IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VIII (May-Jun. 2014), PP 56-61
www.iosrjournals.org

Validation Experimentations of Local Alignment Parameters for Comparing DNA and Protein Sequences

Manhal Elfadil Eltayeeb Elnour[1], Muhammad Shafie Abd Latif[2], Ismail Fauzi Isnin[3]
[1,2,3]*Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia*

Abstract : *A basic issue in aligning DNA and protein sequences is to find similar characters between two or more sequences in order to detect relations between newly defined sequences and well-known sequences stored in genetic databanks. Local Sequences Alignment (LSA) algorithms have been developed to reveal similar regions between compared sequences. LSA algorithms produced optimal alignment using similarity matrix with scoring scheme. Match, mismatch, and gap penalties are identified using substation matrix and affine gap function. The accuracy of the results relies on selecting the best values for these parameters. This paper mainly set out to validate statically parameters for calculating possible alignments.Estimated values are tested using real dataset and the optimal alignment was recorded as well as the parameters. Perfect symmetric results were obtained, when comparing mathematically and statically estimation for the LSA parameters.*
Keywords: *Local Sequences Alignment, DNA, Protein, Substation Matrix, Scoring Scheme.*

I. Introduction

Sequence alignment is a process of finding matching between DNA and Protein sequences in character-to-character level. The main role of sequences comparisons algorithms is to detect homologous regions between compared sequences. Comparing sequences may reveal or predict functional, structural, and evolutionary analogies.Sequence alignment algorithms are classified into global if the sequences compared as a whole. Conversely, a local sequences alignment is to detect regions of similarity between compared sequences. In local sequence alignment, a number of similar subsequent can be produced. Furthermore, sequences in different species may be similar over short conserved regions and dissimilar over remaining regions. A pairwise alignment is dedicatedto aligning two sequences. While, multiple sequence alignments considering more than two sequences. Dynamic Programming (DP) methods [1, 2]are used in aligning global and local sequences, respectively. These algorithms produce an accurate optimal alignment with high computations and memory costs when comparing long sequences.

In order to transform one sequence into another using DP methods, a series of mutation events are considered, which include match, mismatch, insert and delete (indel). These mutation events are calculated using a substations matrix and a scoring scheme, where each pair of residues can be scored, and the similarity score is the sum of the scores of the individual residues. The optimal alignment is then considered with the maximum similarity score for compared sequences.

In substation matrix, sequences compared on a character-to-character level based on scoring records to reveal biologic relevance and evolutionary distance between compared sequences [3, 4]. Scoring record in the substitution matrix depends significantly on the character, whether nucleotide or amino acid. Usually, the number of observed substitutions is less than the real number of substitutions occurring in evolutionary distance, because for a long period the possibility of second substitution increases. To resolve these problems, several attempts have been made to improve substitution matrix, which effect on the scoring systems up-to-date.

1. Nucleotide Substitution Matrix Model:

From biological scenery nucleotides tend to have the same rate of mutation[5], thereby simple substitution scheme adopting such as estimating one for matching residues and zero for a mismatch or gap residues. Numerous statically and mathematical experiments have established to calculate nucleotide differences by considering probability matrices. Markov chain based methods [6], and maximum likelihood methods [7] deduce the substitution probabilities, which lead to the most probable model of evolutionary distance.

2. Amino Acid Substitution Matrix Models

In contrast to nucleotides, different amino acids mutate at different rates [5]. Biologist elucidate that the number of mutation steps needed to transform amino acids ranges from one to three. Therefore, different mutation rates are obtained among amino acids because of the variation in mutational distance. Furthermore, the diversity of the physicochemical properties constitute in the conformation of the protein structure in different

degrees. Amino acid substitution matrixes reflect the fact of different mutation rates for amino acid and record scores for all possible changes. PAM and BLOSUM series are two popular methods for amino acid substitution matrixes until today.

Dayhoff and Schwartz [8], devised Percent Accepted Mutations (PAM), substitution matrix for aligned amino acids to detect regions of similarity in two sequences. PAM based on statistic model consequent from observation of a percent of evolutionary change in a physicochemical of amino acids. One of the assumptions in PAM matrix is that: similar structure and chemical proprieties in any two amino acids involve evolutionary alteration at higher extent between them, while the less similar amino acid alteration could be neglecting. The percent of the evolutionary alteration of amino acids constituted in building a substitution matrix for aligned amino acids. All PAM series was then discovered based on different percent of evolutionary alteration of amino acids, for instance PAM1, PAM30, PAM250.

Henikoff and Henikoff [9], devise BLOcks SUbstitution Matrices (BLOSUM) series. The notion based on observation for any couple of distantly related sequences tendency toward having a highly protected region or blocks occurred by less conserved expansions of sequences. Probability values used in building a substitution matrix for BLOSUM based on physicochemical proprieties for two residues. In BLOSUM, the values of a similarity percent attached to the name of the matrix such as BLOSUM30, BLOSUM62, and BLOSUM80.
PAM and BLOSUM share some characteristics, such as both of them 20 × 20 matrixes. Furthermore, in both models similar residues are considered with high values while replacements considered with lower values. Difference between the two models appears from their formation, where PAM model based on observations of closely related sequences while BLOSUM model based on observations of alignments of distantly related sequences. BLOSUM model tends to output better findings than PAM model [10], especially in aligning distantly related sequences.

The rest of this paper is organized as follows. Section 2 introduces the parameters in local sequences alignment. Mathematically and statically models are addressed and detailed. In section 3, estimated values are tested, and the results were discussed and validated. While, section 4, conclude the paper and highlight the major contributions.

II. Parameters for Local Sequences Alignment

In local sequences alignment, the alignment scores *s(i,j)* is consider as the score of the aligning the subsequence $(A_1,..,A_{i-1})$ and $(B_1,..,B_{j-1})$ in accordance with the substitution matrix $S(A_i,B_j)$. While, on other cases the gap open penalty deducted from the score of the subsequence alignment. These processes can be defined more formally by the following equation:

$$s(i,j) = max\begin{cases} s(i-1,j-1) + S(Ai,Bj) & the\ first\ case \\ s(i-1,j) - g & the\ second\ case \\ s(i,j-1) - g & the\ third\ case \\ 0 & Otherwise \end{cases} \quad \textbf{(1)}$$

The adjacent cells in the matrix *s(i,j)* with the recursive relation and linear gap penalty g, appear in equation (1). However, for implementing local sequences alignment using affine gap model new equations defined and adjusts with affine gap function μ (r). An improvement in equation (1) is added in (2), (3), (4), and (5). A gap-opening penalty ∂ and a gap extension penalty α defined. Furthermore, the alignment preceded by a gap of length $r = 1, \ldots, i$, which represent an insertion in the first and second sequence respectively.

$$s(i,j) = max\begin{cases} s(i-1,j-1) + S(Ai,Bj) \\ s(i-1,j) - \mu(\mathrm{r}) \quad r = 1,\ldots,i \\ s(i,j-1) - \mu(\mathrm{r}) \ \ r = 1,\ldots,j \\ 0 \end{cases} \quad \textbf{(2)}$$

$$s(i,j) = max\begin{cases} s(i-1,j-1) + S(Ai,Bj) \\ v(i,j) \\ w(i,j) \\ 0 \end{cases} \quad \textbf{(3)}$$

$$v(i,j) = max\begin{cases} s(i-1,j) - \partial \\ \\ s(i-1,j) - \alpha \end{cases} \quad \textbf{(4)}$$

$$w(i,j) = max\begin{cases} s(i,j-1) - \partial \\ \\ s(i,j-1) - \alpha \end{cases} \quad \textbf{(5)}$$

In order to calculate gaps according to equation (2) two matrixes introduce in (4) and (5), in addition to the similarity matrix (3). These additional matrixes used to compute the cost of a set of gaps in the two

sequences. A major drawback with affine gap model is that the alignment of sequences requires $O(MN^2)$ in running time and $O(MN)$ in memory space [3], M and N are lengths of any two sequence. Gotoh [11], reduced running time for implementing sequences alignment using affine gap model to $O(MN)$, this valuable study make using affine gap model more convenient and with biological perspective.

1. Mathematical Induction for Local Sequences Alignment:

The most well-known algorithm used in local sequences alignment is Smith and Waterman (SW) algorithm [2]. SW algorithm is controlled by scoring scheme including matches, mismatch, gaps (insert and delete), and substitution matrix. These parameters have several values stemmed by different alignments. The algorithm relies on setting these parameters, which extremely influenced the quality of the resulting alignments [12]. Local sequences alignment based on SW algorithm always considered positive values in building the similarity matrix; however, negative values are rejected and replaced by zero. Parameters are usually adjusts, but not necessary to match $\delta > 0$, mismatch$\varepsilon < 0$, and a gap penalty $\lambda < 0$. Consider two sequences $(A_1,..,A_i)$ and $(B_1,..,B_j)$, the local sequence alignment is defined as follow [13]

$$S_{i,j} = max\begin{cases} S_{i-1,j-1} + \delta & if\ A_i = B_j \\ S_{i-1,j-1} - \varepsilon & if\ A_i \neq B_j \\ S_{i,j-1} - \lambda \\ S_{i-1,1} - \lambda \\ 0 \end{cases} \quad \textbf{(6)}$$

SW algorithm is initialized by $S_{0i} = S_{j0} = 0$for $0 < i < n$, $0 < j < m$. The score of the maximum alignment is calculated by numbers of match, mismatch, and gap columns multiplied by corresponding parameters values as follow

$$\delta \times No.\{Matches\} - \varepsilon \times No.\{Mismatches\} - \lambda \times No.\{Gaps\} \quad \textbf{(7)}$$

An optimal alignment of A and B is one that maximizes equation (7)

$$S = S(A,B) = Max_{i,j} S_{i,j} \quad \textbf{(8)}$$

SW algorithm can be described as identifying relationship between two sequences with diverged of random scoring scheme elements [14]. This relationship appears in calculating alignment scores for two sequences A_i and B_j with linear gap, match, and mismatch, as the following

$$D = W\delta + U\varepsilon + V\lambda \quad \textbf{(9)}$$

Where $D$is the alignment score, $\delta, \varepsilon, and\ \lambda$ are scores of the match, mismatch, and gap, respectively, while W, U, and V are numbers of columns of the match, mismatch, and a gap, respectively. It is clearly identified that D depends on three parameters$\delta, \varepsilon, and\ \lambda$. However, based on (8) there is a unique alignment score D'_{max}, which represents the maximum score of the alignment and depends only on one parameter defined by $\delta, \varepsilon, and\ \lambda$[15]. The total length of the two sequences calculates as follows

$$2W + 2U + V = L_A + L_B \quad \textbf{(10)}$$

Where, L_A and L_B are the sequence lengths. By compensating values from the equations (9) and (10) then D'_{max} written as follows

$$D'_{max} = \frac{D - \frac{\delta}{2}(L_A + L_B)}{\delta - \varepsilon} = -U - V\eta \quad \textbf{(11)}$$

$$\eta = \frac{\delta/_2 - \lambda}{\delta - \varepsilon} \quad \textbf{(12)}$$

The maximum alignment $D'_{max}$is determined byη, which may efficiently use to set an effective scoring scheme. When$\eta > \frac{1}{2}$,$\varepsilon > 2\lambda$ then D'_{max} minimumgapped as mismatch over the gap. When $0 < \eta < \frac{1}{2}$,$\varepsilon < 2\lambda < \delta$, then D'_{max}score higher gap region than mismatches. When$\eta \leq 0, 2\lambda > \delta$, then D'_{max}score higher gap region than matches, which is unreasonable case.

2. Statically Estimation for Local Sequences Alignment:

The possibility of optimal alignment using scoring scheme can be calculated statically using a likelihood values, where a higher likelihood is of great confidence. A maximum likelihood alignment [16] is based on the evolutionary model with insert, delete, and substitutions events can be considered in the sequence alignment problem. The probability of nucleotide to change from the current state into another state is a Hidden Markov Model process (HMM) of the evolutionary model. HMM, is a stochastic model representing the properties of real data using statistic model. It mainly used to predict and simulate the alignment parameters statically [17]. A pair of sequences M and N are derived from a common ancestral sequence Q by a time g as follows:

$$P_g(M,N) = \sum_C P_\infty(Q)\, P_g(M \backslash Q) P_g(N \backslash Q) \quad \textbf{(13)}$$

Where, $P_\infty(Q)$ is the probabilities of sequence Q, and $P_g(M \backslash Q)$ and $P_g(N \backslash Q)$ are the probabilities of changing state from sequence Q to sequence M and N, respectively. The statically representation of the

substitutions model of nucleotides to change state from *i*to *j* in a time *t* with *s* rate ofsubstitution is calculated as follows:

$$f_{ij}(t) = \begin{cases} e^{-st} + \pi_j(1 - e^{-st}) & if\ i = j \\ \pi_j(1 - e^{-st}) & if\ i \neq j \end{cases} \quad \textbf{(14)}$$

In searching for optimal alignment, the sub-sequence length is differing in sizes according to the chance of occurrence. The probability $\lambda_n$of sequence *Q* to spilt in *n* subsequences is written as follows:

$$\lambda_n = \left(1 - \frac{\eta}{\mu}\right)\left(\frac{\eta}{\mu}\right)^n \quad \textbf{(15)}$$

The calculation of the likelihood requires the calculation of changing probabilities from ancestral to descendant sequences, as well as calculation of the probability of the ancestral sequence. Thus, the probability of changing path $P(\psi\backslash\delta)$ by the alignment ψ using δ parameters is written as follows

$$P(\psi\backslash\delta) = P(\psi,\ \psi'\backslash\delta) = P(\psi\backslash\psi',\ \delta)P(\psi'\backslash\delta) \quad \textbf{(16)}$$

In local sequence alignment, to predict SW parameters, sequences are modeled using HMM [15, 18]. Statically model of the SW parameters determines the significance of the alignment. Furthermore, the model is used to prove the reliability of the alignment and to describe alternative sub-optimal alignments. On the proposed modeling, each sequence is represented by path through the model with two states coupled and uncoupled, see Fig.1. Coupled state occurred in ancestral sequences. However, uncoupled state is described independent residues and nucleotide. Transition from coupled to uncoupled state occurs with probability P_G. While, in uncoupled state residues had self-looping. Furthermore, identical pairs in nucleotide sequences emission randomly from the set {A, C, T, and G} with probability P_S.

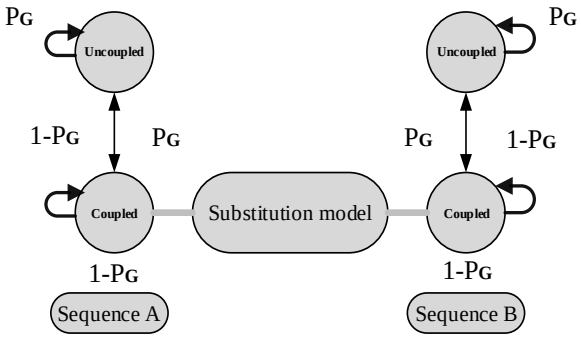

Figure 1: Coupled and uncoupled states of sequences

The probability computed by considering emission and transition along the path of two sequences. Thegoal is to find parameters, which maximize the probability of all sequences in the training set. Holmes [15] and Holmes and Durbin [19], propose log-likelihood equations using Viterbi algorithm [20] for match δ, mismatch ε, and gapλparameters as following:

$$\delta = log\ \frac{(1-P_G)^2(1+3(1-P_S)^2)}{16} \quad \textbf{(17)}$$

$$\varepsilon = log\ \frac{(1-P_G)^2(1-(1-P_S)^2)}{16} \quad \textbf{(18)}$$

$$\lambda = log\ \frac{P_G}{4} \quad \textbf{(19)}$$

Setting random values for P_G and P_S, once can obtain statically estimation of scoring scheme parameters for SW algorithm.

The maximum alignment D'_{max} in the equation (11) also predicting in accordance with new equations (17), (18), and (19) by compensation probabilistic values of $\delta, \varepsilon, and\ \lambda$ in equation (12) in order to define the probabilistic value forη , which denoted byη'. Probabilistic value η'denotes the highest alignment under the generative model as follow:

$$\eta' = \frac{\log\left[\left(\frac{1}{PG}-1\right)\sqrt{1+3(1-PS)^2}\right]}{\log\left[\frac{1+3(1-PS)^2}{1-(1-PS)^2}\right]} \quad \textbf{(20)}$$

III. Testing Values and Discussion

In order to test the estimated parameters, transition *PG* and emission *PS* are set to twenty random values. Furthermore, $\delta, \varepsilon,$ and λ are calculated by equation (17), (18), and (19), respectively. In addition, η and $\eta'$are calculated using equation (12) and (20), respectively. Moreover, maximum alignment$D'_{max}$is obtained by equation (11). These values are listed in Table 1. Negative and/or positive values both are accepted, because nucleotides and residues mutating over evolutionary distance rate [1, 5, 8, 9].

The experiments is conducted using the Code::Blocks (release 13.12), a C++ IDE. The dataset is obtained from the National Center for Biotechnology Information (NCBI) [21] using CLC Sequence Viewer

(version 7.0.2), which is a GUI software enable users to download and analysis bioinformatics data. Some parameters are adjusted in the software such as organism ID, modification, sequence length, and gene location and name. Furthermore, another adjustment of parameters are considered at NCBI genetic databank, where for every dataset a destination is used to allocate the sequence file, as well as the format. A standard FASTA format selected, which based on text format for explaining nucleotide or amino acid sequences.

As illustrated by the Fig.2, values obtained by η and η' are perfectly symmetrical, which in turn indicate the significance and accuracy relation of statically estimation of parameters involved in calculating η and η'.

Another observation in Table 1, is a direct correlation between η and D'_{max}. In general, when η increase, D'_{max} decrease conversely, see Fig. 3. Inverse relation between η and D'_{max} obviously occurs when $(\eta \geq 0)$. However, when $(\eta < 0)$ D'_{max} has fixed values equal to (-1), indicate that D'_{max} scores higher gap region than matches, which is a bottleneck. Highest negative values of D'_{max} occurs when $(\eta \geq \frac{1}{2})$. In this case, D'_{max} minimally gap as mismatch over gaps. The maximum alignment of D'_{max} determines efficiently when setting $0 < \eta < \frac{1}{2}$, thus D'_{max} score higher gap region than mismatches.

Table 1: Values of SW parameters using HMM

ID	*PG*	*PS*	δ	ε	λ	η	η'	D'_{max}
1	0.2	0.7	-1.29	-1.4	-1.36	6.06	6.06	-39
2	0.2	0.6	-1.17	-1.43	-1.41	3.15	3.15	-19.04
3	0.2	0.5	-1.09	-1.49	-1.41	2.13	2.13	-12.98
4	0.4	0.6	-1.48	-1.72	-1	1.1	1.1	-6.5
5	0.3	0.2	-1.05	-1.86	-1.16	0.79	0.79	-4.7
6	0.2	0.1	-0.87	-2.31	-1.24	0.56	0.56	-3.35
7	0.2	0	-0.79	-2.8	-1.34	0.47	0.47	-2.9
8	0.5	0.6	-1.62	-1.89	-0.9	0.35	0.35	-2.67
9	0.5	0.1	-1.23	-2.42	-0.93	0.27	0.27	-2.52
10	0.5	0.2	-1.39	-2.28	-0.88	0.21	0.21	-2.42
11	0.6	0.5	-1.68	-2.05	-0.85	0.03	0.03	-2.25
12	0.5	0.9	-1.79	-1.83	-0.9	0.01	0.01	-2.05
13	0.6	0.6	-1.78	-2.06	-0.83	-0.21	-0.21	-1
14	0.7	0.5	-1.99	-2.36	-0.76	-0.63	-0.63	-1
15	0.9	0.2	-2.99	-4.04	-0.63	-0.82	-0.82	-1
16	1	0	-5.38	-7.12	-0.6	-1.2	-1.2	-1
17	0.7	0.7	-2.15	-2.3	-0.76	-2.16	-2.16	-1
18	0.6	0.8	-2.04	-2.13	-0.79	-2.62	-2.62	-1
19	0.9	0.6	-2.86	-3.09	-0.66	-3.27	-3.27	-1
20	0.9	0.7	-2.79	-2.9	-0.67	-6.29	-6.29	-1

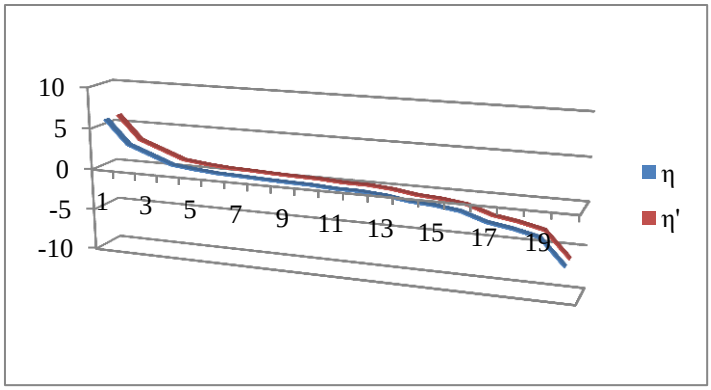

Figure 2: Perfect symmetric of η and η′

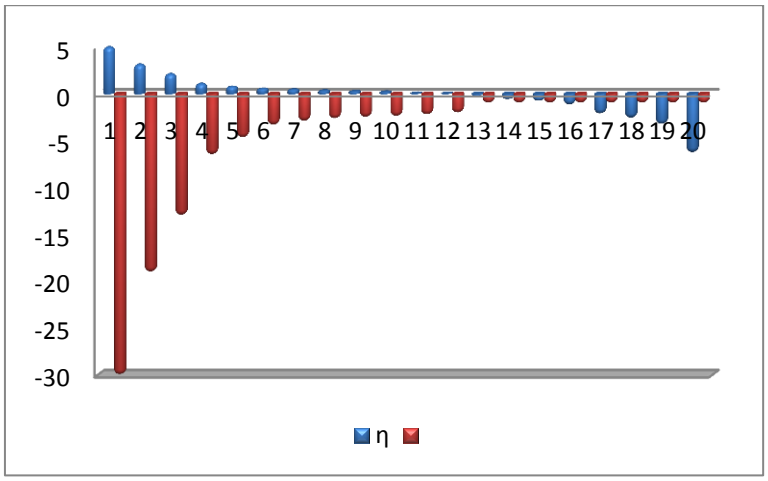

Figure 3: Inverse relationship between η and D'_{max}

IV. Conclusion

This paper describes the methodology used in estimating local sequences alignment parameters.The experiments are conducted using a real environment. Furthermore, accuracy and correctness of the optimal alignment results are evaluated using mathematically and statically model.

The validation process of the optimal alignment is conducted using real dataset obtained from NCBI using CLC Sequence Viewer. Moreover, local sequence alignment parameters are adjusting mathematically, statically, and testing to obtain the optimal alignment. The validation process starts with the mathematical induction for the parameters. Statically estimation using HMM is proposed in order to predict the best alignment parameters and to prove the accuracy of the alignment. In addition, the parameters are adjusted, and the accuracy is tested by comparing mathematically and statically results. Optimal alignment is calculated and compared using Code::Blocks. Finally, optimal alignment parameters reported, and the whole system is evaluated.

References

[1] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, pp. 443-453, 1970.
[2] T. Smith and M. Waterman, "Identification of common molecular subsequences," *J. Mol. Bwl*, vol. 147, pp. 195-197, 1981.
[3] D. Díaz, F. J. Esteban, P. Hernández, J. A. Caballero, G. Dorado, and S. Gálvez, "Parallelizing and optimizing a bioinformatics pairwise sequence alignment algorithm for many-core architecture," *Parallel Computing*, vol. 37, pp. 244-259, 2011.
[4] Y. Zhang, S. Misra, D. Honbo, A. Agrawal, W. Liao, and A. Choudhary, "Efficient pairwise statistical significance estimation for local sequence alignment using GPU," in *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, 2011, pp. 226-231.
[5] S. Brenner, "Optimal Pairwise Alignment," in *Introduction to computational biology: an evolutionary approach*, B. Haubold and T. Wiehe, Eds., ed: Springer, 2006, pp. 11-42.
[6] S. Chib, "Markov chain Monte Carlo methods," *Bayesian Theory and Applications*, p. 87, 2013.
[7] M. Bańbura and M. Modugno, "Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data," *Journal of Applied Econometrics*, vol. 29, pp. 133-160, 2014.
[8] M. O. Dayhoff and R. M. Schwartz, "A model of evolutionary change in proteins," in *In Atlas of protein sequence and structure*, 1978.
[9] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc Natl Acad Sci U S A*, vol. 89, pp. 10915-9, Nov 15 1992.
[10] M. Imelfort, "Sequence comparison tools," *Bioinformatics.*, pp. 13-37, 2009.
[11] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of molecular biology*, vol. 162, pp. 705-708, 1982.
[12] M. S. Waterman, M. Eggert, and E. Lander, "Parametric sequence comparisons," *Proc Natl Acad Sci U S A*, vol. 89, pp. 6090-3, Jul 1 1992.
[13] M. Vingron and M. S. Waterman, "Sequence alignment and penalty choice. Review of concepts, case studies and implications," *J Mol Biol*, vol. 235, pp. 1-12, Jan 7 1994.
[14] R. Olsen, T. Hwa, and M. Lassig, "Optimizing smith-waterman alignments," *arXiv preprint cond-mat/9811225*, 1998.
[15] I. Holmes, "Studies in probabilistic sequence alignment and evolution," Queens' College, 1998.
[16] J. L. Thorne, H. Kishino, and J. Felsenstein, "An evolutionary model for maximum likelihood alignment of DNA sequences," *J Mol Evol*, vol. 33, pp. 114-24, Aug 1991.
[17] K. Smith, "Hidden Markov Models in Bioinformatics with Application to Gene Finding in Human DNA 308-761 Machine Learning Project," ed, 2002.
[18] R. Durbin, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*: Cambridge university press, 1998.
[19] I. Holmes and R. Durbin, "Dynamic programming alignment accuracy," *J Comput Biol*, vol. 5, pp. 493-504, Fall 1998.
[20] G. D. Forney Jr, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, pp. 268-278, 1973.
[21] NCBI. (2014, 15 Feb). *National Center for Biotechnology Information*. Available: http://www.ncbi.nlm.nih.gov/