Report Number: WUCS-90-27

1990-08-01

Information Theoretic Estimation of Clone Overlap Probabilities

Jeffrey C. Beran-Koehn and Will D. Gillett

This technical report describes preliminary research investigating the relationship between information theory and Bayes' theory for estimation of the probability of clone overlap for the use in DNA restriction mapping. A number of languages (along with information theoretic metrics) capable of describing a hypothesized overlap are presented. For each language, the MML criterion is applied to the encoded overlaps of a pair of clones to search for that overlap which is most probable. The objective is to order the pair's encoded overlaps, based on the MML criterion, from the most to the least probable. This ordering is compared to... **Read complete abstract on page 2.**

Information Theoretic Estimation of Clone Overlap Probabilities

Jeffrey C. Beran-Koehn and Will D. Gillett

Complete Abstract:

This technical report describes preliminary research investigating the relationship between information theory and Bayes' theory for estimation of the probability of clone overlap for the use in DNA restriction mapping. A number of languages (along with information theoretic metrics) capable of describing a hypothesized overlap are presented. For each language, the MML criterion is applied to the encoded overlaps of a pair of clones to search for that overlap which is most probable. The objective is to order the pair's encoded overlaps, based on the MML criterion, from the most to the least probable. This ordering is compared to the ordering suggestion by the Bayesian probabilistic approach.

INFORMATION THEORETIC ESTIMATION OF CLONE OVERLAP PROBABILITIES

Jeffrey C. Beran-Koehn and Will D. Gillett

WUCS-90-27

August 1990

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

This work was supported in part by grant T15 LM07049 from the National Library of Medicine and grant 87-24 from the James S. McDonnell Foundation.

Abstract

This technical report describes preliminary research investigating the relationship between information theory and Bayes' theory for estimation of the probability of clone overlap for use in DNA restriction mapping. A number of languages (along with information theoretic metrics) capable of describing a hypothesized overlap are presented. For each language, the MML criterion is applied to the encoded overlaps of a pair of clones to search for that overlap which is most probable. The objective is to order the pair's encoded overlaps, based on the MML criterion, from the most to the least probable. This ordering is compared to the ordering suggested by a Bayesian probabilistic approach.

Outline

1. Introduction

1.1. DNA and Restriction Mapping

A molecule is a group of atoms joined together to form a distinct unit. To form more complex units, molecules join with each other or with single atoms. Long linear molecules of deoxyribonucleic acid (DNA) contain all the information necessary to control the chemistry of life, thereby determining the characteristics of an organism. The basic unit of information is called a nucleotide. DNA is composed of four different kinds of nucleotides represented by the letters A, T, G, and C. Thus, the information contained within a DNA molecule may be viewed as a sequence containing the letters A, T, G, C. Introductory material on this subject may be found in [3].

Restriction maps provide a guide to the location of certain nucleotide subsequences, called recognition sequences, within a DNA molecule. Using the random-clone strategy developed by Olson [5], a restriction map is created from a sample of several identical DNA molecules. These molecules are cut into many randomly overlapping segments, called clones. Each clone is then individually transferred into a bacterium. The clone uses the biological resources of the bacterial cell to reproduce multiple copies of itself. As the bacterial cell repeatedly divides, millions of identical copies of the clone are created. A restriction enzyme (RE) is applied to each of the clones separately. This enzyme is a chemical which cuts the DNA at a specific recognition sequence. The location of this cut is called a RE site and the resultant pieces of DNA are called (restriction) fragments. The restriction enzyme is allowed to work until the clones have been cut at all their recognition sites. The size, or length, of the fragments are measured by a process called gel electorphoresis. This reveals the distance between the RE sites, but their location within the clone (and within the original DNA molecule) is unknown. The entirety of the experimental process destroys the specific knowledge of a clone's location within the original DNA molecule and the order of the RE sites within each clone.

Several techniques exist to infer the positions of the initial RE sites. The key to these techniques is to discover the overlapping portions of each clone. Two clones are said to overlap if portions of their DNA come from the same location in the original DNA molecule. If this is the case, these clones should contain several identical fragments. The process of determining whether or not two fragments from different clones are identical is complicated. However, it is relatively easy to determine if two fragments have measured lengths within experimentally determined measurement error bounds. Two fragments whose lengths are within this error bound are referred to as matching, or matched fragments. The identification of an area of overlap allows the fragments of each clone to be partially ordered. All the matched fragments must be located on adjacent ends of the clones, while the unmatched fragments have their origin on opposite ends of the clones. The identification of matching fragments between two clones and the resulting ordering of the fragments within each clone is referred to as an overlap. Figure 1(a) illustrates the fragment length data of two clones. In describing the order of the fragments, "**x** or **y**" indicates that either **x** is followed by **y**, or **y** is followed by **x**, i.e., the order is not known. Conversely, "**x** and **y**" indicates that **x** occurs before **y**, i.e., the order is known. Figure 1(b) illustrates the ordering imposed by matching fragments **f1** with **f2** and **f3** with **f5**.

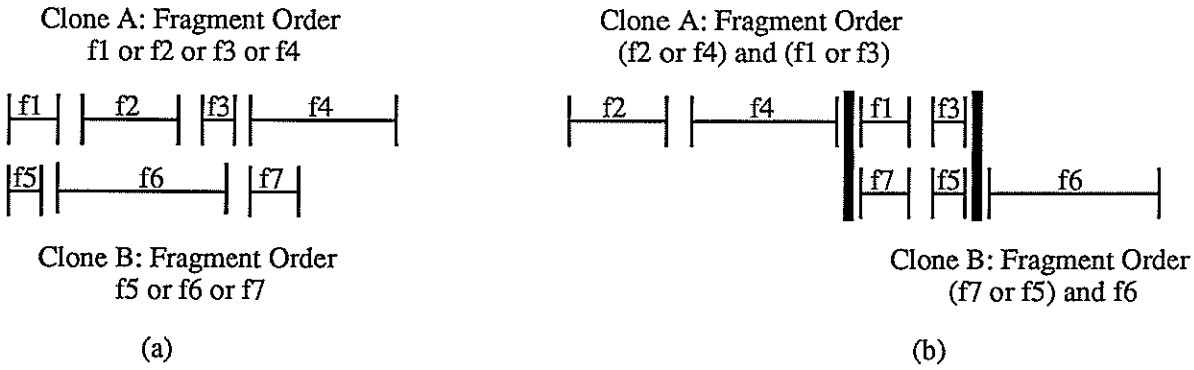

Figure 1: The ordering of fragments due to clone overlap

If the experimental process supplies enough clones which overlap the same area, the order of each fragment will become known revealing the positions of the RE sites. Figure 2 diagrammatically illustrates the entire process of creating a restriction map for a DNA molecule. For simplicity and clarity, the figures show clones with a region of overlap containing only a few matching fragments. In actual practice, many fragments must match (often four is sufficient) before clone overlap is inferred.

There are two sources of error which complicate the search for clone overlaps. First, the fragment length data is subject to measurement error which is proportional to the fragment length. Secondly, the original DNA molecule may contain several different fragments of the same size. Therefore, it is possible that there exists more than one manner in which two clones appear to overlap (only one of which is correct). Because of these errors, it is also possible that two clones which appear to overlap actually do not. Due to the uncertainty caused by these errors, it is necessary to estimate the probability of each of the possible overlaps between two clones.

Barnett [1] estimates the measurement error with a probability distribution **g**, where **g(Y | X)** is the probability that the measured length of a fragment is **Y** given that the true length is **X**. In her work, **g** is a "discretized" normal distribution with mean **X** and standard deviation (s_X). The standard deviation is equal to the product of **X** and the percent measurement error. Because fragment lengths are integers, **g** is discretized by integrating the normal distribution between **Y** - .5 and **Y** + .5.

$$g(Y \mid X) \equiv \int_{y=Y-.5}^{Y+.5} \frac{1}{\sqrt{2\pi}\sigma_x} e^{-.5\left(\frac{y-X}{\sigma_x}\right)^2} \, dy \qquad (1)$$

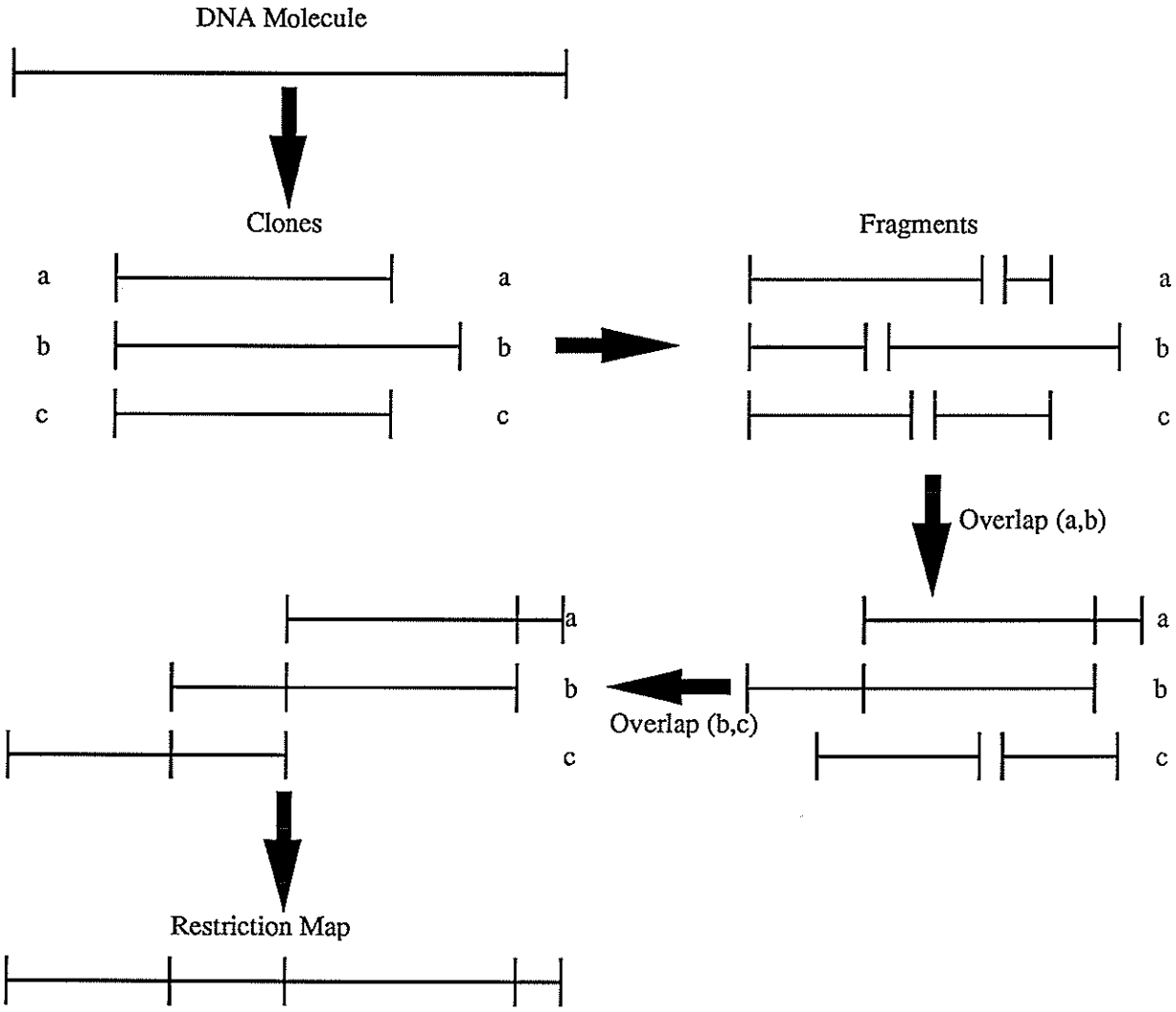

Figure 2: Creating a restriction map of a DNA molecule.

Barnett [1] estimates the error due to the multiplicity of fragment lengths with a probability distribution **f**, where **f(X)** is the probability that the true length of a fragment is **X**. In her work, **f** is a geometric distribution and **p** is the probability of a RE site occurring.

$$f(X) = p(1-p)^{X-1} \quad \text{where } p = \left(\frac{1}{4}\right)^{\#\text{nucleotides in recognition sequence}} \tag{2}$$

(This definition of **p** is based on the assumption that the frequencies of occurrence for the four nucleotides, A, T, G, and C, are equal.)

1.2.1 Bayes' Theory

Bayes' theorem provides a method of computing the probability of an hypothesis given specific evidence [6]. In this situation, we hypothesize a particular overlap between two clones based on the similarity of their fragment length data. If there are **n** possible overlaps between the clones, including the case that they do not overlap, and $\mathbf{overlap_p}$ is the particular overlap of interest, Bayes' theorem states that the probability of $\mathbf{overlap_p}$ given the fragment length **data** is:

$$P(overlap_p \mid data) = \frac{P(data \mid overlap_p)\ P(overlap_p)}{\sum_{i=1}^{n} P(data \mid overlap_i)\ P(overlap_i)} \tag{3}$$

The probability of the fragment length data given an overlap is defined as follows. Let $clone_1$ have **r** fragments, $clone_2$ have **s** fragments, and let there be **m** matched fragments. Number the fragments so that 1...**m** are the matched fragments. If $\mathbf{Y_j}$ and $\mathbf{Z_j}$ are the measured lengths of fragment **j** from from $clone_1$ and $clone_2$ respectively, then Barnett [1] states:

$$P(data \mid overlap_i) = P_1 \times P_2 \times P_3 \tag{4}$$

where P_k equals:

$$\begin{cases} \prod_{j=1}^{m} \sum_{X=1}^{\infty} g(Y_j|X)g(Z_j|X)f(X) & \text{if } k=1 \\ \prod_{j=m+1}^{r} \sum_{X=1}^{\infty} g(Y_j|X)f(X) & \text{if } k=2 \\ \prod_{j=m+1}^{s} \sum_{X=1}^{\infty} g(Z_j|X)f(X) & \text{if } k=3 \end{cases}$$

Barnett [1] calculates the prior probability of a given overlap as follows. If **t** is the number of fragment positions in the original DNA molecule (**t** is one less than the product of **p**, the probability of occurrence of an RE site, and the number of nucleotides in the original DNA molecule) then $P(overlap_i)$ equals:

$$\begin{cases} 1 - \frac{r+s-1}{t} & \text{if } m=0 \\ \frac{2}{t\binom{r}{m}\binom{s}{m}m!} & \text{if } 0<m<\min(r,s) \\ \frac{|r-s|+1}{t\binom{r}{m}\binom{s}{m}m!} & \text{if } m=\min(r,s) \end{cases} \quad (5)$$

In this work, Barnett's probability formulas will be taken to represent the true probability of clone overlap.

1.2.2 Information Theory

Information theory developed from the study of transmitting data over communications media. The data to be sent must first be encoded into a format which is compatible with the media. The encoded data is then sent over the media and decoded when received. To make this operation as efficient as possible, the most frequently occurring data elements should be encoded as compactly as possible. If **pdata** is the probability of a data element occurring, then the theoretical minimum length to encode this data is:

$$-\log (p_{data}) \quad (6)$$

Since probabilities are less than or equal to 1, the logarithm function is negated to insure positive quantities. If the base of the logarithm is 2, then the length is in bits. The Minimum Message Length (MML) criterion states that common data (i.e., highly probable data) are encoded a in few bits while rare data requires many bits for encoding. See [4] for an introduction to information theory.

The application of these principles have been extended to many areas outside of communication theory. Cheeseman [2] utilizes these ideas in his research of machine learning. He states that, for space efficiency, new information should not be stored as is received, but should be modeled based on what is already known. By describing this new information as an instance of an existing abstraction, a more compact encoding of the information may result. If the new information cannot be compactly encoded, this suggests that the previous data has little predictive value for the new information. In this case, the new information may either be considered as "noise", or as a possible new class of knowledge.

1.2.3 Relationship Between Bayes' Theory and Information Theory

Bayesian probability theory can be cast into the MML criterion. By taking the logarithm of (3) and negating the result, we can produce:

$$-\log(P(overlap_p|data)) = -\log(P(data|overlap_p))-\log(P(overlap_p))+constant \quad (7)$$

where the constant term accounts for the logarithm of the denominator of (3). Since we are only interested in the relative probability of different overlaps for a given pair of clones (and not in comparisons between different clone pairs), the constant term may be ignored. Using (4), the Minimum Message Length is the sum of two terms. The first term is the length required to encode the evidence, the fragment length data, supporting the hypothesized overlap. The second term is the length required to encode the hypothesized

overlap. If a suitable language for encoding these two terms is found, the MML criterion will be met, providing an estimation of the overlap probability.

1.3 Research Aims

This technical report presents preliminary research investigating a number of languages capable of encoding clone overlap. The term language is used informally in that it refers not only to a set of symbols which encode clone overlaps, but also to information theoretic measures that describe some characteristic of an overlap. Given a number of possible clone-clone overlaps, the MML criterion is applied to search for that overlap which is most probable. The objective is to evaluate a language's ability, when applied across the hypothesized overlaps, to order each overlap from most to least probable (based on Barnett's [1] probability analysis). Chapter 2 describes the experimental procedure implemented for this study. Each language, along with its performance, is discused in Chapter 3. Chapter 4 contains general observations about the results. Chapter 5 outlines areas of future research, and acknowledgements for this work are given in Chapter 6.

2. Experimental Procedure

2.1 Data Generation

With available software, a simulated DNA molecule consisting of 800,000 nucleotides was created. From this original DNA, 256 unique clones were generated. A restriction enzyme was applied to these clones producing restriction fragments. To simulate measurement error, normal random error with a standard deviation of 1.5 percent was randomly introduced into the fragment length data.

2.2 Data Sampling

Two clones were selected at random from the simulated data and each of their feasible overlaps were calculated. If the largest overlap (the overlap containing the most matched fragments) between the clones contained at least three matching fragments, the clone data and each overlap were saved, otherwise all the data were discarded. This process was repeated until one hundred pairs of clones and their feasible overlaps were obtained.

2.3 Data Evaluation

For each pair of clones retained, the probability (using Barnett's [1] work) and the message length of each overlap (for each of the languages) was calculated. In order to compare the how well the MML criterion estimated the probability, the overlaps (for a clone pair) were sorted in decreasing order by probability. Each overlap was assigned a sequence number corresponding its location in the ordering. The data were then sorted in increasing order by message length. Figure 3 illustrates an example.

Figure 3: The ordering of overlap data.

ML	P
210	.87
251	.53
202	.98
230	.01
245	.65
215	.23

→ Sort on P →

ML	P	SN
202	.98	1
210	.87	2
245	.65	3
251	.53	4
215	.23	5
230	.01	6

→ Sort on ML →

ML	P	SN
202	.98	1
210	.87	2
215	.23	5
230	.01	6
245	.65	3
251	.53	4

2.3.1 Correlation Coefficient

The correlation coefficient (r) provides a quantitative measure of the dependence between two variables [6]. The value of **r** must be between -1 and +1. A negative correlation indicates that as the values of one variable increase, the values for the other variable tend to decrease. Conversely, a positive correlation indicates that as one variable increases the values for the other variable also tend to increase. The closer **r** is to ±1, the more closely the variables are related. If **r** equals ±1, then the value of one variable can be predicted exactly from the other. A value of zero indicates no relationship between the variables.

For each pair of clones, the correlation coefficient was calculated and used to measure the relationship between a overlap's message length and its probability. Since the MML criterion states that as the probability of an overlap increases, its message length decreases, the correlation coefficient of these variables should approach -1.

2.3.2 Evaluation of the Ordering

The correlation coefficient provides a general perspective on the relationship between message length and probability, but it does not given insight into the ordering relationship between individual message lengths and probabilities. The question under investigation is whether a particular language encodes overlaps in such a manner as to adhere to the MML criterion and thus order the overlaps in a manner consistent to that produced by a probabilistic analysis. Thus, it is not desired to be able to predict the probability of an overlap based on its message length, but instead given several overlaps to be able to discern the most probable overlaps. The following metrics were developed to quantify the correctness of an ordering.

2.3.2.1 Ordering Statistic

If after the sort on the message lengths the sequence numbers were still in a strictly increasing order, the MML criterion correctly predicted the relative probabilities of each overlap. If this was not the case, the number of misplaced values in the order was counted. This number is defined as being the minimum number of values which need to be moved to make the ordering correct. For example, the following order has two misplaced values (5 and 6 or equivalently 3 and 4).

$$1\,2\,5\,6\,3\,4 \Rightarrow_5 1\,2\,6\,3\,4\,5 \Rightarrow_6 1\,2\,3\,4\,5\,6$$

The ordering statistic quantifies the quality of an ordering as the percentage of misplaced values, and is defined as follows:

$$\frac{\text{Number of misplaced values}}{\text{Number of values - 1}} \tag{8}$$

Thus, the ordering statistic will have a value ranging between zero and one. A value of zero indicates that no mistakes were made in the ordering, while a value of 1 indicates that every mistake possible was made in the ordering.

2.3.2.2 Distance Statistic

The distance statistic provides a second metric to describe the quality of an order by indicating how far, on average, each value is misplaced. If **n** is the number of values in an order, the distance statistic is defined as:

$$\frac{\sum_{i=1}^{n} \text{abs}(\text{sequence\_number}_i - i)}{n} \tag{9}$$

Thus, large values of the distance statistic indicate that sever errors exist in an ordering.

2.3.2.3 Uncertainty Statistic

In general, the probabilities of two different overlaps will not be the same. This property does not necessarily hold for message length values. Since the data are first sorted by probability, two overlaps with equal message lengths will always be placed in the correct relative order. In the absence of the probability data, this is an additional source of error. Therefore, languages which encode several different overlaps into messages of the same length will have artificially good ordering and distance statistics. The uncertainty statistic is designed to identify this situation.

The uncertainty statistic provides a measure of the uniqueness of the message lengths of correctly placed overlaps. An overlap is defined to be uncertain if there exists another overlap with an equal message length such that their probabilities are not equal and both overlaps have been correctly ordered. The uncertainty statistic is defined as:

$$\frac{\text{Number of uncertain overlaps}}{\text{Number of correctly placed overlaps}} \tag{10}$$

The uncertainty statistic has a range of [0,1). A value of zero indicates that each correctly placed overlap has a unique message length and the order would be correct even without knowledge of the probabilities. As the value increases, the number of possible additional errors in the order also increases. Thus, this metric serves as a flag to warn that the ordering and distance statistics may be misleading. Languages that have very good ordering and distance statistics, but have an uncertainty statistic which approaches 1, are able to partition the feasible overlaps. Each block of the partition, referred to as a group, contains overlaps with identical message lengths. As the ordering and distance statistics indicate, there are very few overlaps which are placed in the wrong group, but without the probability data there is no way to correctly order the overlaps within each group. In Figure 4, uncertain overlaps are indicated by a "?" and misplaced overlaps by a "*".

Figure 4: An example illustrating the uncertainty within an order.

ML	P	SN	
202	.98	1	
210	.87	2	
210	.65	3?	Uncertainty = 2/5 = 0.4
210	.53	4?	
215	.01	6*	
230	.23	5	

3. Language Description and Evaluation

A number of different languages used to encode the overlap of two clones have been investigated. A description of each languages is provided below. For each language, the ordering, distance, uncertainty, and correlation statistics for the sample of 100 clone pairs are presented in tabular form. The table contains the minimum, maximum, and average values and the standard deviation for each statistic. In addition to the table, the percentage of the orderings which contain no misplaced overlaps is given. Also, the average location (rounded) in the sequence of the first misplaced overlap is provided. A clone pair has been selected as an example to illustrate the performance of each language. For each language, a figure similar to Figure 4 shows the ordering produced by applying the MML criterion to the feasible overlaps of the selected clone pair.

Each language is listed in the order it was developed. It is hoped that this will provide insight into some of the key concepts of each language. Figure 5 provides an example overlap which is used to illustrate many of the concepts discussed. The left hand side of the figure illustrates the overlap between two clones. Based on this overlap, the right hand side illustrates the implied actual lengths of the fragments in the original DNA molecule.

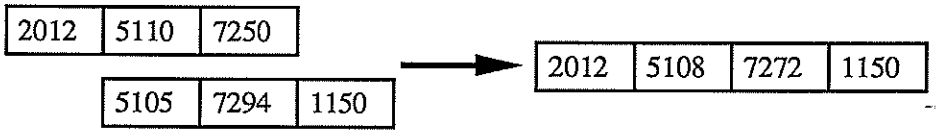

Figure 5: The overlap of two clones and the implied actual fragment lengths.

Messages are implemented as strings of characters. The length of a message, denoted by the function ML(), is simply the number of characters in the message. In order for the length of a message to adhere more closely to the MML criterion, numeric values are expressed in binary. When represented in base 10, the numeric values 11 and 99 are encoded with the same number of characters (2). When represented in base 2, 11 (1011) requires 4 characters and 99 (1100011) requires 7 characters. Thus, the message length more accurately represents the semantics of the numeric symbols within a message, when those symbols are from a binary alphabet.

3.1 Reconstruction

The idea of the reconstruction language is to create a message for each overlap from which the original data may be reconstructed. A overlap is encoded in the following manner. Unmatched fragments are encoded as <fragment length>. Matched fragments are encoded

as <ave_fragment_length,offset1,offset2>, where the offsets are defined as the difference between the length of the fragment from the appropriate clone and the ave_fragment_length. The overlap of Figure 5 would be encoded as (numeric values are represented in base 10 to improve the readability):

<2012><5108,-2,+3><7272,+22,-22><1150>

Note that this work deals with the overlap between only two clones and thus, both offsets are equal, or differ by 1. However, when the overlap between more than two clones is considered, the offsets will not necessarily be identical.

Table 1: Performance of the reconstruction language.

0% of the orderings contained no errors.
On average, the 2nd of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering Distance	0.200000	0.875000	0.522552	0.026735
Uncertainty	0.333333	11.120000	2.858606	3.591480
	0.000000	0.875000	0.273030	0.059911
Correlation	-0.831816	0.948892	-0.234548	0.124075

Figure 6: Ordering of the overlaps for the example clone pair by the reconstruction language.

ML	P	SN
2.510000e+02	6.016591e-03	5*
2.520000e+02	5.892324e-01	1
2.530000e+02	7.201576e-03	3
2.560000e+02	6.685164e-07	9*
2.570000e+02	6.596663e-03	4
2.570000e+02	3.860438e-03	7*
2.580000e+02	3.780705e-01	2*
2.590000e+02	5.304813e-03	6
2.610000e+02	3.715861e-03	8
2.620000e+02	4.289416e-07	10

As Table 1 indicates this method performs poorly. One of the primary factors in the probability calculations is the number of fragments which are matched in the overlap. The more fragments matched, the higher the probability. In this language, the length used to encode the ave_fragment_length and the two offsets is almost equivalent to that required to encode both of the original fragments lengths. Thus, a match does not significantly reduce the message length and therefore the corresponding poor results are obtained.

3.2 Hybrid

The hybrid language attempts to address the short-comings of the reconstruction language. The basic idea of encoding the length of each fragment and the amount by which this length

changes is retained. Instead of encoding the change in length of both original fragments only the change of one is encoded. Thus, the overlap of Figure 5 is encoded as:

<2012><5108,+2><7272,-22><1150>

Table 2: Performance of the hybrid language.

8% of the orderings contained no errors.
On average, the 4th of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering	0.000000	0.833333	0.348319	0.033209
Distance	0.000000	8.720000	1.741030	2.443042
Uncertainty	0.000000	0.937500	0.334062	0.057170
Correlation	-0.996255	0.876356	-0.783164	0.153448

Figure 7: Ordering of the overlaps for the example clone pair by the hybrid language.

ML	P	SN
2.310000e+02	5.892324e-01	1
2.340000e+02	3.780705e-01	2
2.370000e+02	6.016591e-03	5*
2.380000e+02	7.201576e-03	3
2.400000e+02	6.596663e-03	4
2.400000e+02	3.860438e-03	7?
2.400000e+02	6.685164e-07	9*
2.410000e+02	5.304813e-03	6*
2.420000e+02	3.715861e-03	8
2.430000e+02	4.289416e-07	10

The results from Table 2 show a dramatic increase in performance over the reconstruction language. Each matched fragment reduces the message length, while the offset term differentiates between matches. As the difference in the length of matched fragments decreases, so does the space required to encode the offset, thus producing shorter messages for better matches. Although this language provides an improvement, the orderings produced contain too many errors and have too many uncertain elements. Improvements are still needed.

3.3 Hybrid Sum

Many of the messages produced by the hybrid method are of the same length (33% by the uncertainty statistic). Since the numeric values of a message are encoded in binary, a more precise message length may be obtained if these values are encoded in unary. If **n** is the number of fragments and **m** is the number of matched fragments, the message length calculated by the hybrid sum is:

$$\sum_{i=1}^{n} \text{fragment\_length}_i + \sum_{j=1}^{m} \text{abs}(\text{offset}_j) \tag{11}$$

The hybrid sum message length for the data of Figure 5 is:

$$(2012 + 5108 + 7272 + 1150) + (2+22)$$

Table 3: Performance of the hybrid sum language.

57% of the orderings contained no errors.
On average, the 4th of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering	0.000000	0.545455	0.106783	0.024506
Distance	0.000000	3.920000	0.342417	0.409490
Uncertainty	0.000000	0.500000	0.057441	0.016833
Correlation	-0.967291	-0.376870	-0.677272	0.017980

Figure 8: Ordering of the overlaps for the example clone pair by the hybrid sum language.

ML	P	SN
3.023700e+04	5.892324e-01	1
3.024900e+04	3.780705e-01	2
3.175800e+04	7.201576e-03	3
3.177000e+04	6.596663e-03	4
3.179200e+04	5.304813e-03	6
3.179200e+04	3.715861e-03	8*
3.198600e+04	6.016591e-03	5*
3.199800e+04	3.860438e-03	7
3.433100e+04	6.685164e-07	9
3.434300e+04	4.289416e-07	10

The results in Table 3 show that the representation of numeric values in unary greatly improves the resultant ordering. Note that the correlation coefficient has decreased from that of the hybrid language. Since hybrid messages create many messages of the same length, the value of the length function is constant at these points. For virtually all overlaps, the value of the length function for the hybrid sum is unique. It is easier to predict the value of a variable that is constant than one which is not and thus, the correlation coefficient is greater for the hybrid language.

3.4 Match Information Content

From (4), the information content of a symbol is the negative logarithm of the probability of that symbol occurring. Thus, the information contained within a symbol increases as the frequency with which that symbol occurs decreases. This concept may be applied to overlap data if the length of a fragment is abstracted to a symbol from an alphabet. The match information content language addresses errors due to the non-uniqueness of the fragment length data. If an overlap contains **m** matching fragments, the average length of the $\mathbf{j^{th}}$ pair of matching fragments is $\mathbf{fragment\_length_j}$, and the probability of a fragment of length $\mathbf{fragment\_length_j}$ is given by (2), the match information content is:

$$\sum_{j=1}^{m} (\log \quad P(\text{fragment\_length}_j)) \tag{12}$$

The negative sign is dropped so that as the information content increases, the value decreases. The information content for the data of Figure 5 is:

$$\log (f(5108)) + \log(f(7272))$$

Table 4: Performance of the match information content language.

51% of the orderings contained no errors.
On average, the 3rd of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering	0.000000	0.636364	0.156924	0.039647
Distance	0.000000	5.120000	0.473022	0.624153
Uncertainty	0.000000	0.538462	0.030276	0.009803
Correlation	-0.999961	-0.633936	-0.942682	0.007779

Figure 9: Ordering of the overlaps for the example clone pair by the match information content language.

ML	P	SN
-3.936289e+01	5.892324e-01	1
-3.936289e+01	3.780705e-01	2?
-2.956538e+01	7.201576e-03	3
-2.956495e+01	6.596663e-03	4
-2.956434e+01	3.715861e-03	8*
-2.956392e+01	5.304813e-03	6
-2.955061e+01	6.016591e-03	5*
-2.955061e+01	3.860438e-03	7
-2.940876e+01	6.685164e-07	9
-2.940876e+01	4.289416e-07	10?

3.5 Compaction

The idea of the compaction language is that the overlaps which result in the most compact restriction map are most likely to produce the correct map. Thus, this language only encodes the average length of each fragment of the overlapping clones. The data of Figure 5 is encoded as:

<2012><5108><7272><1150>

Table 5: Performance of the compaction language.

93% of the orderings contained no errors.
On average, the 4th of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering	0.000000	0.200000	0.008458	0.001169
Distance	0.000000	0.666667	0.022278	0.009622
Uncertainty	0.000000	0.920000	0.460856	0.063465
Correlation	-0.999180	-0.738038	-0.960211	0.002412

Figure 10: Ordering of the overlaps for the example clone pair by the compaction language.

ML	P	SN
2.100000e+02	5.892324e-01	1
2.100000e+02	3.780705e-01	2?
2.230000e+02	7.201576e-03	3
2.230000e+02	6.596663e-03	4?
2.230000e+02	6.016591e-03	5?
2.230000e+02	5.304813e-03	6?
2.230000e+02	3.860438e-03	7?
2.230000e+02	3.715861e-03	8?
2.240000e+02	6.685164e-07	9
2.240000e+02	4.289416e-07	10?

Notice that Table 5 shows that 93 percent of the orderings contained no misplaced overlaps. This is due to the high uncertainty statistic and the very good ordering and distance statistics. This language is able to correctly group the overlaps, but does not differentiate between overlaps within the same group. Variations of this language that tweak the message lengths to remove the uncertainty may provide excellent results.

3.6 Compaction Sum

The compaction sum language is based on the same principle as the hybrid sum language. In order to increase the precision of the message length, numerical values are represented in unary instead of binary. Let **n** be the number of fragments. If the $\mathbf{i^{th}}$ fragment is in a region of overlap then $\mathbf{fragment\_length_i}$ is the average length of the matched fragments, otherwise, $\mathbf{fragment\_length_i}$ is equal to the length of the original fragment. The length of the resulting message is calculated as:

$$\sum_{i=1}^{n} \text{fragment\_length}_i \qquad (13)$$

The length of the message created for the data of Figure 5 is:

$$2012 + 5108 + 7272 + 1150$$

Table 6: Performance of the compaction sum language.

51% of the orderings contained no errors.
On average, the 3rd of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering	0.000000	0.636364	0.144886	0.033965
Distance	0.000000	4.880000	0.442022	0.561367
Uncertainty	0.000000	0.571429	0.058008	0.019707
Correlation	-0.968036	-0.376626	-0.676730	0.017885

Figure 11: Ordering of the overlaps for the example clone pair by the compaction sum language.

ML	P	SN
3.019600e+04	5.892324e-01	1
3.019600e+04	3.780705e-01	2?
3.172800e+04	7.201576e-03	3
3.173500e+04	6.596663e-03	4
3.174500e+04	3.715861e-03	8*
3.175200e+04	5.304813e-03	6
3.196950e+04	6.016591e-03	5*
3.196950e+04	3.860438e-03	7
3.429450e+04	6.685164e-07	9
3.429450e+04	4.289416e-07	10?

The results of Table 6 show that this language is successful in reducing the uncertainty. It is interesting to note that the results are not as good as those for the hybrid sum language. The only differences between the languages is the inclusion of the offset data in the hybrid sum language. This confirms the importance of including the difference in fragment lengths between the original fragments and the implied actual fragment length in the message length.

3.7 Compaction with Coefficient of Variation

The coefficient of variation (CV) is defined to be the ratio of the standard deviation to the arithmetic mean [6]. The compaction with coefficient of variation languages uses the CV to measure the amount by which the size of a matched fragment changes due to an overlap. If an overlap contains **m** matching fragments, and the $\mathbf{j^{th}}$ matching fragment pair has an average length $\mathbf{fragment\_length_j}$ and a standard deviation $\mathbf{s_j}$, the message length is:

$$ML(\text{compaction}) + \sum_{j=1}^{m} \frac{\sigma_j}{\text{fragment length}_j} \qquad (14)$$

The length of the message created for the data of Figure 5 is:

$$ML(<2012><5108><7272><1150>) + \frac{2.5}{5108} + \frac{22}{7272}$$

Table 7: Performance of the compaction with CV language.

39% of the orderings contained no errors.
On average, the 4th of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering Distance	0.000000	0.578947	0.165042	0.025732
Uncertainty	0.000000	3.200000	0.517621	0.415925
	0.000000	0.333333	0.013563	0.003206
Correlation	-0.999139	-0.738782	-0.960888	0.002327

Figure 12: Ordering of the overlaps for the example clone pair by the compaction with CV language.

ML	P	SN
2.100032e+02	5.892324e-01	1
2.100043e+02	3.780705e-01	2
2.230013e+02	6.016591e-03	5*
2.230022e+02	7.201576e-03	3
2.230024e+02	3.860438e-03	7*
2.230027e+02	6.596663e-03	4
2.230031e+02	5.304813e-03	6
2.230038e+02	3.715861e-03	8
2.240031e+02	6.685164e-07	9
2.240042e+02	4.289416e-07	10

The results, listed in Table 7, show that this language reduces the uncertainty of the compaction language. The fact that this measure compares favorably to the hybrid language illustrates the importance of normalizing the change in fragment length. Based on the improvement between the hybrid and the hybrid sum languages, the combination of compaction sum and the coefficient of variation should improve these results further.

3.8 Compaction with Information Content

The final metric used to reduce the uncertainty of the compaction language is to consider the information content of the matched fragments in the overlap. If an overlap contains **m** matching fragments, the average length of the **j**th pair of matching fragments is **fragment_length**$_j$, and the probability of a fragment of length **fragment_length**$_j$ is given by (2), the length of a compaction with information content message is:

$$\text{ML(compaction)} + \sum_{j=1}^{m} \log(P(\text{fragment\_length}_j)) \tag{15}$$

The length of the message created for the data of Figure 5 is:

$$\text{ML(<2012><5108><7272><1150>)} + \log(f(5108)) + \log(f(7272))$$

Table 8: Performance of the compaction with information content language.

77% of the orderings contained no errors.
On average, the 5th of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering Distance	0.000000	0.500000	0.037487	0.007403
Uncertainty	0.000000	1.250000	0.108435	0.059261
	0.000000	0.880000	0.289090	0.072534
Correlation	-0.999178	-0.738050	-0.960198	0.002412

Figure 13: Ordering of the overlaps for the example clone pair by the compaction with information content language.

ML	P	SN
2.100213e+02	5.892324e-01	1
2.100213e+02	3.780705e-01	2?
2.230158e+02	7.201576e-03	3
2.230158e+02	6.596663e-03	4?
2.230158e+02	5.304813e-03	6?
2.230158e+02	3.715861e-03	8*
2.230159e+02	6.016591e-03	5*
2.230159e+02	3.860438e-03	7
2.240166e+02	6.685164e-07	9
2.240166e+02	4.289416e-07	10?

The results in table 8 show that this metric reduces the average uncertainty of the compaction language by almost 20 percentage points, while increasing the average number of misplaced overlaps by only 3 percentage points. The uncertainty is still high, but a possible solution would consist of using the compaction sum and information content to calculate the message length.

3.9 Mutual Overlap Statistic

The mutual overlap statistic (MOS) estimates the quality of an overlap by the proportion of fragments which are matched. If a clone of $\mathbf{n_1}$ fragments overlaps a clone of $\mathbf{n_2}$ fragments and there are $\mathbf{m}$ matching fragments, the MOS value is:

$$\frac{m^2}{n_1 * n_2} \tag{16}$$

This value is negated to provide the desired property that more likely overlaps have smaller values. The MOS value for the data of Figure 5 is:

$$-\frac{2^2}{3 * 3}$$

Table 9: Performance of the MOS language.

100% of the orderings contained no errors.
There were no elements misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering Distance	0.000000	0.000000	0.000000	0.000000
Uncertainty	0.000000	0.000000	0.000000	0.000000
	0.600000	0.960000	0.757881	0.013069
Correlation	-0.999976	0.000000	-0.868867	0.095067

Figure 14: Ordering of the overlaps for the example clone pair by the MOS language.

ML	P	SN
-1.600000e-01	5.892324e-01	1
-1.600000e-01	3.780705e-01	2?
-9.000000e-02	7.201576e-03	3
-9.000000e-02	6.596663e-03	4?
-9.000000e-02	6.016591e-03	5?
-9.000000e-02	5.304813e-03	6?
-9.000000e-02	3.860438e-03	7?
-9.000000e-02	3.715861e-03	8?
-9.000000e-02	6.685164e-07	9?
-9.000000e-02	4.289416e-07	10?

The MOS message length value did not produce any errors in the ordering, but the uncertainty is very high. Table 9 shows that more than 75 percent of the overlaps have the same message length. Variations of this language follow to reduce the uncertainty.

3.10 MOS of Message Lengths

In order to improve the performance of the MOS language, the two overlapping clones and the implied segment of the DNA molecule are encoded into a compaction message and the MOS is calculated based on message lengths, not on the number of fragments. If $\mathbf{m_{c1}}$ and $\mathbf{m_{c2}}$ are the messages representing the clones and $\mathbf{m_{dna}}$ is the message of the implied DNA segment, the MOS of message lengths value is:

$$-\frac{ML(m_{dna})^2}{ML(m_{c1}) * ML(m_{c2})} \qquad (17)$$

The MOS of message lengths value for the data of Figure 5 is:

$$\frac{ML(<2012><5108><7272><1150>)^2}{ML(<2012><5110><7250>) * ML(<5105><7294><1150>)}$$

Table 10: Performance of the MOS of message lengths language.

92% of the orderings contained no errors.
On average, the 4^{th} of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering Distance	0.000000	0.454545	0.013004	0.003158
Uncertainty	0.000000	2.333333	0.045611	0.063017
	0.000000	0.920000	0.459020	0.062577
Correlation	-0.999503	-0.742605	-0.961745	0.002317

Figure 15: Ordering of the overlaps for the example clone pair by the MOS of message lengths language.

ML	P	SN
-2.096571e-01	5.892324e-01	1
-2.096571e-01	3.780705e-01	2?
-1.440911e-01	7.201576e-03	3
-1.438197e-01	6.596663e-03	4
-1.434324e-01	3.715861e-03	8*
-1.431615e-01	5.304813e-03	6
-1.348740e-01	6.016591e-03	5*
-1.348740e-01	3.860438e-03	7
-6.172116e-02	6.685164e-07	9
-6.172116e-02	4.289416e-07	10?

As Table 10 indicates, this metric reduced the uncertainty of the MOS language by 30 percentage points while only increasing the number of misplaced elements by 1 percentage point. The uncertainty is still very high and a finer metric is needed.

3.11 MOS of Fragment Lengths

As with other languages, when greater precision is required it is useful to encode the numeric symbols of a message in unary instead of binary. Let **n1** and **n2** be the number of fragments from each clone and **m** be the number of matching fragments. For the matching fragments, **fragment_length** is the average length of the matched fragments, otherwise, **fragment_length** is equal to the length of the original fragment. The MOS of fragment lengths value is:

$$\frac{\left(\sum_{j=1}^{m} fragment\_length_j\right)^2}{\sum_{i=1}^{n1} fragment\_length_i * \sum_{k=1}^{n2} fragment\_length_k} \qquad (18)$$

Table 11: Performance of the MOS of fragment lengths language.

51% of the orderings contained no errors.
On average, the 3rd of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering Distance	0.000000	0.636364	0.144886	0.033965
Uncertainty	0.000000	4.880000	0.442022	0.561367
	0.000000	0.571429	0.058008	0.019707
Correlation	-0.974873	-0.394418	-0.718033	0.015532

Figure 16: Ordering of the overlaps for the example clone pair by the MOS of fragment lengths language.

ML	P	SN
-1.686524e-01	5.892324e-01	1
-1.686524e-01	3.780705e-01	2?
-9.722383e-02	7.201576e-03	3
-9.722383e-02	6.596663e-03	4?
-9.722383e-02	6.016591e-03	5?
-9.722383e-02	5.304813e-03	6?
-9.722383e-02	3.860438e-03	7?
-9.722383e-02	3.715861e-03	8?
-8.796992e-02	6.685164e-07	9
-8.796992e-02	4.289416e-07	10?

Table 11 reveals that this metric successfully reduces the uncertainty and produces results very similar to the compaction sum. This is due to the fact that the primary factor which effects the messages produced by both the compaction sum language and the MOS of fragment length language is the number of fragments matched by a overlap.

3.12 MOS of Information Content

Another method to remove uncertainty from an ordering is to include the information content of the matching fragments. Let **n1** and **n2** be the number of fragments from each clone and **m** be the number of matching fragments. For the matching fragments, **fragment_length** is the average length of the matched fragment pair, otherwise, **fragment_length** is equal to the length of the original fragment. The probability of a fragment of length **fragment_length** is given by (2). The MOS of information content value is:

$$\frac{\left(\sum_{j=1}^{m} \log(P(\text{fragment\_length}_j))\right)^2}{\sum_{i=1}^{n1} \log(P(\text{fragment\_length}_i)) * \sum_{k=1}^{n2} \log(P(\text{fragment\_length}_k))} \qquad (19)$$

Table 12: Performance of the MOS of information content language.

51% of the orderings contained no errors.
On average, the 3rd of 12 elements was the first to be misplaced.

Statistic	Min	Max	Ave	Std Dev
Ordering	0.000000	0.636364	0.157833	0.040277
Distance	0.000000	5.360000	0.477089	0.649303
Uncertainty	0.000000	0.538462	0.024736	0.008871
Correlation	-0.999962	-0.634434	-0.942625	0.007755

Figure 17: Ordering of the overlaps for the example clone pair by the MOS of information content language.

ML	P	SN
-1.605616e-01	5.892324e-01	1
-1.605616e-01	3.780705e-01	2?
-9.058050e-02	7.201576e-03	3
-9.057789e-02	6.596663e-03	4
-9.057414e-02	3.715861e-03	8*
-9.057153e-02	5.304813e-03	6
-9.049007e-02	6.016591e-03	5*
-9.049007e-02	3.860438e-03	7
-8.962330e-02	6.685164e-07	9
-8.962330e-02	4.289416e-07	10?

Notice from the results given in Table 12, that the uncertainty statistic for the MOS of information content is lower than it is for the compaction with information content. This result may be explained by the manner in which the fragment length data is encoded. The compaction with information content language represents fragment lengths in binary, while the MOS of information content uses the probability that a fragment of a given length occurs. The use of the probability produces a "finer" representation of the fragment length which in turn reduces the value of the uncertainty statistic.

4. Observations

4.1 Categorization of the Languages

The investigated languages may be placed into one of three categories based on the type of ordering they produce. The first category is comprised of those languages that do not perform well. These languages are characterized by very poor ordering, distance and uncertainty statistics and are summarized in Table 13.

Table 13: Summary of languages which perform poorly.

Heuristic Types	Ave Ordering	Ave Distance	Ave Uncert	Ave Corr	% Correct	1st Error
Reconstruction	0.522552	2.858606	0.273030	-0.234548	0	2
Hybrid	0.348319	1.074103	0.334062	-0.783164	8	4

The second category of languages is comprised of those languages that group overlaps correctly, but are not able to distinguish between different overlaps within the same group. These languages are characterized by having very good ordering and distance statistics, but a poor uncertainty statistic. These languages are useful for the gross categorization of overlaps. Table 14 summarizes the grouping languages.

Table 14: Summary of languages which group overlaps.

Heuristic Types	Ave Ordering	Ave Distance	Ave Uncert	Ave Corr	% Correct	1st Error
Compaction	0.008458	0.022278	0.460856	-0.960211	93	4
Compaction and Info Content	0.037487	0.108435	0.289090	-0.960198	77	5
Mutual Overlap Statistic	0.000000	0.000000	0.757881	-0.868867	100	-
MOS of Message Lengths	0.013004	0.045611	0.459020	-0.961745	92	4

The final category contains languages whose message lengths are unique for virtual all overlaps. As a result the uncertainty statistic for these languages is very good, but the ordering and distance statistics are slightly poorer than those of the grouping languages. These languages are able to differentiate overlaps on a fine level. Table 15 summarizes the fine languages.

Table 15: Summary of languages which order overlaps on a fine level.

Heuristic Types	Ave Ordering	Ave Distance	Ave Uncert	Ave Corr	% Correct	1st Error
Hybrid Sum	0.106783	0.342417	0.057441	-0.677272	57	4
Match Information Content	0.156924	0.473022	0.030267	-0.942682	51	3
Compaction Sum	0.144886	0.442022	0.058008	-0.676730	51	3
Compaction and Coeff of Var	0.165042	0.517621	0.013563	-0.960888	39	4
MOS of Fragment Lengths	0.144886	0.442022	0.058008	-0.718033	51	3
MOS of Information Content	0.157833	0.477089	0.024736	-0.942625	51	3

4.2 Similarity of the Sum and Information Content Languages.

Table 15 indicates that languages which use the sum of the fragment length data and those that use the information content of the fragment length data produce very similar orderings. In each case 51 percent of the orderings contained no misplaced overlaps. The other statistics are comparable as well. (Note that the hybrid sum language also contains information on the amount by which a match changes a

fragment's length and thus, 57 percent of these orderings were completely correct.) In general, large fragments occurs less frequently than small fragments. Thus as the size of a fragment increases, so does its information content. Therefore, instead of incurring the computational costs of calculating the information content, the use of the fragment length data alone provides equivalent results.

4.3 Dependency of Results and Language Choice

The only factors which effect the value of a MML language are the overlap, the given fragment data and the encoding language. On the other hand, several factors, such as measurement error, the size of the original DNA molecule, and the average clone size, affect the overlap probability. Varying any of these factors may change the probabilities and possibly the ordering of the overlaps, but the message length will remain constant. Thus a universal MML language will not be found. The aim of further research should be to discover a language which, by the MML criterion, correctly describes the fragment length data and the feasible overlaps between clones.

5. Further Work

This research has shown that a information theoretic based language shows promise in its ability to estimate the probability of clone overlap. MML languages are particularly successful in the categorization of overlaps into groups. The results also look favorable on a fine level, but further investigation is required to quantify this. The true test of effectiveness will be to compare the restriction maps created using a MML criterion versus those created using Bayesian probabilities in terms of accuracy and execution time. Furthermore, this study only considered the problem of evaluating the overlap between two clones. The results need to be extended to verify the effectiveness of these methods when considering the overlap between a portion of the map and a clone and also between two portions of the map.

6. Acknowledgements

This work was supported in part by grant T15 LM07049 from the National Library of Medicine and grant 87-24 from the James S. McDonnell Foundation.

References

[1] L. Barnett. Probabilistic analysis of random clone restriction mapping. Department of Computer Science, Washington University, Technical Report, WUCS-90-28.

[2] P. Cheeseman. On Finding the Most Probable Model. To appear in P. Langley and J. Shrugger, editors, *Methods fo Scientific Discovery.*

[3] K. Drlica. *Understanding DNA and Gene Cloning.* John Wiley & Sons, Inc., 1984.

[4] F. Ingels. *Information and Coding Theory.* International Textbook Company, 1971.

[5] M. Olson et al. Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Aca. Sci.*, 83:7826-7830, October 1986.

[6] B. Rosner. *Fundamentals of Biostatistics, 2nd edition.* Duxbury Press, 1986.