Estimating Pairwise Statistical Significance of Protein Local Alignments Using a Clustering-Classification Approach Based on Amino Acid Composition

Ankit Agrawal[1], Arka Ghosh[2], and Xiaoqiu Huang[1]

[1] Department of Computer Science, Iowa State University,
226 Atanasoff Hall, Ames, IA 50011-1041, USA
{ankitag,xqhuang}@iastate.edu
[2] Department of Statistics, Iowa State University, 303 Snedecor Hall
Ames, IA, 50011-1210, USA
apghosh@iastate.edu

Abstract. A central question in pairwise sequence comparison is assessing the statistical significance of the alignment. The alignment score distribution is known to follow an extreme value distribution with analytically calculable parameters K and λ for ungapped alignments with one substitution matrix. But no statistical theory is currently available for the gapped case and for alignments using multiple scoring matrices, although their score distribution is known to closely follow extreme value distribution and the corresponding parameters can be estimated by simulation. Ideal estimation would require simulation for each sequence pair, which is impractical. In this paper, we present a simple clustering-classification approach based on amino acid composition to estimate K and λ for a given sequence pair and scoring scheme, including using multiple parameter sets. The resulting set of K and λ for different cluster pairs has large variability even for the same scoring scheme, underscoring the heavy dependence of K and λ on the amino acid composition. The proposed approach in this paper is an attempt to separate the influence of amino acid composition in estimation of statistical significance of pairwise protein alignments. Experiments and analysis of other approaches to estimate statistical parameters also indicate that the methods used in this work estimate the statistical significance with good accuracy.

Keywords: Clustering, Classification, Pairwise local alignment, Statistical significance.

1 Introduction

Sequence alignment is extremely useful in the analysis of DNA and protein sequences [1]. Sequence alignment forms the basic step of making various high level inferences about the DNA and protein sequences - like homology, finding protein function, protein structure, deciphering evolutionary relationships, etc.

I. Măndoiu, R. Sunderraman, and A. Zelikovsky (Eds.): ISBRA 2008, LNBI 4983, pp. 62–73, 2008.

There are many programs that use some well known algorithms [2,3] or their heuristic version [1,4,5]. Recently, some enhancements in alignment program features have also become available [6,7] using difference blocks and multiple scoring matrices. Quality of a pairwise sequence alignment is gauged by the statistical significance rather than the alignment score alone, i.e., if an alignment score has a low probability of occurring by chance, the alignment is considered statistically significant.

For ungapped alignments, rigorous statistical theory for the alignment score distribution is available [8], and it was shown that the statistical parameters K and λ can be calculated analytically for a pair of sequences with given amino acid composition and scoring scheme. However, no perfect theory currently exists for gapped alignment score distribution, and for score distributions from alignment programs using additional features like difference blocks [6], and which use multiple parameter sets [7]. The problem of accurately determining the statistical significance of gapped sequence alignment has attracted a lot of attention in the recent years [9,10,11,12,13,14,15]. There exist a couple of good starting points for statistically describing gapped alignment score distributions [16,17], but a complete mathematical description of the optimal scores distribution remains far from reach [17]. Some excellent reviews on statistical significance in sequence comparison are available in the literature [18,19,20].

The statistical significance of a pairwise alignment depends upon various factors sequence alignment method, scoring scheme, sequence length, and sequence composition [19]. The straightforward way to estimate statistical significance of scores from an alignment program for which the statistical theory is unavailable is to generate a distribution of alignment scores using the program with randomly shuffled versions of the pair of sequences, and compare the obtained score with the generated score distribution, either directly or by fitting an extreme value distribution (EVD) curve (explained in the next section) to the generated distribution to get the EVD parameters K and λ, and using the EVD formula with the estimated K and λ to calculate the statistical significance of the obtained score. However, the parameters thus obtained are ideally valid only for the specific sequence pair under consideration, and for any other sequence pair, the parameters should be recomputed by generating another distribution, which is very time-consuming and impractical.

Thus, BLAST2.0 [1] uses a lookup method wherein the parameters K and λ are pre-computed for different scoring schemes assuming average amino acid composition of both sequences. PRSS program in the FASTA package [4,5,9] calculates the statistical significance of an alignment by aligning them, shuffling the second sequence up to 1000 times, and estimating the statistical significance from the distribution of shuffled alignment scores. It uses maximum likelihood to fit an EVD to the shuffled score distribution. A similar approach is also used in HMMER [21]. It also uses maximum likelihood fitting [22] and also allows for censoring of data left of a given cutoff, for fitting only the right tail of the histogram. A heuristic approximation of the gapped local alignment score distribution is also available [10], and based on these statistics, accurate formulae

for statistical parameters K and λ for gapped alignments are derived and implemented in a program called ARIADNE [11]. These methods can provide an accurate estimation of statistical significance for gapped alignments, but currently do not incorporate the additional features of sequence alignment, like using difference blocks and multiple parameter sets [6,7].

The problem of estimating the statistical significance of the database searches has been addressed in much detail over the past two decades as discussed earlier. However, accurate estimation of the statistical significance of specific pairwise alignments needs directed research efforts. It is an important problem critical in comparison of various alignment programs, and especially with new alignment programs coming up with additional features to suit the features of the real biological sequences, this problem of estimating statistical significance for pairwise sequence alignments becomes particularly important. It has also been shown recently [23] that pairwise statistical significance is a better indicator of homology than database statistical significance. The method used in [23], although was shown to be accurate, but is also very time-consuming, as it involves generating a score distribution of tens of thousands of alignments. The need for faster methods for estimating pairwise statistical significance was also stressed in [23].

In this paper, we propose and implement a simple clustering-classification approach that clusters the universe of protein sequences based on amino acid composition, and estimates the parameters K and λ for all cluster pairs for different scoring schemes and alignment methods. In this way, we attempt to separate the dependence of the statistical parameters K and λ on amino-acid composition from other factors like alignment method and scoring schemes. The task of estimating statistical significance thus reduces to classifying the sequences to appropriate clusters, and using the corresponding K and λ values of the classified cluster pair. This approach is similar to the lookup method used in BLAST2.0 [1] but takes into account the features of the specific sequence pair being aligned. For simple alignment methods, the results are also presented using other approaches (PRSS [4,5,9] and ARIADNE [11]), and for advanced alignment methods [6,7] currently no other quick methods are available to estimate pairwise statistical significance except the method described in this paper.

2 The Extreme Value Distribution for Ungapped and Gapped Alignments

Just as the distribution of the sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution (central limit theorem), the distribution of the maximum of a large number of i.i.d. random variables tends to an extreme value distribution (EVD). This is an important fact, because it allows us to fit an EVD to the score distribution from any local alignment program, and use it for estimating statistical significance of scores from that program. The distribution of Smith-Waterman local alignment score between random, unrelated sequences is approximately a Gumbel-type EVD [8]. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP

(High-scoring Segment Pairs which correspond to the ungapped local alignment) scores are characterized by two parameters, K and λ. The probability that the optimal local alignment score S exceeds x is given by

$$\Pr(S > x) \sim 1 - exp[-Kmne^{-\lambda x}]$$

This is valid for ungapped alignments [8], and the parameters K and λ can be computed analytically from the substitution scores and sequence compositions. An important point here is that this scheme allows for the use of only one substitution matrix. For the gapped alignment, no perfect statistical theory has yet been developed, although there is ample evidence that it also closely follows an extreme value distribution [9,11,24,7].

3 Clustering-Classification Approach

This paper presents a simple clustering-classification approach based on amino acid composition for estimating statistical significance of pairwise protein local alignments, which is essentially an enhanced lookup method, where K and λ values are pre-computed for each cluster pair by simulation. Subsequently, for a given sequence pair, the sequences are individually classified to the corresponding clusters based on their amino acid composition, and the K and λ parameters for the cluster pair are used for statistical significance calculation of alignments of the sequence pair.

3.1 Clustering

There are many algorithms available for clustering like hierarchical clustering, k-means clustering, etc. [25]. Here we are dealing with clustering the universe of protein sequences whose number is in hundreds of thousands. Therefore, we use k-means clustering as hierarchical methods typically involve the computation of a distance matrix of quadratic complexity with respect to the input size. In this work, we have used the k-means implementation in R package [26]. Each of the k clusters of sequences is represented by a single representative sequence (central sequence), and subsequently the parameters K and λ are computed for each pair of the k representative sequences. Given below a pseudo code for the clustering module:

```
alphabet = "ACDEFGHIKLMNPQRSTVWY"   #protein alphabet (amino acids)
sequences4R = set of sequences to be clustered
nSeq = number of sequences
for (i in 1:nSeq) {
  seqArray = sequences4R[i]
  lenSeq=length(seqArray)
  for (j in 1:lenAlphabet-1) {
    AACounts[i,j] = number of occurrences of amino acid alphabet[j] in seqArray
  }
  AAComposition[i,]=AACounts[i,]/lenSeq
}
```

```
k = number of clusters
seqClusters = clustered sequences based on AAComposition
for (i in 1:k) {
  clust_reprSeq[i] = representative sequence of cluster[i]
}
for (i in 1:k) {
  for (j in 1:i) {
    Compute the value of K and lambda by empirical simulation
    K_clusters[i,j] = K_clusters[j,i] = estimated K
    lambda_clusters[i,j] = lambda_clusters[j,i] = estimated lambda
  }
}
```

3.2 Classification

Given two protein sequences for estimation of statistical parameters, they are classified individually to the appropriate clusters. Each of the k clusters obtained in the clustering step have their center, which corresponds to the central amino acid composition for that cluster. A sequence is classified to the cluster that minimizes the sum of squares of differences between the amino acid composition of the sequence and the central amino acid composition of the cluster. Subsequently, the pre-computed K and λ values for the classified cluster pair are used for the statistical significance estimation of alignments of the two input sequences. Given below a pseudo code for the classification module:

```
alphabet = "ACDEFGHIKLMNPQRSTVWY"
sequences4R = set of two sequences for which K and lambda is to be estimated
nSeq = 2
for (i in 1:nSeq) {
  seqArray = sequences4R[i]
  lenSeq=length(seqArray)
  for (j in 1:lenAlphabet-1) {
    AACounts[i,j] = number of occurrences of amino acid alphabet[j] in seqArray
  }
  AAComposition[i,]=AACounts[i,]/lenSeq
}
k = number of clusters
for (j in 1:nSeq) {
  classifiedCluster[j] = classified cluster based on AAComposition
}
estimatedK = K_clusters[classifiedCluster[1],classifiedCluster[2]]
estimatedLambda=lambda_clusters[classifiedCluster[1],classifiedCluster[2]]
```

4 Tools and Programs Used

We worked with the alignment programs SIM [27], which is an ordinary alignment program (similar to SSEARCH), and GAP4 [7], which allows dynamically

finding similarity blocks and difference blocks [6], as well as using multiple parameter sets (scoring matrices, gap penalties, difference block penalties) to generate a single pairwise alignment. For estimating the statistical parameters K and λ, we used several programs. First is PRSS from the FASTA package [4,5,9], which takes two protein sequences and one set of parameters (scoring matrix, gap penalty), generates the optimal alignment, and estimates the K and λ parameters by aligning up to 1000 shuffled versions of the second sequence, and fitting an EVD using Maximum Likelihood. In addition to uniform shuffling, it also allows for windowed shuffling. We also used ARIADNE [11], that uses an approximate formula to estimate gapped K and λ from ungapped K and λ, which are calculable analytically as described before. Both these methods are currently applicable only for alignment methods using one parameter set. We also used the Linear Regression fitting program used in [7] to estimate K and λ from an empirical distribution of alignment scores. Finally, we also used the Maximum likelihood method [22] and corresponding routines in the HMMER package [21] to fit an EVD to the empirical distribution. Here type-I censoring is defined as the one in which we fit only the data right of the peak of the histogram [22], and type-II censoring is defined as one where the cutoff is set to the score that corresponds to a normalized E-value of 0.01. We used all these methods to estimate K and λ values for a pair of representative sequences for a given alignment scheme.

5 Experiments and Results

We downloaded all the available 261513 Swissprot protein sequences from http://www.ebi.ac.uk/FTP/. The statistics of the lengths of the sequences are given in Table 1, and the histogram of sequence lengths less than 1000 is shown in Fig. 1. Clearly, the variation in sequence length is very extreme, although the length of most of the sequences is in the range of 150 to 450. To minimize the influence of variation in length, we only select the sequences with length between the 1st and 3rd quartile for clustering. The number of sequences between 1st and 3rd quartile is 131486. The amino acid composition of all these sequences is calculated, and an implementation of the k-means clustering algorithm in R package [26] is used to cluster the sequences into k=5 clusters, based on their amino acid composition. The k-means implementation in R returns for each of the k clusters its center, its within-sum-of-squares, its size, and of course, the classification of the input data points in one of the k clusters.

Fig. 2 is an attempt to visualize the clusters by representing the 20 dimensional amino-acid-composition vector as a point in $x-y$ plane using the first two amino-acid-compositions. Although it does not give a full picture of the clusters and their separation, it nonetheless gives some idea of how the clusters are located. One representative sequence for each of the k clusters is then selected by choosing the one whose amino-acid-composition vector is the closest to the center of the cluster (i.e., which gave the minimum sum of square of differences). Then, for each pair of the representative sequences, the parameters K and λ are

estimated using the methods described earlier. This work presents the preliminary analysis taking k as 5 to study the effectiveness of this method. However, no detailed study on the number of clusters has been presented in this work. For k

Table 1. Statistics of lengths of sequences

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
2	165	296	365.7	460	34350

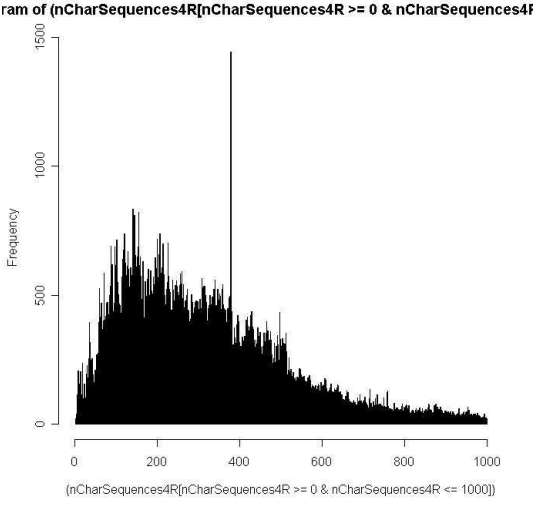

Fig. 1. Histogram of length of sequences with length ≤ 1000

$= 5$, there exist 15 ($=^5 C_2 + 5$) different pairwise cluster combinations. Table 2 gives the K and λ estimates for one of the 15 pairwise clusters ($\langle 3, 2 \rangle$). Here, we used several options for the alignment parameters. For substitution matrices, we used all possible combinations of BLOSUM45, BLOSUM62, and BLOSUM100 matrices. The alignment program GAP4 [7] is capable of using multiple substitution matrices to produce a single optimal alignment of two sequences. It requires all substitution matrices to be in the same scale, and thus all matrices were used in 1/3 bit scale. Other parameters like gap penalties, etc. were the same as the default values used in GAP4 [7] for matrices in 1/3 bit scale. We used the various programs for statistical parameter estimation as described earlier. Rows in first half of Table 2 show the K and λ estimates from ARIADNE [11] and PRSS [4,5,9], and the second half of the table show the estimates from ML and LR. As pointed out earlier, ARIADNE and PRSS currently can work only with one parameter set, and cannot estimate the pairwise statistical significance parameters for alignment programs that use multiple parameter sets, and hence, the corresponding entries in Table 2 are not available.

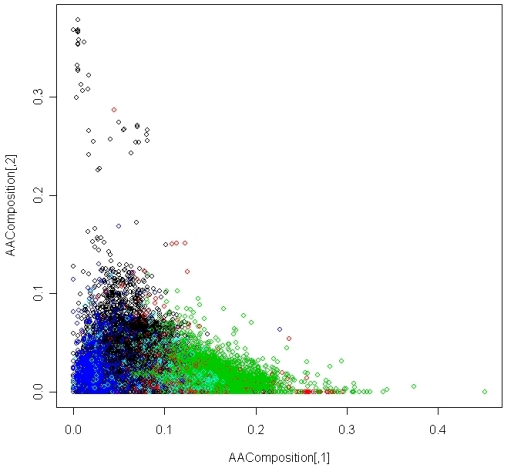

Fig. 2. k-means clusters (k=5). Sequences in each cluster are represented by different colors. This visualization represents the 20-dimensional amino-acid-composition vector by a 2-dimensional vector (corresponding to the first two entries of the 20-dimensional amino-acid-composition vector), and hence is not complete, but gives an overall idea of how the clusters are located.

Table 2. K and λ estimates for the cluster pair $\langle 3, 2 \rangle$

Substitution Matrix	Gap Open	Gap Ext	ARIADNE		PRSS(1000 shuffles)					
					uniform		-w 10		-w 20	
			K	λ	K	λ	K	λ	K	λ
BLOSUM45	12	2	0.01795	0.184148	0.0329	0.1869	0.03736	0.1941	0.0381	0.1974
BLOSUM62	14	3	0.06445	0.200311	0.0956	0.2104	0.1108	0.2154	0.1212	0.2181
BLOSUM100	16	4	0.15101	0.210326	0.1888	0.224	0.2624	0.2328	0.1564	0.2198
BL45,62,100	12,14,16	2,3,4	NA	NA	NA	NA	NA	NA	NA	NA
BL45,BL62	12,14	2,3	NA	NA	NA	NA	NA	NA	NA	NA
BL45,BL100	12,16	2,4	NA	NA	NA	NA	NA	NA	NA	NA
BL62,BL100	14,16	3,4	NA	NA	NA	NA	NA	NA	NA	NA

Substitution Matrix	Gap Open	Gap Ext	Maximum Likelihood (100000 shuffles)						LinearRegr. (100000 shfls)	
			Full		Censored-I		Censored-II			
			K	λ	K	λ	K	λ	K	λ
BLOSUM45	12	2	0.03387	0.189248	0.0316	0.1876	0.089487	0.204045	0.1083	0.2063
BLOSUM62	14	3	0.08757	0.205953	0.0875	0.2058	0.045709	0.196304	0.2389	0.2195
BLOSUM100	16	4	0.18503	0.2191	0.1761	0.2179	0.358664	0.228915	0.4009	0.2304
BL45,62,100	12,14,16	2,3,4	0.10576	0.194163	0.0967	0.1923	0.096223	0.192396	0.2358	0.2044
BL45,BL62	12,14	2,3	0.06773	0.194176	0.0919	0.1982	0.123769	0.202883	0.1551	0.2057
BL45,BL100	12,16	2,4	0.09969	0.195183	0.0911	0.1932	0.147417	0.200051	0.3205	0.2102
BL62,BL100	14,16	3,4	0.15685	0.207436	0.1570	0.2074	0.243203	0.21407	0.2807	0.2157

It was reported in [23] that that Maximum likelihood fitting with type-I censoring gives the most accurate estimates of K and λ for estimation of pairwise statistical significance. Therefore, we report the corresponding the K and λ estimates for all cluster-pairs in Table 3, presenting the final result of this work. There are 7 sub-tables in Table 3, each showing the K and λ estimates for all cluster pairs for a unique scoring scheme (7 scoring schemes are presented here).

Table 3. Pairwise cluster statistical significance parameters for a variety of scoring schemes

Parameters	Substitution Matrix: BLOSUM45; Gap Open Penalty: 12; Gap Extension Penalty: 2									
	λ					K				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0.1358571					0.036457				
2	0.212847	0.1642628				0.055198	0.022692			
3	0.2076972	0.187666	0.1501918			0.052404	0.031631	0.020645		
4	0.2439074	0.1868858	0.2037552	0.1584317		0.070919	0.031597	0.040751	0.018778	
5	0.1948708	0.1909384	0.190503	0.1971262	0.1733189	0.041417	0.033241	0.032547	0.034713	0.024396

Parameters	Substitution Matrix: BLOSUM62; Gap Open Penalty: 14; Gap Extension Penalty: 3									
	λ					K				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0.155231					0.053164				
2	0.2214101	0.1904642				0.108926	0.076897			
3	0.2173865	0.2058921	0.1772875			0.11316	0.087505	0.059218		
4	0.2461108	0.209065	0.2199229	0.1976537		0.12987	0.091884	0.104014	0.085541	
5	0.2085433	0.2057725	0.2063069	0.2142532	0.193569	0.09491	0.087126	0.089158	0.095499	0.069993

Parameters	Substitution Matrix: BLOSUM100; Gap Open Penalty: 16; Gap Extension Penalty: 4									
	λ					K				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0.1866353					0.154048				
2	0.228544	0.2064503				0.200226	0.193781			
3	0.2242455	0.2179898	0.2045788			0.192326	0.17616	0.167797		
4	0.2456977	0.2182969	0.2276359	0.2107703		0.209087	0.173654	0.183061	0.167819	
5	0.221009	0.2162649	0.2202855	0.2244157	0.2126472	0.188667	0.164509	0.173582	0.179575	0.164874

Parameters	Substitution Matrix: BL45, BL62,BL100; Gap Open Penalty: 12,14,16; Gap Extension Penalty: 2,3,4									
	λ					K				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0.1368407					0.049183				
2	0.213582	0.1701292				0.159863	0.063276			
3	0.2079975	0.192342	0.154174			0.14941	0.096799	0.040994		
4	0.2373249	0.1928895	0.207206	0.1650268		0.198173	0.095714	0.123883	0.043014	
5	0.1994941	0.1937772	0.1957841	0.2015482	0.1787635	0.12346	0.099895	0.104908	0.10812	0.066458

Parameters	Substitution Matrix: BL45, BL62; Gap Open Penalty: 12,14; Gap Extension Penalty: 2,3									
	λ					K				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0.1402432					0.05231				
2	0.2145139	0.1712422				0.100205	0.045569			
3	0.2093419	0.1982169	0.1540202			0.101861	0.091968	0.031788		
4	0.2411866	0.1931565	0.2063857	0.1642365		0.125784	0.063434	0.075368	0.030885	
5	0.1987875	0.1941648	0.1946307	0.2020014	0.1783505	0.079681	0.064921	0.066602	0.072043	0.046346

Parameters	Substitution Matrix: BL45, BL100; Gap Open Penalty: 12,16; Gap Extension Penalty: 2,4									
	λ					K				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0.1382574					0.05416				
2	0.2141082	0.1717717				0.14858	0.065682			
3	0.2092158	0.1932967	0.1537908			0.141856	0.091108	0.036667		
4	0.2404284	0.1954375	0.2076836	0.1648767		0.208871	0.104467	0.113331	0.041522	
5	0.2003162	0.1949345	0.1968851	0.2041312	0.1789941	0.11604	0.094415	0.096769	0.109758	0.059355

Parameters	Substitution Matrix: BL62, BL100; Gap Open Penalty: 14,16; Gap Extension Penalty: 3,4									
	λ					K				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0.1612023					0.088929				
2	0.2201957	0.1928412				0.185743	0.137453			
3	0.2210159	0.2074159	0.1823947			0.251905	0.157085	0.103554		
4	0.2406616	0.210198	0.2198564	0.1994292		0.212245	0.164609	0.180414	0.144506	
5	0.2090656	0.2068202	0.2068121	0.2145398	0.1982235	0.159461	0.152743	0.142889	0.161868	0.12984

The K and λ estimates in table 3 are for 1/3-bit scaled substitution matrices. For each scoring scheme, there is a wide variation in the estimated K and λ values. For instance, in the first sub-table, λ values range from 0.1358571 to 0.2439074, and K values range from 0.018778 to 0.070919, although all the pairwise alignments of random sequences for getting the K and λ estimates in the

first sub-table were done using the SIM program with BLOSUM45 substitution matrix, gap open penalty 12, and gap extension penalty 2, i.e. using the same scoring scheme. Since the only contributing factor for the difference between K and λ values for different cluster pairs is the amino acid composition, we can observe that the statistical parameters heavily depend on the amino acid composition. Clustering the protein sequences into groups of similar amino acid composition has therefore to some degree separated the dependence of the statistical significance parameters on the amino acid composition, which is very helpful for quick and accurate estimates of statistical significance for specific pairwise alignments. Once parameter estimation for the cluster-pairs is done for a given scoring scheme, subsequent statistical significance estimation for any sequence pair using the same soring scheme is very quick, since it only involves classification of the sequences to corresponding clusters, and using the statistical parameters for the corresponding cluster-pair.

6 Conclusion and Future Work

The implementation of a clustering-classification based approach for estimating the statistical parameters K and λ for estimating the statistical significance of pairwise alignments is done and is experimented with. The clusters are based on the amino-acid composition and the estimates of the statistical parameters K and λ for each cluster-pair are calculated by simulation. Given two sequences, the estimate of K and λ for that pair is given by the K and λ values corresponding to the cluster-pair to which the given sequences are classified based on the amino-acid-composition.

The estimated values of K and λ for different clusters show a considerable variability, even for the same alignment scoring scheme, which suggests that the influence of amino acid composition on statistical parameters K and λ is very strong, and it is imperative to use different K and λ values for different sequences. The clustering technique used in this work has therefore separated the influence of amino acid composition on statistical parameters, which is the main contribution of this paper. Another major significance of this work is that this method can be applied to any new alignment program with any scoring scheme without the knowledge of the statistics of the alignment procedure (which is in general difficult to determine). Once the influence of amino acid composition on statistical significance parameters is separated from other factors, all that needs to be done is the accurate estimation of the statistical parameters for all cluster pairs using the new alignment program, and subsequently use those values for any pair of sequences with individually similar amino acid composition as that of the clusters to which the pair of sequences are individually classified. Especially with a number of new alignment methods being developed, this technique is expected to be very useful in comparing them.

Although the simple idea is very promising, it is unclear how well it works for an application where statistical significance is used, like homology detection. This approach is just a beginning of the efforts to separate the influence of amino

acid composition, and clustering is just one of the many methods which can do so. It may be possible that it is not the exact composition clusters that two protein sequences under comparison fall into that matters, but instead, simply the difference between the composition distributions of the two proteins, which needs further exploration. Another shortcoming of this work is that by clustering hundreds of thousands of sequences in to just five clusters, we lose a lot of information about the amino acid composition distribution across the real protein sequences. An analytical study of the amino acid composition distribution may be required to get the optimal number of clusters. Hence, this method can be further looked into in detail to evaluate the performance of clustering. Another improvement can be to use a small scale simulation along with the proposed approach to increase the accuracy of the statistical significance estimates.

References

1. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Research 25(17), 3389–3402 (1997)
2. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology 147(1), 195–197 (1981)
3. Sellers, P.H.: Pattern Recognition in Genetic Sequences by Mismatch Density. Bulletin of Mathematical Biology 46(4), 501–514 (1984)
4. Pearson, W.R.: Effective Protein Sequence Comparison. Methods in Enzymology 266, 227–259 (1996)
5. Pearson, W.R.: Flexible Sequence Similarity Searching with the FASTA3 Program Package. Methods in Molecular Biology 132, 185–219 (2000)
6. Huang, X., Chao, K.M.: A Generalized Global Alignment Algorithm. Bioinformatics 19(2), 228–233 (2003)
7. Huang, X., Brutlag, D.L.: Dynamic Use of Multiple Parameter Sets in Sequence Alignment. Nucleic Acids Research 35(2), 678–686 (2007)
8. Karlin, S., Altschul, S.F.: Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. Proceedings of the National Academy of Sciences, USA 87(6), 2264–2268 (1990)
9. Pearson, W.R.: Empirical Statistical Estimates for Sequence Similarity Searches. Journal of Molecular Biology 276, 71–84 (1998)
10. Mott, R., Tribe, R.: Approximate Statistics of Gapped Alignments. Journal of Computational Biology 6(1), 91–112 (1999)
11. Mott, R.: Accurate Formula for P-values of Gapped Local Sequence and Profile Alignments. Journal of Molecular Biology 300, 649–659 (2000)
12. Altschul, S.F., Bundschuh, R., Olsen, R., Hwa, T.: The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Research 29(2), 351–361 (2001)
13. Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F.: Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-based Statistics and Other Refinements. Nucleic Acids Research 29(14), 2994–3005 (2001)
14. Bundschuh, R.: Rapid Significance Estimation in Local Sequence Alignment with Gaps. In: RECOMB 2001: Proceedings of the fifth annual International Conference on Computational biology, pp. 77–85. ACM, New York (2001)

15. Poleksic, A., Danzer, J.F., Hambly, K., Debe, D.A.: Convergent Island Statistics: A Fast Method for Determining Local Alignment Score Significance. Bioinformatics 21(12), 2827–2831 (2005)
16. Kschischo, M., Lässig, M., Yu, Y.: Toward an Accurate Statistics of Gapped Alignments. Bulletin of Mathematical Biology 67, 169–191 (2004)
17. Grossmann, S., Yakir, B.: Large Deviations for Global Maxima of Independent Superadditive Processes with Negative Drift and an Application to Optimal Sequence Alignments. Bernoulli 10(5), 829–845 (2004)
18. Pearson, W.R., Wood, T.C.: Statistical Significance in Biological Sequence Comparison. In: Balding, D.J., Bishop, M., Cannings, C. (eds.) Handbook of Statistical Genetics, pp. 39–66. Wiley, Chichester (2001)
19. Mott, R.: Alignment: Statistical Significance. Encyclopedia of Life Sciences (2005), http://mrw.interscience.wiley.com/emrw/9780470015902/els/article/a0005264/current/abstract
20. Mitrophanov, A.Y., Borodovsky, M.: Statistical Significance in Biological Sequence Analysis. Briefings in Bioinformatics 7(1), 2–24 (2006)
21. Eddy, S.R.: Multiple Alignment Using Hidden Markov Models. In: Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T., Wodak, S. (eds.) Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, pp. 114–120. AAAI Press, Menlo Park (1995)
22. Eddy, S.R.: Maximum Likelihood Fitting of Extreme Value Distributions (1997), unpublished manuscript, citeseer.ist.psu.edu/370503.html
23. Agrawal, A., Brendel, V., Huang, X.: Pairwise Statistical Significance Versus Database Statistical Significance for Local Alignment of Protein Sequences. In: Măndoiu, I., Sunderraman, R., Zelikovsky, A. (eds.) ISBRA 2008. LNCS(LNBI), vol. 4983, pp. 50–61. Springer, Heidelberg (in press, 2008)
24. Olsen, R., Bundschuh, R., Hwa, T.: Rapid Assessment of Extremal Statistics for Gapped Local Alignment. In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pp. 211–222. AAAI Press, Menlo Park (1999)
25. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 2nd edn. Wiley-Interscience, Chichester (2003)
26. Language, R.A.: Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2006)
27. Huang, X., Miller, W.: A Time-efficient Linear-space Local Similarity Algorithm. Advances in Applied Mathematics 12(3), 337–357 (1991)