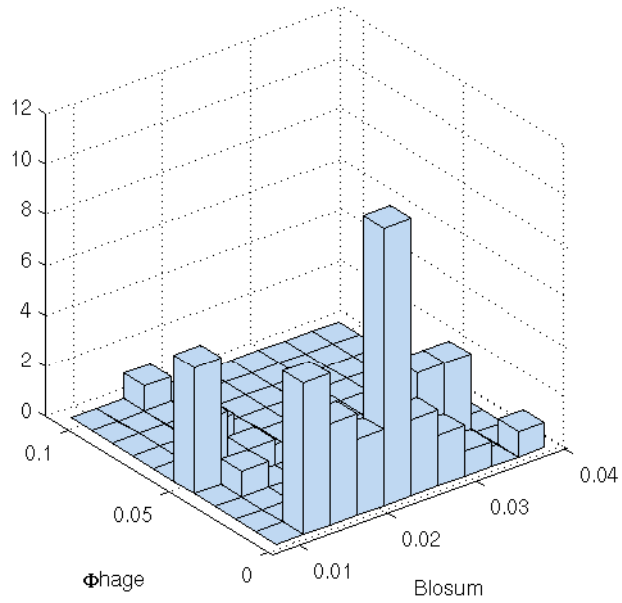

Figure 6: Scoring Sequence Convergence

the average information available for each position, in order to distinguish the alignment from chance is not same in the two groups of alignments. Another reason for the observed difference may be that the selected Blosum matrix is incorrectly matched to the evolutionary distance of the Φhage protein sequences. One other possibility is that some of the Φhage protein sequences may have diverged under a nonstandard evolutionary process. This weak correlation was well captured by the Φhage matrix in that it obtained large values for some alignments for which it's Blosum counterpart was much less. The analysis may be different for gapped alignments, as the choice of gap values exercise an essential influence on the $I(X,Y)$.

4 Conclusions and Future Work

Our goal was to construct and evaluate the Φhage Blosum Scoring Matrix with respect to differentiating the amino acid substitution patterns of Φhage proteins from that of non Φhage proteins and whether an optimal alignment of sequences in the twilight zone could be obtained.
It was found that in most cases the performance of the Φhage Matrix was better than that of the Blosum series in providing added biological information for Φhage proteins. This was evident from the fact that the BLAST output using the Φhage matrix, with the same gap parameter contained a larger number of similar sequences. A topic for future work will be to analyze the statistical behavior of the Φhage scoring scheme as a function of gap penalties. To this end, a database homology search pro-

gram will be implemented using the in house scoring matrix and a variety of other available scoring systems. One possibility may be to evaluate the remote homology detection ability and alignment quality of generalized affine gap costs [5]. Alternatively one may extend and apply the theory introduced in [12].

References

[1] Piotr Pokarowski, Andrzej Kloczkowski, Szymon Nowakowski, Maria Pokarowska, Robert L. Jernigan and Andrzej Kolinski *Ideal amino acid exchange forms for approximating substitution matrices* pp 379-393 69/2007

[2] Francesco Fabris, Andrea Sgarro and Alessandro Tossi *Splitting the BLOSUM Score into Numbers of Biological Significane* EURASIP Journal on Bioinformatics and Systems Biology Volume 2007

[3] I. Mihalek, I. Res and O. Lichtarge *Background frequencies for residue variability estimates : BLOSUM revisited* 8/2007 BMC Bioinformatics

[4] Yi-Kuo Yu and Stephen F. Altschul *The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions* pp 902-911 21/2005 Bioinformatics

[5] Marcus A. Zachariah, Gavin E. Crooks, Stephen R. Holbrook and Steven E. Brenner *A Generalized Affine Gap Model Significantly Improves Protein Sequence Alignment Accuracy* pp 329-338 58/2005 PROTEINS : Structure, Function, and Bioinformatics

[6] J. C. Gelly, L. Chiche, and J. Gracy *"EvDTree : structure-dependent substitution profiles based on decision tree classification of 3D environments* 6/2005. BMC Bioinformatics

[7] Akira R.Kinjo and Ken Nishikawa *Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of the mode of sequence conservation in protiens* pp 2504-2508 20/2004 Bioinformatics Discovery Note

[8] Robert C. Edgar *MUSCLE : a multiple sequence alignment with high accuracy and high throughput* 32/2004 pp 1792-1797 Nucleic Acids Research

[9] Eddy S. *Where did the BLOSUM62 alignment score matrix come from ?* 22/2004 pp. 1035 - 1036 Nature Biotechnology

[10] Warren J. Ewens and Gregory R. Grant *Statistical Methods in Bioinformatics An Introduction* Springer 2001 Berlin

[11] Altschul SF, Mark S. Boguski, Warren Gish and John C. Wootton *Issues in searching molecular sequence databases*, 6/1994 Nature

[12] Stephen F. Altschul *Generalized Affine Gap Costs for Protein Sequence Alignment* pp 88-96 32/1998 PROTEINS : Structure, Function and Bioinformatics

[13] Altschul S. *A Protein Alignment Scoring System Sensitive at All Evolutionary Distances* pp 290-300 36/1993 J Mol Evol

[14] Henikoff S. and Henikoff J *Amino acid substitution matrices from protein blocks* pp 10915-10919 89/1992 Proc. Natl. Acad. Sci. USA

[15] Steven Henikoff and Jorja G. Henikoff *Performance Evaluation of Amino Acid Substitution Matrices* pp 49-61 17/1993 PROTEINS : Structure, Function and Genetics

[16] David T. Jones, William R. Taylor, and Janet M. Thornton *The rapid generation of mutation data matrices from protien sequences* pp 275-282 8/1992 CABIOS

[17] States D J, Warren Gish and Stephen F Altschul *Improved sensitivity of nucleic acid database searches using application specific scoring matrices* pp 66-70 3/1991 A Companion to Methods in Enzymology

[18] W. Feller *An Introduction to Probability and Its Applications* John Wiley & Sons, NY USA 1968

[19] C.E. Shannon *A mathematical theory of communication - part 1* pp 623-656 27/1948 Bell System Technical Journal