Optimization of Substitution Matrix for Sequence Alignment of Major Capsid Proteins of Human Herpes Simplex Virus

Vipan Kumar Sohpal[1*], **Amarpal Singh**[2], **Apurba Dey**[3]

[1]*Department of Chemical & Bio Technology*
Beant College of Engineering & Technology
Gurdaspur, 143521 Punjab, India
E-mail: Vipan752002@gmail.com

[2]*Department of Electronics & Communication Engineering*
Beant College of Engineering & Technology
Gurdaspur, 143521 Punjab, India

[3]*Department of Bio Technology*
National Institute of Technology
Durgapur, West Bengal, India

[*]*Corresponding author*

Received: June 02, 2011 **Accepted: December 27, 2011**

Published: January 31, 2012

Abstract: *Protein sequence alignment has become an informative tool in modern molecular biology research. A number of substitution matrices have been readily available for sequence alignments, but it is challenging task to compute optimal matrices for alignment accuracy. Here, we used the parameter optimization procedure to select the optimal Q of substitution matrices for major viral capsid protein of human herpes simplex virus. Results predict that Blosum matrix is most accurate on alignment benchmarks, and Blosum 60 provides the optimal Q in all substitution matrices. PAM 200 matrices results slightly below than Blosum 60, while VTML matrices are intermediate of PAM and VT matrices under dynamic programming.*

Keywords: *Human herpes simplex virus, Sequence alignment, Blosum, Parameter optimization procedure.*

Introduction

Sequence alignment is one of the important tools used in bioinformatics and computational biology. Optimization of substitution matrices using mathematical and bioinformatics software is next issue in field of sequence alignment and phylogenetic analysis. A number of alignment algorithms have been developed for sequence alignment on the basis of dynamic programming (N-W, S-W) and heuristics approach. Most of these algorithmic formulations and designed on the problem seek an alignment maximizing, a function known as the objective score. Objective scores are a sum of terms for matching pairs of letters and penalties for gaps. In the most of classic approach to alignment, scoring matrix and gap costs are parameters which determine alignment scores. Classic as well as heuristic techniques have their limitation. Later approach on the basis of an ad-hoc algorithm without strong foundation and former having cut off score.

Previous work in this area has included Clustal W [5, 6] is one of the best sequence alignment tools based on progressive approach. The main problem of Clustal W is that the initial pairwise alignments are fix, and early errors cannot be corrected later, even if those alignments conflict with sequences added later [11]. T-Coffee is another popular sequence alignment tool and can be viewed as a variant of the progressive method. They report to getting the highest scores on BAliBASE bench mark database [10]. The significant improvement achieve by pre-processing a data set of all pair-wise alignments and thus allowing for much better use of information in early stages. Moreover, the different commercial software tool and their availability have been documented in literature for sequence alignment with conventional dynamic programming [13, 14]. Parameter optimization procedure mathematically described for the problems of optimizing gap penalties and selecting substitution matrices for protein alignment [2]. Surprisingly, we could find only two publications that attempts to work on protein UL-56 and UL-11 of human herpes virus, which cause infection associated with arterial myxoma [7, 16]. Most importantly, there are no rigorous assessment of scoring schemes of protein sequence and their utilization for drug design.

In this paper, we used the two-parameter global alignment method for optimization substitution matrices. The objective for this work is optimization of substitution matrices using combination techniques by varying extension penalty with fixed open gap for of major viral capsid protein of human herpes virus (HHV-1/HHV-2). This paper has unique approach to use POP algorithm for optimization of substitution matrices using muscle tool for protein causing cerebral disease.

Method

The best alignment, of two amino acid sequences has become almost the standard technique for sequence comparison in molecular biology. It uses to find similarity between two sequences and, evolutionary history between species, to find consensus sequences, and other significant functions. There are numbers of papers on this topic and its applications to biology [3, 12, 17]. There is no unique method mentioned for the parameters such as to gap penalties and, extension.

We used Parameter optimization procedure for optimization of substitution matrices in three steps. In the first step, an N-dimensional hyper cuboid is explored by evaluating $Q(w)$ at each point in a regularly spaced grid. Local maxima is target as points at which Q is greater than all neighboring points, and the best of these are use as starting points for the second stage.

Let $w = w_i$, $i = 1, ..., N$ be the parameters of interest (e.g., gap open and extend penalties), and $Q(w)$ be the function to be optimized (an alignment accuracy). The goal is to find values $wOPT$ = argmax$(w)Q(w)$ that maximize Q. It is typically expensive to compute, being requiring enormous time to evaluate at a single point, and is non-convex with many local maximums (Fig. 1).

The second and third steps use a hill-climbing strategy to approach a local maximum given a starting point. The final result is the best maximum found by the third step. The three steps are use for large subsets of the training samples, with the purpose to save the computational time. First two steps in POP method using randomly chosen subsets and the third steps the entire training set [2]. Then computational analyses of sequence using Heuristic methods, as MUSCLE or exact approaches based on dynamic programming, such as the Smith-Waterman algorithm are use as tools to analyze protein sequences. In this work, the substitution matrix is regarded as fixed while open gap penalty and extension penalty parameters are included. If a gap includes the first or last column of an alignment it is described as terminal, rest of all other gaps are internal. The objective score is than the sum of substitution scores minus gap penalties and a maximum-scoring global alignment are found using standard dynamic programming techniques.

These substitution matrix types are considered: BLOSUM [4], PAM [1] and VT/VTML [8, 9]. Each matrix family is a series with members defined by a measure of evolutionary distance: percent identity cutoff in the case of BLOSUM, PAM distance for the rest. Conventionally, integer valued matrices are used in which log-odds scores in fractional bits have been rounded to one or two significant figures.

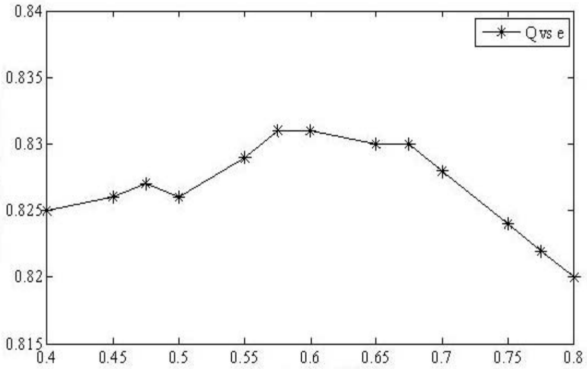

Fig. 1 Alignment accuracy (Q) vs. e.
The graphs show alignment accuracy on all pairs in as a function of extension penalty e with fixed gap-open penalty g = 7.0 and the BLOSUM 60.

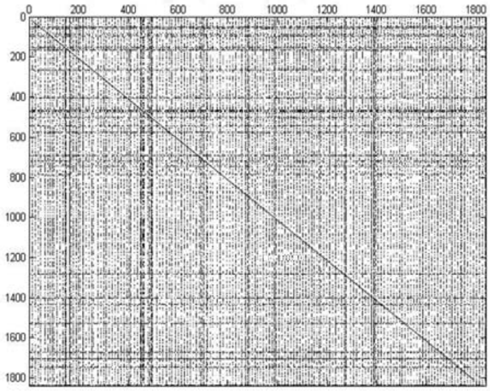

Fig. 2 Sequence dot plot.
The graphs show similarity between pairs of major capsid protein of HHV-1 (x-axis) and HHV-2 (y-axis).

Reference data access from two protein alignment benchmarks. The proteomic data for sequence alignment of major capsid protein of herpes virus were obtained from the National Centre of Biotechnology Information (NCBI) and protein data base bank of Uniprot KB. Data accessed from NP_044620.1 and Uniprot KB/Swiss-Prot P06491 and NP_044488.1 and Uniprot KB P89442 for HHV-1 and HHV-2 respectively. There are 1789 pair-wise reference alignments by Muscle using MEGA [15] and optimal Q calculated using align tool of Emboss. Evolutionary distance measure, by of four subsets of HHV proteins were constructed 1600 randomly selected pairs, with identities in the range 0-25% (Id0_25), 25-50% (Id25_50), 50-75% (Id50_75) and 75-99% (Id75_99) respectively through Blastp. The result obtained using randomly chosen subsets and the entire training set shown in Fig. 3 and Fig. 4. The matrices equivalence is require plotting the value on 2 dimensional graphs which is shown in Table 1.

Table 1. Matrices equivalence used to plot Blosum/PAM substitution matrices versus optimal Q

Blosum	30	40	50	60	70	80	90
PAM	350	300	250	200	150	100	50

Results

(a) One of the simplest methods for evaluating similarity between two sequences is to visualize regions of similarity using dot plots. So a straightforward dot, plot constructed for major capsid protein (UL-19) of HHV-1 on the horizontal axis of plot space and data of sequence HHV-2 assigned to the vertical axis. This dot plot for two protein sequences that share extensive regions of similarity (Fig. 2). From the result, of dot plot higher similarity shown in Fig. 2, which pushes the research toward protein alignment accuracy, optimum substitution matrix, and gap penalties vary with sequence identity.

(b) Fig. 3(a) shows the results of optimizing Model 1 having extension penalty (*e* = 0.6) with fixed gap-open penalty (*g* = 7.0) on the Molecular Evolutionary Genetic Analysis (MEGA) sets using inbuilt Muscle alignment. Similarly, Model 2 has extension penalty (*e* = 0.4) with fixed gap-open penalty (*g* = 7.0) aligned as Model 1 and observed result shown in Fig. 3(b). The results indicated in Fig. 3(a) and 3(b) are qualitatively similar on the two sets, giving confidence that they indicate general trends rather than artifacts to benchmark construction, overtraining significantly suboptimal local maxima. This is further confirmed by cross-training (Fig. 3(c) and 3(d)), which again gives similar, results, as would be expected. The results show Blosum to be the best matrix series, with BLOSUM > PAM > VTML > VT holding for most members though the differences between VTML, VT and BLOSUM are small except at the extreme high- and low-distance ends of the series. The PAM series is a minor inferior to Blosum, giving accuracy scores around 1% lower.

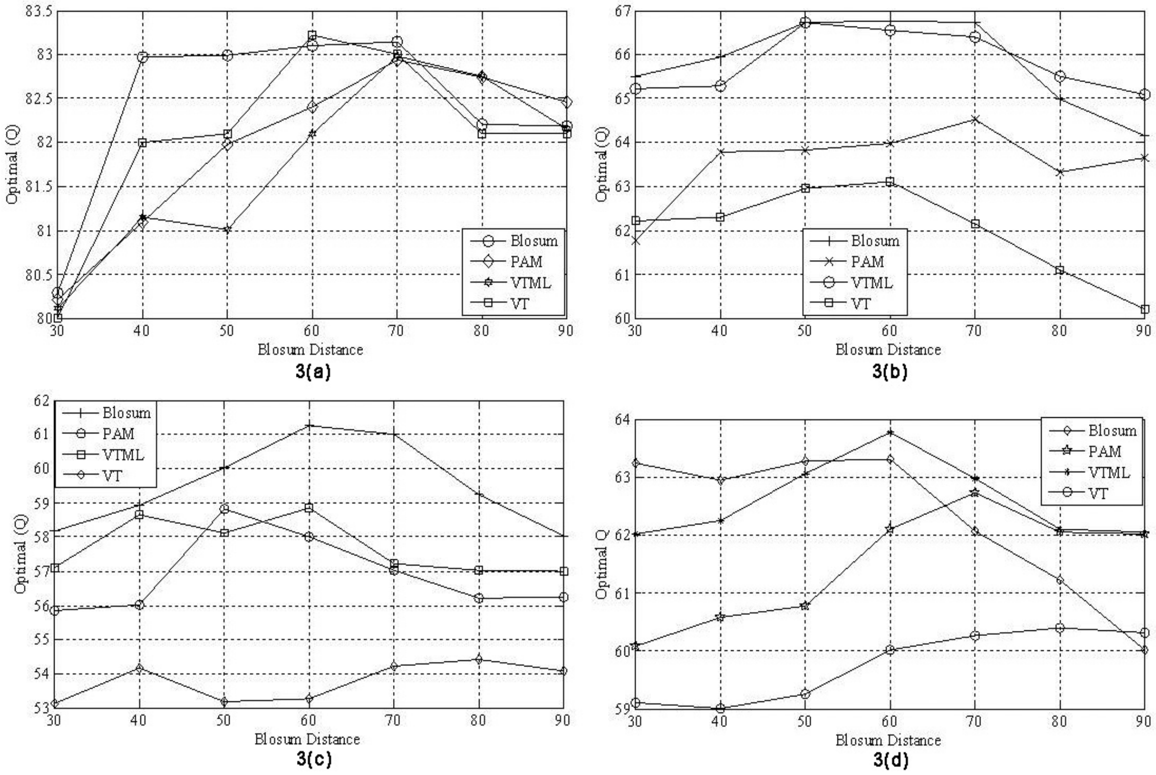

Fig. 3 (a) and (b) are graphs show matrix accuracy *Q* (vertical axis) in percentage vs. substitution matrices Blosum on horizontal axis for Model 1 and Model 2.
(c) and (d) show cross trained results for optimization of Model 1 and Model 2 *Q* (vertical axis) in percentage vs. substitution matrices Blosum on horizontal axis.

(c) Previous works revel that different substitution matrix more effective at different evolutionary distances rate. To investigate this, we optimized Model 1 on the Id0_25, Id25_50, Id50_75 and Id75_99 sub sets, with the results shown in Fig. 4. Interestingly, the plots are qualitatively similar for the four sets despite increasing pair-wise identity and the increasingly narrow variation in optimal *Q*. The lowest alignment accuracy observed in Subset 1 of Id0_25 (Fig. 4 a) while optimal score for major capsid protein in Subset 2 and Subset 3 are intermediate. Remarkably, in the case of Id75_99 accuracies are all above 93.40% and the difference between the best matrices score is 2.30%. The peak in each curve that identifies the best matrix in each family is at approximately the same evolutionary distance on each set, showing that the best choice of the matrix is almost independent of sequence divergence. The results shown in Fig. 4(a), 4(b), 4(c) and 4(d) indicate Blosum to be the best matrix series. Fig. 3 represents the optimal alignment score in a similar pattern as observed in Fig. 4.

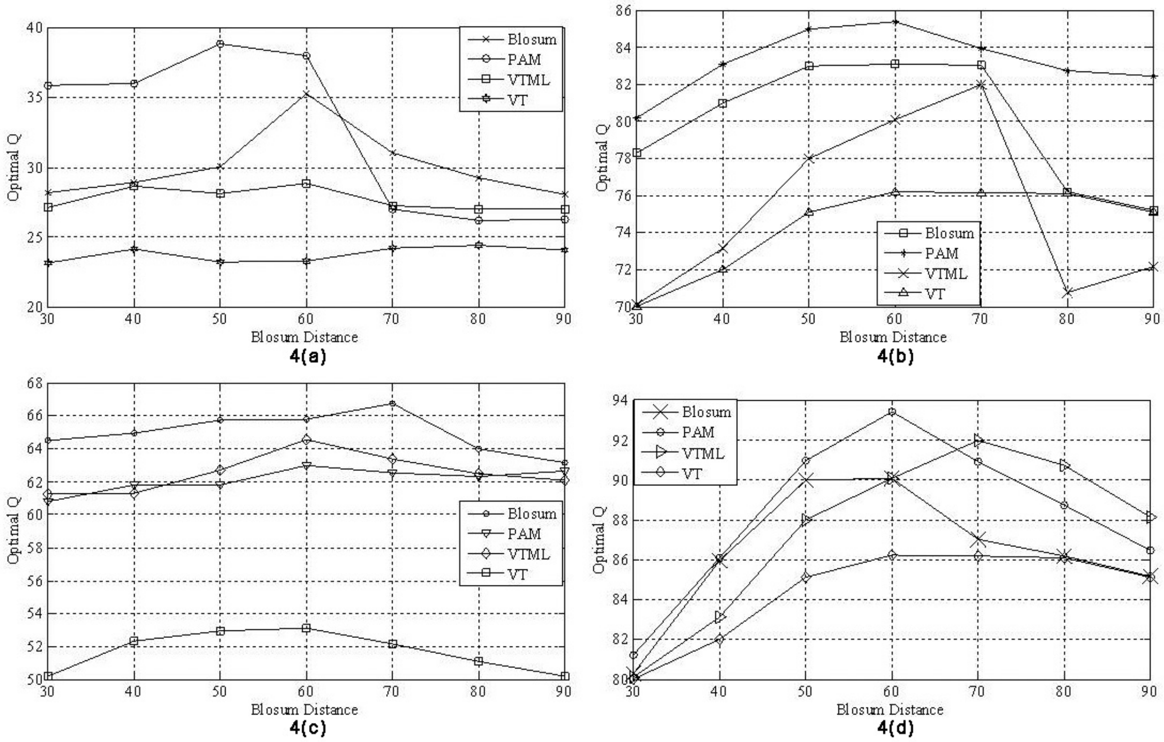

Fig. 4 Results of four subsets by sequence identity, for *Q* (vertical axis) in percentage vs. substitution matrices Blosum on horizontal axis: (a) represents subset Id0_25, (b) depicts the subset Id50_75, (c) reflects subset Id25_50 and (d) shows the subset having Id75_99

(d) In all the tests (matrix accuracy, cross trained and subset alignment) reported in paper, Blosum 60 is the best or close to the best choice and is recommended as the alignment of major capsid protein of human herpes virus. In the PAM and VTML series, PAM 200 and VTML150 respectively are recommended. It indicating that VT series is inferior as compare to other substitution matrices.

Discussion

POP used for optimization of substitution matrices for major capsid protein of human herpes virus, but it can be used for any protein sequence data. POP preliminary requires an understanding of the input data and some trial and error. However, consistency of trends using

Muscle through MEGA benchmark and the differences in optimal score vs. substitution matrices are remarkably small. This consistency also observed when run cross trained and subsets of data. It suggests that parameter optimization procedure may find a global optimum in the pair-wise global alignment tests considered here. From results shown in Fig. 3 and Fig. 4, Blosum matrices proved to give the best accuracy, in alignment. The PAM series were not far behind. VTML series was marginally better but poor than PAM series.

We have made several alignments combinations. First of all, in the pairwise alignment case as shown in Fig. 3(a) for model1 has optimal $Q \approx 83\%$. The range of optimal Q varied ≈ 80-83% for all substitution matrices. With an increasing extension penalty, especially for Model 2, the quantitative results change tremendously Model 2 having Q optimal from 61.75-66.75% in all scoring matrices indicated in Fig. 3(b). Cross trained sequence yielding optimal Q average. Alignment programs have scores around ≈ 53-65%. This is quite remarkable, that PAM show a similar performance as Blosum in terms of scoring.

However, computes Q for pair wise alignments of four subsets (Id0_25, Id25_50, Id50_75 and Id75_99) of major capsid protein constructed from base pairs and results of alignments that are highly consistent with all subset alignments shown in Fig. 4. As we move from subset1 to Subset 4 (lower similarity to higher similarity percentage) then POP generates better alignments accuracy, Q score. Another astonishing observation is that alignment accuracy a purely sequence-based results. The Fig. 4(c) and Fig. 4(d) instances show a comparable performance for instances above ≈ 72%, with a growing number of input instances the optimal Q becomes even better.

Conclusion

The analysis of this work suggests that the Blosum 60 matrix is recommending as the most appropriate for major capsid protein for global alignment accuracy with extension penalty 0.6. Overall results analysis shows that BLOSUM series of matrices has higher alignment accuracy as compare to other substitution matrices. Model 1 reflects the better optimal Q with reference to Model 2. Subset Id75_99 has alignment accuracy 8% higher than Model 1.

References

1. Dayhoff M. O. (Ed.), L. T. Hunt, W. C. Barker, R. M. Schwartz, B. C. Orcutt, C. L. Young (1978). Atlas of Protein Sequence and Structure, 5(3), 470-496.
2. Edger R. C. (2009). Optimizing Substitution Matrix Choice and Gap Parameters for Sequence Alignment, BMC Bioinformatics, 10, 396-405.
3. Fitch W. M., T. F. Smith (1983). Optimal Sequence Alignments, Proc Nat Acad Sci USA, 80, 1382-1386.
4. Henikoff S., J. G. Henikoff (1992). Amino Acid Substitution Matrices from Protein Blocks, Proc Natl Acad Sci USA, 89, 10915-10919.
5. Higgins D., P. Sharp (1988). CLUSTAL: A Package for Performing Multiple Sequence Alignment on a Microcomputer, Gene, 73, 237-244.
6. Hompson J., D. Higgins, T. Gibson (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting Position Specific Gap Penalties and Weight Matrix Choice, Nucleic Acids Research, 22(22), 4673-4690.
7. Li Y., Z. Pan, Y. Ji, M. Sheppard, D. J. Jeffries, L. C. Archard, H. Zhang (2003). Herpes Simplex Virus Type 1 Infection Associated with Artial Myxoma, American Journal of Pathology, 163(6), 2407-2412.
8. Muller T., M. Vingron (2000). Modeling Amino Acid Replacement, Journal of Computational Biology, 7(6), 761-776.

9. Muller T., R. Spang, M. Vingron (2002). Estimating Amino Acid Substitution Models: A Comparison of Dayhoff's Estimator, the Resolvent Approach and a Maximum Likelihood Method, Mol Biology Evolution, (19), 8-13.
10. Notredame C. (2002) Recent Progress in Multiple Sequence Alignment: A Survey, Pharmacogenomics, 3(1), 131-144.
11. Notredame C., D. Higgins, J. Heringa (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment, Journal of Molecular Biology, 302, 205-217.
12. Pearson W. R., D. J. Lipman (1988). Improved Tools for Biological Sequence Compare, Proc Nat Acad Sci USA, 85, 2444-2448.
13. Sohpal V. K., A. Dey, A. Singh (2010). MEGA Biocenteric Software for Sequence & Phylogenetic Analysis: A Review, Int J Bioinformatics Research and Applications, 6(3), 230-240.
14. Sohpal V. K., A. Dey, A. Singh (2010). Sequence Alignment and Phylogenetic Analysis of Human Herpes Simplex Virus using Bioinformatics Tool: A Review, Int Journal of Computational Biology & Drug Design, 3(1), 68-88.
15. Tamura K., J. Dudley, M. Nei, S. Kumar (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0, Molecular Biology and Evolution, 24, 1596-1599.
16. Ushijima Y., T. Koshizuka, F. Goshima, H. Kimura, Y. Nishiyama (2008). Herpes Simplex Virus Type 2 UL56 Interacts with the Ubiquitin Ligase Nedd4 and Increases its Ubiquitination, Journal of Virology, 82(11), 5220-5233.
17. Waterman M. S. (1989). Sequence Alignments, in Mathematical Methods for DNA Sequences, CRC Press, Boca, 53-92.

Vipan Kumar Sohpal
E-mail: vipan752002@gmail.com

Vipan Kumar Sohpal received his B.Tech. in Chemical and Biotechnology Engineering from PTU Jalandhar and his ME from NIT Jalandhar in 1999 and 2006, respectively. Currently, he is an Assistant Professor in the Department of Chemical Engineering, Beant College of Engineering and Technology, Gurdaspur, Punjab, India. His present research interests are bioinformatics and Biofuel. Even though he is a young person he already has published around 20 papers.

Amarpal Singh, Ph.D.
E-mail: s_amarpal@yahoo.com

Amarpal Singh received his Ph.D. in Electronics and communication Engineering in 2008. He is a full time Associate Professor of Electronics & Communication Engineering. His present research interests are optical and wireless communication systems and networks and fuzzy logic. He has published more than 50 papers in international and national journals and conferences. He is a reviewer of many international journals.

Apurba Dey, Ph.D.
E-mail: apurbadey2007@yahoo.com

Apurba Dey received his B.Tech. and M.Tech. from Jadavpur University and Ph.D. from IIT Delhi. Currently, he is a Professor in the Department of Biotechnology, National Institute of Technology, and Durgapur West Bengal, India. His present research interests are biochemical engineering, environmental biotechnology and plant tissue culture. He has published more than 30 papers in international, national journals and conferences.