Washington University in St. Louis
Washington University Open Scholarship

All Computer Science and Engineering Research | Computer Science and Engineering

Report Number: WUCS-90-28

1990-12-01

Probabilistic Analysis of Random Clone Restriction Mapping

Laurie J. Barnett

Several current DNA mapping projects are based on detection of overlaps between cloned DNA molecules. This thesis places the problem of overlap detection in a probabilistic framework by deriving, for each of the relevant overlap types, expressions for the probability that a postulated overlap is correct. In addition, computationally feasible approximations for the probability expressions are developed. These expressions have been implemented and, using the implementations, the validity of both the original and the approximated probability expressions is verified.

Follow this and additional works at: https://openscholarship.wustl.edu/cse_research

Recommended Citation
Barnett, Laurie J., "Probabilistic Analysis of Random Clone Restriction Mapping" Report Number: WUCS-90-28 (1990). *All Computer Science and Engineering Research.*
https://openscholarship.wustl.edu/cse_research/703

Department of Computer Science & Engineering - Washington University in St. Louis
Campus Box 1045 - St. Louis, MO - 63130 - ph: (314) 935-6160.

PROBABILISTIC ANALYSIS OF RANDOM CLONE RESTRICTION MAPPING

Master's Thesis

Laurie J. Barnett

WUCS-90-28

December 1990

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

WASHINGTON UNIVERSITY
SEVER INSTITUTE OF TECHNOLOGY

ABSTRACT

PROBABILISTIC ANALYSIS OF RANDOM CLONE
RESTRICTION MAPPING

by Laurie J. Barnett

ADVISOR: Professor Will Gillett

December, 1990

Saint Louis, Missouri

Several current DNA mapping projects are based on detection of overlaps between cloned DNA molecules. This thesis places the problem of overlap detection in a probabilistic framework by deriving, for each of the relevant overlap types, expressions for the probability that a postulated overlap is correct. In addition, computationally feasible approximations for the probability expressions are developed. These expressions have been implemented and, using the implementations, the validity of both the original and the approximated probability expressions is verified.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Thesis Topic

Several current DNA mapping projects are based on detecting overlaps between cloned DNA molecules. This thesis derives expressions for the probability that a postulated overlap is correct, and addresses issues of computational feasibility and verification of the validity of these probability expressions.

1.2 Genetics Concepts and Terminology

1.2.1 Genetics Background

DNA is the genetic material which supplies the blueprint for an organism's development. A DNA molecule is composed of nucleotides, each nucleotide consisting of a sugar, a phosphate, and a "base." There are four types of bases, called A (Adenine), T (Thymine), C (Cytosine), and G (Guanine). Nucleotides are distinguished by the base they contain. Sugar-phosphate bonds can bind the nucleotides into strands. The bases on one strand can pair with bases on another strand by hydrogen bonding, however, the only base pairings allowed are A-T and G-C. Thus, A and T are complementary bases, as are G and C. A DNA molecule is made of two complementary nucleotide strands bound together by base pairing. The base sequence of one DNA strand determines the base sequence of the other because of the A-T, C-G complementation restrictions. For example, Figure 1 shows a schematic of a DNA molecule ten base pairs long, in which the sugar-phosphate bonds are represented

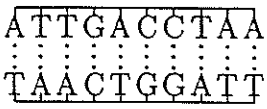


Figure 1: A Double-Stranded DNA Molecule. A schematic of a double-stranded DNA molecule ten base pairs in length is shown. The horizontal lines represent the sugar-phosphate backbones of the molecule, the letters A, T, G, and C represent the bases for each of the constituent nucleotides, and the hydrogen bonds between the complementary base pairs are represented by the dotted lines.

as the solid horizontal lines and the hydrogen bonds as the vertically aligned dots. A DNA molecule can be abstracted as a sequence of base pairs, with its size given by the number of base pairs.

DNA molecules can be cut into fragments using **restriction enzymes**. A restriction enzyme recognizes only one or a few specific short base pair sequences (usually between 4 and 6 base pairs long) called the **recognition sequence(s)** of the enzyme. The length of the recognition sequence will be referred to as the **sitesize.**[1] For the class of restriction enzymes generally used in the laboratory, there is one position within each recognition sequence at which the enzyme will cut. This position is called the **cutsite**. A **complete digest** of a DNA molecule consists of exposing the DNA to a restriction enzyme for enough time to allow cleavage of all of the enzyme's cutsites which occur in the molecule. A **partial digest** exposes the DNA for a shorter period of time so that only a randomly chosen fraction of the cutsites are actually cleaved.

The DNA fragments resulting from a digestion can be separated and sized using **agarose gel electrophoresis** [20]. DNA is negatively charged at normal pH; when placed in an agarose gel across which an electric current is maintained, a

[1] "Sitesize" is *not* a term in general use.

DNA molecule will migrate toward the anode. Gels usually contain several **lanes**; each lane generally contains a single DNA sample. A DNA molecule's speed of migration, called its **mobility**, is a linear function of the logarithm of the length of the molecule, such that the smaller molecules migrate more quickly than the larger molecules. Therefore, during a set time period of exposure to the electric current, the smaller molecules will move farther through the gel than the larger molecules, while molecules of identical lengths will co-migrate. If enough identical copies of a DNA molecule are present in a sample, then the co-migrating DNA can be seen as a **band** on the gel by using appropriate staining techniques. If a sample of DNA contains fragments of different sizes then several bands can be seen in the lane, each band corresponding to fragments of a particular length. For example, Figure 2 is a photograph of a digitized image of an agarose gel containing fourteen lanes. Counting from the top of the gel, the arrow in the figure is pointing to the 3^{rd} band in the 6^{th} lane. (The very first mark in each lane is the well in which the DNA sample was placed. Counting of bands begins at the top band, below the well.) The bands containing the largest DNA fragments are at the top of the gel, with bands containing progressively smaller DNA fragments located progressively nearer the bottom of the gel.

In general, a few of the gel lanes are used as standards and contain DNA fragments of known lengths. To measure the length of the fragments in a band, the band position is compared to band positions in the standard lanes. Inaccuracy in the measured length can result from variability in either the speed of migration of the fragments or in the measurement and comparison of the band positions.

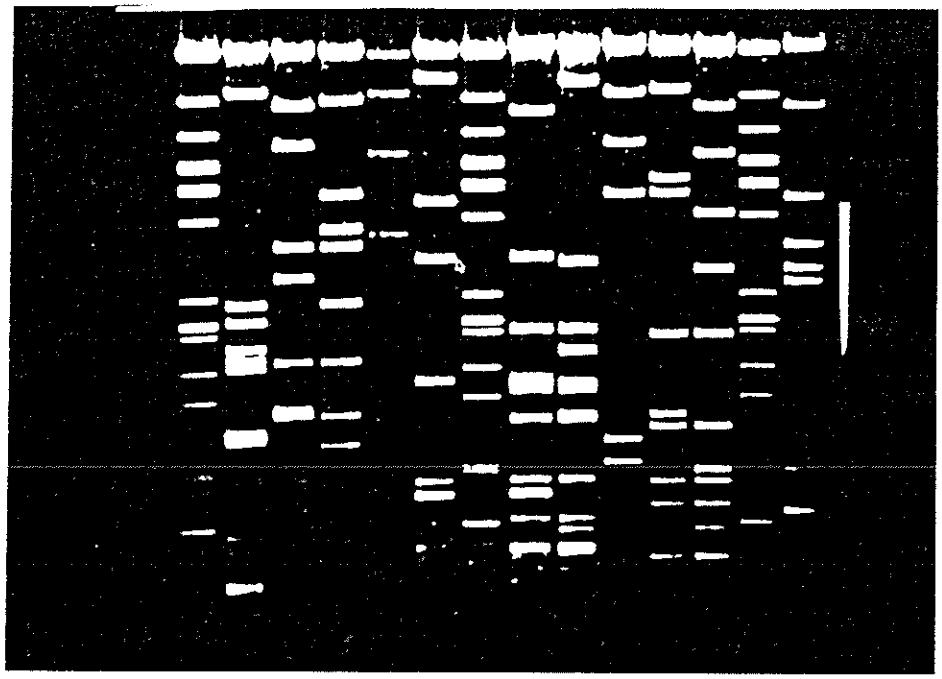


Figure 2: Photograph of a Digitized Image of Agarose Gel Electrophoresis. The fourteen lanes in the gel correspond to fourteen DNA samples.

Note that electrophoresis will generally *not* separate DNA molecules which differ in base pair sequence but not in length. Only DNA molecules of *different* lengths can be separated using agarose gel electrophoresis.

1.2.2 Map Building

A map is a representation of the DNA in terms of the distances between occurrences of particular "markers" along the molecule. In a **physical map**, the distance is in terms of the number of base pairs between the markers. The markers are usually defined in terms of the underlying base sequence of the DNA. For example, restriction enzyme cutsites are commonly used as markers, and the resultant physical map is called a **restriction map**. In contrast, **genetic linkage maps** are constructed by studies of inheritance patterns using family pedigrees: the distance between markers

is in terms of the **recombination distance**, measured in units based on the frequency with which crossovers occur between the markers during meiosis. In general, recombination distances only grossly approximate the number of base pairs between the markers [15].

Genome mapping refers to mapping *all* of an organism's DNA. One powerful approach to physical genome mapping is to create an ordered set of **overlapping random clones**. Genomic DNA is partially digested, and the resultant DNA segments are cloned.[2] Different clones which originated from the same genome location are said to **overlap**. (This definition of overlap implies that the overlapping sections of clones have the same base pair sequence.[3] However, the converse is *not* implied. If sections of DNA molecules have the same base pair sequence, it is not necessarily true that these sections actually overlap.) The clones are considered random because it is assumed that the cutsites which are cleaved during the partial digestion are randomly chosen. The clones can be ordered in terms of the genome locations from which they originated.

Figure 3 is a schematic of genomic DNA with an ordered set of overlapping clones beneath it. In the figure, the clones are drawn directly below the genome locations from which they originated, vertically aligning the overlapping sections

[2]A clone is defined as "a large number of cells or molecules all identical to some ancestral cell or molecule" [15]. To clone a DNA molecule, it is first inserted into a vector. The vector is a DNA molecule which possesses characteristics which allow it to be maintained as part of a host organism, for example as part of a bacterium or a virus. After the DNA is inserted into the vector, the "vector + new DNA" is inserted into the host organism. Many identical copies of the original DNA molecule (the clone of the DNA) are obtained by allowing the host to multiply. The clone can be separated from the host and vector DNA using standard laboratory procedures. Cloning is necessary to obtain enough copies of a piece of DNA for laboratory procedures such as gel electrophoresis.

[3]This assumes no mutations occured during the cloning process.

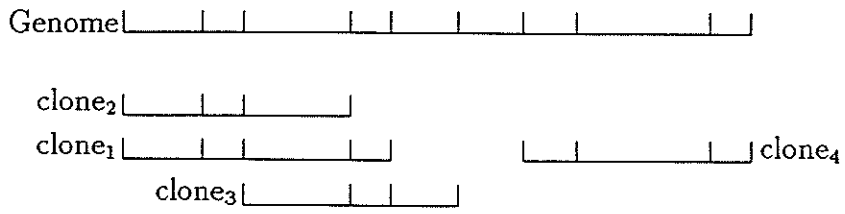


Figure 3: Genomic DNA and a Set of Overlapping Random Clones. Genomic DNA is represented by the top horizontal line, with each smaller horizontal line beneath representing one clone. Cutsites are marked by vertical lines. Overlapping sections of clones are vertically aligned.

of the clones. The **depth** of overlap for a section of the genome is the number of clones which overlap that section. Clearly, the depth can vary in different genome locations. For example, in Figure 3 there are overlap depths of zero, one, two, and three. A genome is considered to be **x% covered** by a given set of clones when x% of the genome is overlapped by at least one of the clones. For example, in Figure 3 the genome is approximately 90% covered.

The clone order is lost during the digestion/cloning process, but can be recovered by detection of clone overlaps. The clones are **fingerprinted**, and overlaps are inferred when there are significant similarities in the fingerprints. Fingerprinting refers to a laboratory procedure which can produce characteristic data for each clone. The measured lengths of the fragments resulting from restriction enzyme digestion [5, 13, 18], the base pair sequence of the ends of the fragments resulting from restriction enzyme digestion [2], and probe hybridization patterns[4] [9, 16] are examples of different types of fingerprints. When the fingerprint consists of the

[4]A probe is a short single-stranded piece of DNA. In probe hybridization fingerprinting, the clone is separated into single strands and mixed with the probe. The probe will bind (hybridize) to the clone when it is complementary to a subsequence of the clone DNA. The probe hybridization sites are the markers along the DNA.

measured lengths of the fragments which result from a complete digest of a clone, the fingerprints are considered similar when they possess fragments of the same measured lengths, within experimental error.

Physical genome mapping using overlapping random clones has been carried out or is in progress for several different organisms including the bacterium *E. coli* [13, 25], the yeast *S. cerevesiae* [18], the nematode *C. elegans* [5, 24], and man [4, 3, 9, 21]. A major goal of these physical genome mapping projects is to integrate the physical maps with known genetic linkage maps. This integration will result in a more accurate representation of the relative locations of genes and the overall organization of the genome than is currently available in genetic linkage maps alone. In addition, it will allow ready access to the DNA of a new gene and its flanking sequences[5] if a closely linked gene can be identified. One method which has been used to align genetic and physical maps involves comparing the base pair sequences of the genes in the genetic linkage map[6] with the recognition sequences in the physical map [19].

1.3 Statement of the Problem

1.3.1 Single Enzyme Digest Fingerprinting

This thesis addresses the fingerprint methodology used by Olson et al. in mapping the genome of *S. cerevesiae* [18]. The fingerprint consists of the sizes of the

[5] Flanking sequences are important for studies on gene regulation and genome organization.

[6] Many genes in the linkage maps have been either partially or completely sequenced.

fragments[7] resulting from a complete single enzyme digestion of a clone[8], as measured by agarose gel electrophoresis. A postulated overlap between two clones consists of **fragment pairings**, where fragments that are paired are postulated to overlap and fragments that are not paired are postulated not to overlap. Because fragments which actually do overlap *must* have the same true length, the only fragment pairings which should be postulated are those between fragments for which, within experimental error, the measured lengths are the same. It is possible for two fragments which do not actually overlap to have the same true length and/or the same measured length, within experimental error. Therefore, a postulated overlap between two clones should only be inferred to be correct if it includes several fragment pairings.

Each inferred overlap imposes a partial order on the fragments of the clones involved. For the purposes of this thesis, a partial order will be viewed as a sequence of sets. The fragments within each set are unordered, but the sets themselves are ordered. Let { } denote a set and [ ] denote a sequence. As an example, let clone_1 contain fragments with measured lengths of 300, 800, 1200, 4300, 6200 and clone_2 contain fragments with measured lengths of 303, 650, 1206, 4290. The fragment

[7]Although the term "fragment" is generally applicable to any DNA molecule resulting from cleavage of a larger molecule, this thesis will constrain its use specifically to those DNA molecules resulting from a *complete* digestion.

[8]In practice, more than one enzyme may actually be used to digest the clone in order to increase the likelihood of a cutsite occurrence. However, the fragments resulting from the digestion are *not* distinguished with respect to the enzyme used, and so the cutsites of the different enzymes are also *not* distinguished. Essentially, the multiple enzymes are treated as a single unit in terms of data analysis. For ease of notation, subsequent discussions will refer to this type of fingerprinting as a "single enzyme digest" because the fragments from the multiple enzymes are not distinguished with respect to the enzyme. The term "multiple enzyme digest" will be reserved for situations when the fragments resulting from the different enzymes *are* distinguished with respect to the enzyme.

pairings of 300-303, 1200-1206, and 4300-4290 will impose one of the two following partial orders on the fragments:

1. [{800, 6200}, {300-303, 1200-1206, 4300-4290}, {650}],
2. [{650}, {300-303, 1200-1206, 4300-4290}, {800, 6200}].

These partial orders are based on the fact that all of the fragments from one clone must be contiguous in the genome – there can be no "gaps" between the fragments of a clone. The only partial orders which are consistent with this requirement place the overlapping fragments in the middle set, with the unpaired fragments from one clone in the first set and the unpaired fragments from the other clone in the last set.

A map (formally defined in Section 2.1.3) is built by repetitively inferring clone overlaps using the fragment length fingerprints of each clone. This process is illustrated in Figure 4.[9] The measured fragment lengths for each of the four clones are shown in Figure 4(a). Inferred overlap_1, between clone_3 and clone_4, results in map_1 as shown in Figure 4(b). Inferred overlap_2, between map_1 and clone_2, results in map_2 as shown in Figure 4(c). Inferred overlap_3, between map_2 and clone_1, results in map_3 of Figure 4(d).

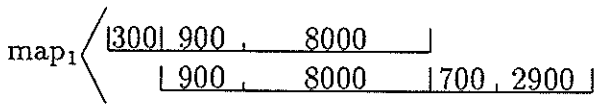
As shown in the figure, a map may contain *multiple* fragments which overlap each other. For example, in map_3 there are four fragments with measured lengths of 900 which have been inferred to overlap each other, and there are three fragments of length 8000 which have also been inferred to overlap each other. In addition, inferred overlaps may extend a map and/or partition sets already in the map into

[9] For simplicity, the inferred overlaps in the figure are based on very few fragment pairings, with the fragment pairings consisting of fragments with identical measured lengths. In practice, the measured fragment lengths would not be identical and more fragment pairings would be required before an overlap inference was made.

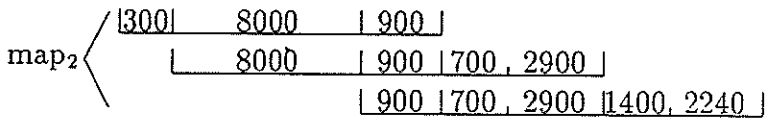
(a) Restriction Fragment Lengths for Four Clones:

$clone_1$: {300, 600, 900, 1260, 2900, 4200, 8000}
$clone_2$: {700, 900, 1400, 2240, 2900}
$clone_3$: {300, 900, 8000}
$clone_4$: {700, 900, 2900, 8000}

(b) Inferred $overlap_1$, between $clone_3$ and $clone_4$. Results in map_1:
[{300}, {900-900, 8000-8000}, {700, 2900}].



(c) Inferred $overlap_2$, between map_1 and $clone_2$. Results in map_2:
[{300}, {8000-8000}, {900-900-900}, {700-700, 2900-2900}, {1400, 2240}].



(d) Inferred $overlap_3$, between map_2 and $clone_1$. Results in map_3:
[{600, 1260, 4200}, {300-300}, {8000-8000-8000}, {900-900-900-900},
{2900-2900-2900}, {700-700}, {1400, 2240}].

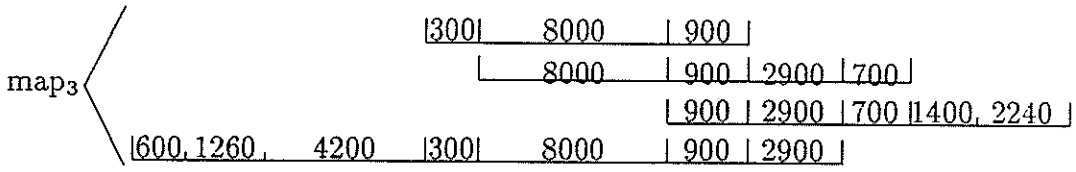


Figure 4: Example of Map Building Based on Inferring Overlaps. Each horizontal line represents a clone, with fragments separated by vertical lines. Each fragment's measured length is written above the fragment. The *ordered* sequence of sets is denoted by the tall vertical lines, and the *unordered* fragments within each set are denoted by the short vertical lines. For notational convenience, the unordered fragments within each set are written in ascending order with respect to their measured lengths. The paired fragments within each set are vertically aligned.

ordered subsets. For example, $overlap_3$ partitions the set in map_2 which contains the fragments of lengths 700 and 2900 into two ordered subsets, the first containing the fragment of length 2900 and the second containing the fragment of length 700. Also, for any map there are two possible left-right **orientations** of the map with respect to the genome, because the partial order present in the map distinguishes between the "left" and "right" ends of the map. Figure 4 only shows one of the two possible orientations for each map. For example, the alternate orientation for map_2 would be [{1400, 2240}, {700-700, 2900-2900}, {900-900-900}, {8000-8000}, {300}]. Finally, note that there are many other overlap inferences which could have been made between the clones shown in Figure 4. If other overlaps had been inferred, then a different map would have resulted.

To place the map-building process in perspective with respect to genome mapping projects, the genomic DNA of the yeast *S. cerevesiae* is approximately 1.5×10^7 base pairs long. In the mapping project conducted by Olson et al., a random clone library of approximately 5000 clones was used for which the average clone size was 13,700 base pairs. The average number of fragments per clone was 8.4. [18]. Thus, the library covered the genome to a depth of 4.5. The human genome mapping project is even more complex, in large part due to the 100-fold increase in the genome size; the human genome is approximately 3.0×10^9 base pairs long.

1.3.2 The Problem: Probabilistic Assessment of Postulated Overlaps

In order to use fingerprint comparisons to infer overlaps, rules must be established for quantifying fingerprint similarity and judging its significance. These rules must

be stringent enough to minimize the possibility of a false positive overlap inference (incorrectly inferring that an overlap exists when it actually does not). However, a more stringent rule will, in general, require a bigger overlapping region before an overlap can be inferred. Many postulated overlaps, although actually correct, may not contain a large enough overlap region to be inferred as correct by the more stringent rules. This results in false negative overlap inferences (inferring that a postulated overlap is wrong when it actually is correct). Both false positive and false negative errors cause problems in map building. False positives result in maps that are incorrect, while false negatives result in maps that are incomplete. In addition, high false negative rates have been shown to substantially slow the progress of mapping projects [14].

Therefore, it is important that the rules used for assessing fingerprint similarity limit the number of both false positive and false negative inferences. The rules should also provide an accurate assessment of the risk that a false positive overlap inference may occur, because this will allow evaluation of the resultant map's quality in terms of the amount of error it is likely to contain.

The rules used by Olson et al. are based on statistics involving the size of the postulated overlap region (in terms of the number of paired fragments) and the differences in the measured lengths of the paired fragments [18]. These metrics are not as sensitive as desired, and there is no way to accurately predict the expected frequency of false positive inferences when using them. In addition, setting the statistical thresholds for inferring overlaps can be based only on empiric grounds.

This thesis proposes that basing the rules for inferring overlaps on a precise

probabilistic model of the sources of error in the laboratory methodology should result in a more sensitive metric than the statistics used by Olson et al. In addition, the thresholds for overlap inference could be based directly on a known, acceptable risk of false positives.[10] In this way, finer control of the false positive and false negative rates could be established; by choosing the largest acceptable false positive rate, the resultant false negative rate would be as small as possible within the limits of the laboratory methodology.

Accordingly, the problem that this thesis addresses is to derive expressions for the probability that a postulated overlap is correct, to ensure that the expressions are computationally feasible, and to verify that the expressions are valid.

1.4 Thesis Content and Organization

This thesis is organized into five chapters. This introduction is Chapter One. Chapter Two contains the problem analysis. The types of overlaps are categorized, probabilistic models are developed for the sources of error in the laboratory methodology, and the basic probabilistic approach is described.

Chapter Three contains the derivations for the probability expressions for each type of overlap. Probability expressions are first derived for a simplified model of the sources of error in the laboratory methodology, and then extended to a general model. Computationally feasible approximations to the probability expressions are also developed in this chapter.

[10]For example, if a false positive rate of two errors per one hundred overlap inferences is acceptable, then a threshold could be used in which a postulated overlap must have at least a 98% probability of being correct.

Chapter Four contains the verification of the probability expressions. Using simulated data, the accuracy of the approximations is demonstrated, the expected changes in probabilities paralleling changes in various biologic parameters are documented, and it is verified that the probabilities do accurately predict the percentage of postulated overlaps which are actually correct.

Chapter Five is the conclusion. It contains a summary of the work done in this thesis. In addition, a discussion is given of the expressions' validity in various environments, their usefulness for the particular fingerprinting methodology of Olson et al., and their importance to genome mapping research in general. The conclusion ends with the future directions in which this work could be pursued.

CHAPTER 2

Problem Analysis

There are three components to the problem of deriving expressions for the probability that a postulated overlap is correct given the measured fragment lengths. First, the relevant types of overlaps must be defined, because it will be necessary to derive probability expressions for each overlap type. Second, because the probability expressions should be based on the sources of error in the laboratory methodology, these sources of error must be modelled probabilistically. Third, the basic approach to be used in deriving the probability expressions must be determined. These three components are addressed in this chapter.

2.1 Hierarchy of Overlap Types

There are four relevant types of overlaps. They can be categorized into the following hierarchy: **fragment-fragment overlap**, **clone-clone overlap**, **clone-map overlap**, and **map-map overlap**. These overlap types will be defined in the following sections, with examples based on Figure 5. The figure shows overlap relations which are correct, based on the genomic DNA shown in (a). In the following discussion, a postulated overlap is considered incorrect when it is inconsistent with the biological reality shown in Figure 5. The measured lengths of the fragments in Figure 5 have been abstracted to letters.

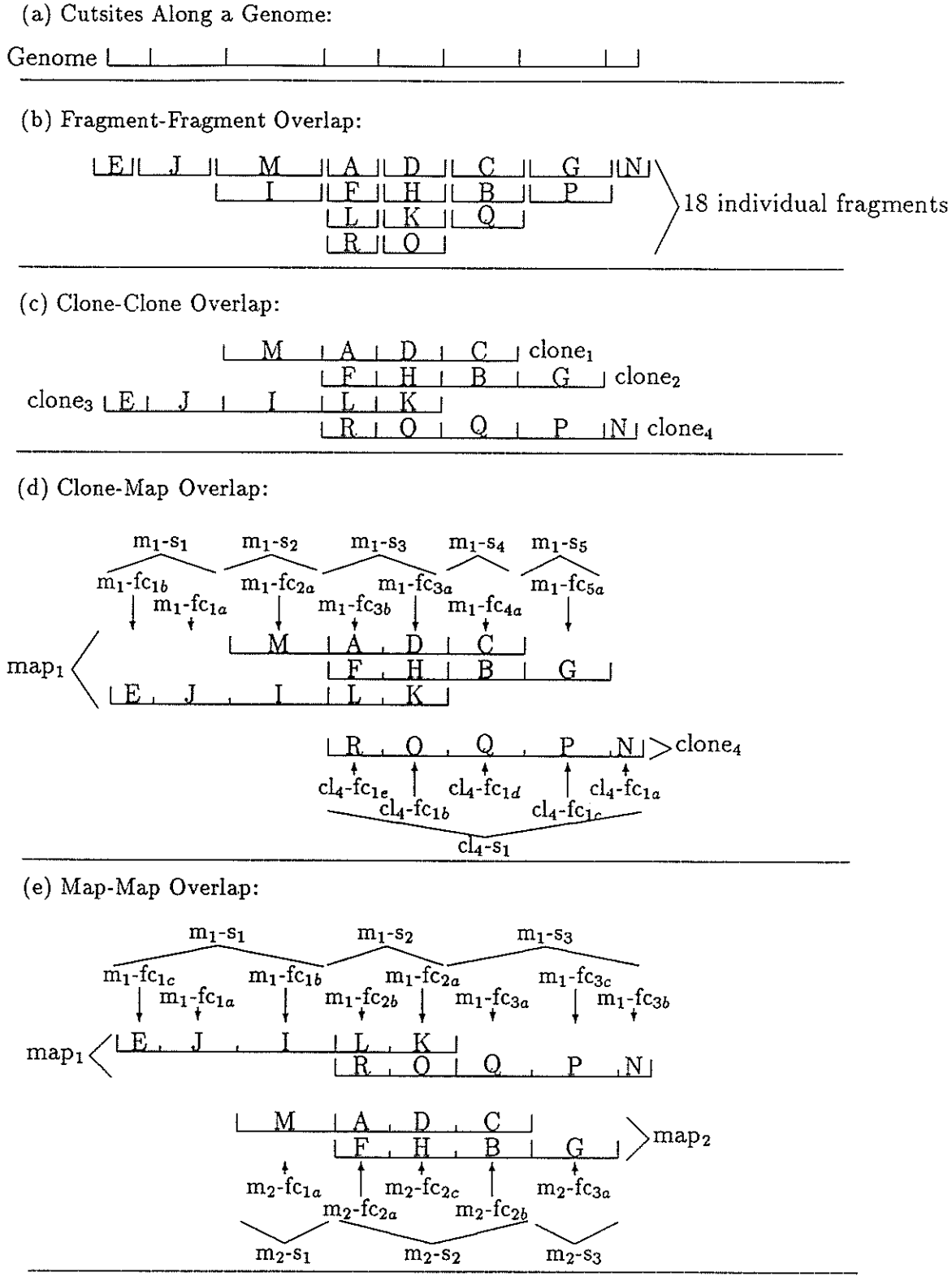


Figure 5: The Four Overlap Types. Discussion of the figure is given in the text. "fc"=fragment column, "s"=set, "m"=map, "cl"=clone.

Table 1: Fragment Data Available to the Researcher for the Fragment-Fragment Overlap Type. Fragment data shown in the table correspond to Figure 5(b).

Fragment-Fragment Overlap Type																	
Eighteen Randomly Chosen Fragments																	
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R

2.1.1 Fragment-Fragment Overlap

Assume the genomic DNA shown in Figure 5(a) is completely digested. Out of the resultant pool of fragments, let Figure 5(b) represent eighteen randomly chosen fragments. In Figure 5(b), the correct fragment ordering is shown by the vertical alignment of each fragment under the portion of the genome from which it originated. However, this ordering information is not available to the researcher because it is lost during the digestion. Table 1 shows the fragment data as the researcher would see it, as a list of the measured fragment lengths with no ordering information.[11]

For any two of the fragments, there are exactly two possibilities which can be postulated; either the two fragments do overlap or they do not overlap. For example, using Table 1 it can be postulated that fragments F and G overlap, that D and K overlap, or that M and I do not overlap. According to Figure 5(b), of these three different proposals only the second is actually correct.

The term **f-f overlap** will be used as an abbreviation for fragment-fragment overlap.

[11]For notational convenience, the fragments in the table are listed in alphabetical order.

Table 2: Fragment Data Available to the Researcher for the Clone-Clone Overlap Type. Fragment data shown in the table correspond to Figure 5(c).

Clone-Clone Overlap Type																
Four Randomly Chosen Clones																
$Clone_1$				$Clone_2$				$Clone_3$					$Clone_4$			
A	C	D	M	B	F	G	H	E	I	J	K	L	N	O	P	Q	R

2.1.2 Clone-Clone Overlap

Assume the genomic DNA shown in Figure 5(a) is partially digested and cloned. Out of the resultant set of random clones, let Figure 5(c) represent four fingerprinted clones with the vertical lines delimiting the fragments within the clones. The correct ordering of both the clones and the fragments within the clones, as shown in Figure 5(c), is not available to the researcher because it is lost during the digestion and cloning process. Table 2 shows the fragment data as the researcher would see it, as a list of clones and, for each clone, a list of fragments.[12]

A postulated clone-clone overlap consists of three sets: a set of constituent f-f overlaps and, for each clone, a set of fragments which do not overlap the other clone.[13] Many different overlaps can be postulated between any two clones. For example, using the information in Table 2, two different overlaps which could be postulated between $clone_1$ and $clone_3$ are:

$overlap_1$: {A-K, C-I}, {D, M}, {E, J, L},
$overlap_2$: {A-L, D-K, M-I}, {C}, {E, J}.

$Overlap_2$ is a *different* overlap than $overlap_1$ – the two sets of constituent f-f overlaps are not the same. According to Figure 5(c), $overlap_2$ is actually correct.

[12]For notational convenience, the clones are shown in numerical order and the fragments within the clones are shown in alphabetical order.

[13]Any, but not all, of these sets may be empty.

Let $overlap_p$ be a postulated clone-clone overlap. A **suboverlap** of $overlap_p$ is defined as a clone-clone overlap for which the set of constituent f-f overlaps is a subset of $overlap_p$'s set of constituent f-f overlaps. For example, if $overlap_3$ between $clone_1$ and $clone_3$ is defined as {A-L, D-K}, {C, M}, {E, I, J}, then $overlap_3$ is a suboverlap of $overlap_2$. According to Figure 5(c), $overlap_3$ is incorrect but $overlap_2$ is correct. For the clone-clone overlap type, the **null overlap** is the overlap for which the set of constituent f-f overlaps is empty.

The clone-clone overlap type reduces to the fragment-fragment type when each clone consists of only one fragment.

2.1.3 Clone-Map Overlap

Assume the genomic DNA shown in Figure 5(a) is partially digested and cloned, and let Figure 5(c) represent four randomly chosen clones. In addition, assume overlaps have been inferred between $clone_1$, $clone_2$, and $clone_3$, resulting in map_1 of Figure 5(d). Map_1 and $clone_4$ of Figure 5(d) will be used to illustrate the clone-map overlap type.

Before discussing clone-map overlaps, a formal definition of a **map** is needed. For this thesis, a map is defined as a partial order of **fragment columns (fc)**. A fragment column is defined as a set of fragments, each of which is inferred to overlap all of the other fragments in the set. In the schematics of maps which have been given (see Figures 4 and 5(d)), a fragment column is a set of vertically aligned fragments. The cardinality of a fragment column is defined as the number of fragments in the column. For example, in Figure 5(d), map_1 contains seven fragment columns, each

of which has a cardinality of either one, two, or three. The partial order of fragment columns in a map can be represented as a sequence of sets of fragment columns. The cardinality of a set in a map is defined as the number of fragment columns it contains (not as the number of fragments it contains). In Figure 5(d), the j^{th} set of the i^{th} map is denoted by "m_i-s_j," and column k of set j in map i is denoted by "m_i-fc_{jk}." Because a map implies no ordering information of the fragments within each set, k denotes a letter, not a number. Using this notation, map_1 contains a sequence of five sets, set m_1-s_3, which contains columns m_1-fc_{3b} and m_1-fc_{3a}, has a cardinality of two, and column m_1-fc_{3b} has a cardinality of three. In Figure 5(d) the unordered fragment columns are delimited by the short vertical lines, and the ordered sets are delimited by the tall vertical lines.

A clone can be viewed as a map which consists of only one set, with all fragment columns having cardinality one. Figure 5(d) shows $clone_4$ from this viewpoint.

The correct ordering of the fragment columns within each set and of the clone with respect to the map, and the correct orientation of the map, as shown in Figure 5(d), are not available to the researcher. Only the inferred order of the sets and the fragments within each column would be known. Table 3 shows the fragment data and the ordering information as the researcher would see it.[14] Note that because the orientation of the map with respect to the genome is not known to the researcher, the sets of map_1 in the table could also have been listed in the alternate orientation, as [m_1-s_5, m_1-s_4, ..., m_1-s_1].

[14]For notational convenience, both the fragment columns within each set and the fragments within each column are listed in alphabetical order.

Table 3: Fragment Data Available to the Researcher for the Clone-Map Overlap Type. Fragment data shown in the table correspond to Figure 5(d). "fc"=fragment column, "cl"=clone, "s"=set, "m"=map.

Clone-Map Overlap Type						
Map_1						
m_1-s_1		m_1-s_2	m_1-s_3		m_1-s_4	m_1-s_5
m_1-fc_{1a}	m_1-fc_{1b}	m_1-fc_{2a}	m_1-fc_{3a}	m_1-fc_{3b}	m_1-fc_{4a}	m_1-fc_{5a}
J	E	M I	D H K	A F L	C B	G

$Clone_4$				
cl_4-s_1				
cl_4-fc_{1a}	cl_4-fc_{1b}	cl_4-fc_{1c}	cl_4-fc_{1d}	cl_4-fc_{1e}
N	O	P	Q	R

A postulated clone-map overlap consists of three sets: a set of constituent **fragment-fragment column (f-fc) overlaps**, a set of fragments from the clone which do not overlap the map, and a set of columns from the map which do not overlap the clone. All overlaps which can be postulated between a clone and a map must be consistent with the partial order already present in the map. For example, let there be a postulated clone-map overlap for which at least one fragment in the clone does not overlap the map, and let a column in the i^{th} fragment set of the map be postulated to overlap the clone. Then, *all* of the fragment columns in either sets $1 \ldots i-1$ or $i+1 \ldots n$, where n is the number of sets in the map, must also be postulated to overlap the clone. Using the data in Table 3, the overlap of

{O–m_1-fc_{3b}, Q–m_1-fc_{4a}, P–m_1-fc_{5a}}, {N, R},
{m_1-fc_{1a}, m_1-fc_{1b}, m_1-fc_{2a}, m_1-fc_{3a}}

can be postulated. However, the overlap of

{O–m_1-fc_{3b}, Q–m_1-fc_{4a}}, {N, P, R},
{m_1-fc_{1a}, m_1-fc_{1b}, m_1-fc_{2a}, m_1-fc_{3a}, m_1-fc_{5a}}

cannot be postulated because either m_1-fc_{5a} or all three of the columns m_1-fc_{1b}, m_1-fc_{1a} and m_1-fc_{2a} must also be postulated to overlap the clone. Otherwise, either the

contiguity of the fragments within $clone_4$ or the partial ordering in map_1 would be violated.

Let $overlap_p$ be a postulated clone-map overlap. A **suboverlap** of $overlap_p$ is defined as an overlap for which the set of constituent f-fc overlaps is a subset of $overlap_p$'s set of f-fc overlaps. For the clone-map overlap type, the **null overlap** is defined as the overlap for which the set of constituent f-fc overlaps is empty.

The clone-map overlap type reduces to the clone-clone type when the map contains only one clone.

2.1.4 Map-Map Overlap

Assume $clone_1$, $clone_2$, $clone_3$, and $clone_4$ are obtained by random selection of clones created from partial digestion of the genomic DNA of Figure 5(a). Assume the overlaps shown in Figure 5(e), between $clone_3$ and $clone_4$ resulting in map_1, and between $clone_1$ and $clone_2$ resulting in map_2, have been inferred. Figure 5(e) shows the correct ordering of the maps and the columns within each set of each map, and the correct orientation of both maps. Table 4 shows the fragment data and ordering information which is actually available to the researcher.[15] Again, no orientation information for either map would be available; for either of the maps in the table, the sets could also have been listed in the alternate orientation, [m_1-s_3, m_1-s_2, m_1-s_1] and [m_2-s_3, m_2-s_2, m_2-s_1] for map_1 and map_2, respectively.

[15]For notational convenience, both the fragment columns within each set and the fragments within each column are listed in alphabetical order.

Table 4: Fragment Data Available to the Researcher for the Map-Map Overlap Type. Fragment data correspond to Figure 5(e). "fc"=fragment column, "s"=set, "m"=map.

Map-Map Overlap Type							
Map_1							
m_1-s_1			m_1-s_2		m_1-s_3		
m_1-fc_{1a}	m_1-fc_{1b}	m_1-fc_{1c}	m_1-fc_{2a}	m_1-fc_{2b}	m_1-$4fc_{3a}$	m_1-fc_{3b}	m_1-fc_{3c}
J	I	E	K O	L R	Q	N	P

Map_2				
m_2-s_1	m_2-s_2			m_2-s_3
m_2-fc_{1a}	m_2-fc_{2a}	m_2-fc_{2b}	m_2-fc_{2c}	m_2-fc_{3a}
M	A F	C B	D H	G

A postulated map-map overlap consists of three sets: a set of **fragment column-fragment column (fc-fc) overlaps** and, for each map, a set of fragment columns which do not overlap the other map. All overlaps postulated between two maps must be consistent with the partial order already present in *each* map. For example, the map-map overlap of

$\{m_1\text{-}fc_{1c}\text{–}m_2\text{-}fc_{1a}, m_1\text{-}fc_{2b}\text{–}m_2\text{-}fc_{2a}, m_1\text{-}fc_{2a}\text{–}m_2\text{-}fc_{2c}, m_1\text{-}fc_{3c}\text{–}m_2\text{-}fc_{2b}, m_1\text{-}fc_{3a}\text{–}m_2\text{-}fc_{3a}\}, \{m_1\text{-}fc_{1a}, m_1\text{-}fc_{1b}, m_1\text{-}fc_{3b}\}, \{\}$

can be postulated. However, the overlap of

$\{m_1\text{-}fc_{3b}\text{–}m_2\text{-}fc_{1a}, m_1\text{-}fc_{2b}\text{–}m_2\text{-}fc_{2a}, m_1\text{-}fc_{2a}\text{–}m_2\text{-}fc_{2c}, m_1\text{-}fc_{3c}\text{–}m_2\text{-}fc_{2b}, m_1\text{-}fc_{3a}\text{–}m_2\text{-}fc_{3a}\}, \{m_1\text{-}fc_{1a}, m_1\text{-}fc_{1b}, m_1\text{-}fc_{1c}\}, \{\}$

cannot be postulated because if m_2-fc_{1a} were to overlap m_1-fc_{3b} in addition to the other fc-fc overlaps specified, the partial order of one of the maps would have to be violated. In addition, particular orientations of the two maps relative to each other are usually imposed when a map-map overlap is inferred.

The definition for a map-map **suboverlap** is the same as that for a clone-map suboverlap except that constituent fc-fc overlaps are of concern, rather than the f-fc overlaps of the clone-map case. Also, for the map-map overlap type, the null

overlap is defined as that overlap for which the set of constituent fc-fc overlaps is null.

The map-map overlap type reduces to the clone-map type when one of the maps contains only one clone.

Comments on overlap notation In subsequent sections, clone-clone, clone-map, and map-map overlaps will be defined by explicitly stating only the set containing all of the constituent f-f, f-fc, and fc-fc overlaps, respectively. Because the two remaining sets of non-overlapping fragments and columns are uniquely determined by the explicitly stated set, they may be left implicit.

2.2 Sources of Error to Model Probabilistically

There are three main sources of error in evaluating fingerprint similarities based on the single enzyme digest laboratory methodology: **random size measurement error**, **correlated size measurement error**, and **fragment length multiplicity**. Probabilistic models of these sources of error will be needed to derive the probability expressions. Accordingly, this section discusses the probabilistic models which will be used.

2.2.1 Random Size Measurement Error

Agarose gels have limited resolution, thus there is error associated with determining the precise position of a band on a gel. This results in error in the measured fragment lengths, because measured length is based on the band position. These

errors are termed random because the error associated with a band is assumed to be independent of the error associated with any other band. In practice, the resultant error in the measured length of large fragments tends to be larger than that of smaller fragments. This occurs because fragment mobility on a gel is a linear function of the logarithm of the fragment length.[16] Accordingly, this thesis models random measurement error by assuming it is proportional to the true fragment length.

Define the **percent random error** as the proportionality constant which determines the range in which the majority of random measurement errors will fall. Specifically, if there is a $k\%$ random error and there is no other type of measurement error, then the majority of measured lengths of a fragment of true length X should fall within $X \pm \frac{kX}{100}$. Subsequent derivations assume that the same percent random error is applied to all fragments.[17] For example, if the random error was 1%, there was no other form of measurement error, and the true lengths of a clone's fragments were 400, 980, 2300, 3600, then the majority of measurements of these fragments should result in lengths within 400 ± 4, 980 ± 10, 2300 ± 23, 3600 ± 36.

---

[16]The log-linear relationship results in less distance on the gel between large fragments than between small fragments. For example, fragments with true lengths between 6500 and 6768 may occupy 1 millimeter on a gel, but fragments with true lengths between 400 and 415 may also occupy 1 millimeter on a gel [8]. (In [8], measurements are in terms of pixels. A conversion of 4.88 pixels per mm [7] was used to convert the pixels into millimeters.) If the band position of a fragment of length 6500 is mismeasured by .5 millimeters, then the resultant measured length of the fragment would be off by 134 base pairs. If a band position of a fragment of length 400 is mismeasured by .5 millimeters, then the resultant measured length of the fragment would be off by only 7 base pairs. The measured length of the larger fragment (6500) is associated with a greater measurement error than that of the smaller fragment (400), even the both band positions were mismeasured by the same amount.

[17]In practice, the percent random error associated with small fragments is larger than that associated with other fragments. However, because the percent random error is a constant in the subsequent probability expression derivations, use of a different value for smaller fragments will not invalidate the derivations. Therefore, this assumption can be made without loss of generality, but with increased notational convenience.

A smaller percent random error implies increased accuracy in the measured lengths of the fragments. The percent random error will vary between laboratories.

In this work, the random error is modelled using a probability function g, where $g(Y \mid X)$ is the probability that the measured length (ml) is Y given that the true length (tl) is X,

$$P(ml = Y \mid tl = X) \stackrel{\text{def}}{=} g(Y \mid X).$$

Further, g is assumed to be a discretized normal distribution with mean X (μ_x) and standard deviation proportional to X (σ_x). A discrete function is used because fragment lengths are integers. A normal distribution is used because a unimodal, two-parameter distribution is needed (to parameterize both the fragment true length and the percent random error), and, in addition, the normal distribution possesses certain mathematical properties which will allow subsequent approximations to be made (see Sections 3.4.2 and 3.4.3). One recent publication does claim to have experimentally verified that random measurement error is normally distributed [1].

The normal distribution is discretized by integration between $Y - .5$ and $Y + .5$. Therefore,

$$g(Y \mid X) \stackrel{\text{def}}{=} \int_{y=Y-.5}^{Y+.5} \mathcal{N}(y \mid \mu_x = X, \sigma_x)dy = \int_{y=Y-.5}^{Y+.5} \frac{1}{\sqrt{2\pi}\sigma_x} e^{\frac{-1}{2\sigma_x^2}(y-X)^2} dy.$$

For this thesis, σ_x is defined such that $\mu_x \pm 2\sigma_x = X \pm \frac{kX}{100}$, where k is the percent random error. In other words, $\sigma_x \stackrel{\text{def}}{=} \frac{kX}{2\times 100}$. For example, if there is a 1% random error and a fragment's true length is 5000, then σ_x is 25. By defining σ_x in this way,

95% of all of the random measurement errors will fall within the range defined by the percent random error.

With this model, the probability of obtaining a measured length of Y increases as Y approaches the true length X.

2.2.2 Correlated Size Measurement Error

An additional source of measurement error is the correlated size measurement error. Conditions leading to correlated measurement error include unintentional non-uniformity in electrophoretic conditions, such as varying gel temperature, or in sample conditions, such as varying DNA or salt concentration. In these situations, *all* of the fragment mobilities in a lane may be simultaneously affected, resulting in atypical band positions. This error is considered correlated because the amount of error associated with one band position in a lane is related to the error associated with the other band positions in the same lane.

The fragment mobilities may either all increase, in which case all of the bands in the lane move faster than usual, or they may all decrease, in which case all of the bands in the lane move slower than usual. In either case, all of the band positions in the lane will appear "shifted" in the same direction. It is also possible for the large fragments to move faster and the small fragments slower, in which case the lane appears "compressed," or for the large fragments to move slower and the small fragments faster, in which case the lane appears "stretched." Figure 6 shows these four types of correlated errors individually. It is also possible for a lane to appear both shifted and compressed, or shifted and stretched.

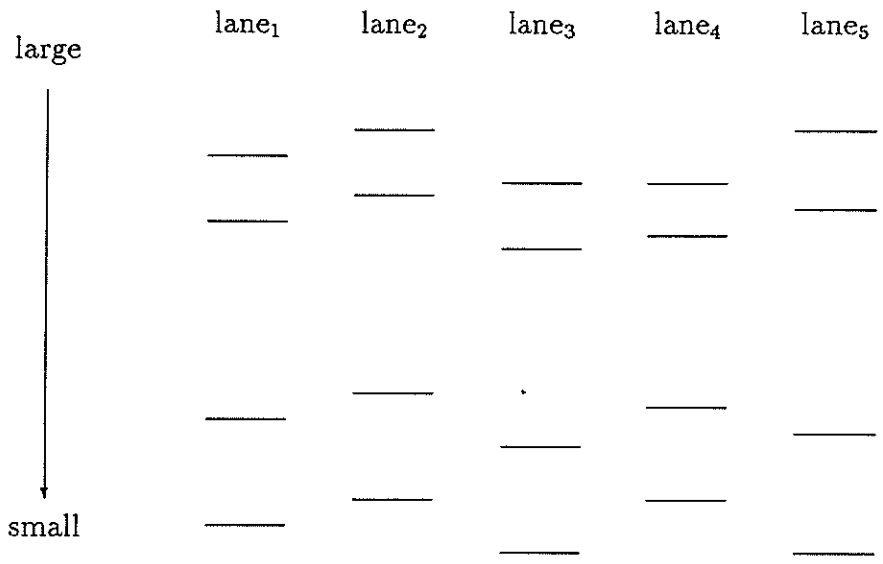


Figure 6: Examples of Correlated Measurement Error. Assume that the same DNA sample was used for each of the five lanes. All of the bands in $lane_1$ have moved at the typical speed, all of the bands in $lane_2$ have moved slower (shifting the lane up), all of the bands in $lane_3$ have moved faster (shifting the lane down), in $lane_4$ the small bands have moved slower and the large bands faster (compressing the lane), and in $lane_5$ the small bands have moved faster and the large bands slower (stretching the lane).

In practice, each lane in a gel contains DNA from one clone. Therefore, the correlated error in a gel lane will affect the measured lengths of the fragments in only one clone, but it will affect *all* of the measured fragment lengths in that clone.

Olson et al. have modelled the correlated error with a two-parameter model, in which $Y_j = \alpha(X_j)^\beta$ for $1 \leq j \leq r$, where r is the number of fragments in a clone, X_j is the true length of fragment j in that clone, and Y_j is its measured length. The parameter α models the shifting component, while β models the stretch/compression component [18]. All of the fragments from one clone are assumed to be associated with the same α and β, but α and β for different clones can be different.

In this work, a method was not found for applying the two-parameter model to the probability expressions in a manner which would allow certain mathematical approximations necessary to achieve computational feasibility (see Section 3.4.3). Therefore, a simplification of the two-parameter model to a one-parameter model for which $Y_j = \alpha X_j$ for $1 \leq j \leq r$ was used. Recall that the mobility of a fragment is a linear function of its length. Because $\log(\alpha Y) = \log\alpha + \log Y$, this simplified version does model the additive constant for the band position of a fragment, which is the shifting component of the correlated error, but it does not model the multiplicative constant, which is the stretch and compression component of the correlated error. Therefore, this one-parameter model assumes that a lane may be shifted up or down, but that it will not be stretched or compressed. In this model all of the fragments from one clone are assumed to be associated with the same value of α, but different clones may have different α values.

To verify that this one-parameter model is reasonable, it has been compared

to the two-parameter model using real data consisting of 81 overlapping clones comprising a map of five columns, where each column had a depth of 81. (The clones overlapped exactly. For each clone, every fragment overlapped a fragment from each of the remaining 80 clones.) The comparisions were done as follows. For each column, the true length of the fragments within the column was estimated by the mean of the 81 measured lengths. Then, for each of the 81 clones, the effect of the correlated error on the five measured lengths was minimized using the two-parameter model by determining the α and β values that would optimize the least squares fit between the measured lengths in the clone and the paired estimated true lengths.[18] The clone's measured length data were then adjusted using these α and β. For each of the 81 clones, the effect of the correlated error was also minimized using the one-parameter model by determining the α which would optimize the probability of obtaining the clone's measured length data given the pairing between the clone's measured lengths and the estimated true fragment lengths (Section 3.2 describes this probability, which is called $P_{d|oc}$, and Section 3.4.3 describes the method of determining the α which optimizes $P_{d|oc}$.). The clone's measured length data were then adjusted using this α.

The models were assessed in two ways. First, for each of the two models the variability in the *adjusted* measured lengths for each of the five columns was determined. The models were compared on the basis of the reduction in this variability. Table 5 shows the results. In the table, the mean of the unadjusted measured lengths for each of the five columns is shown, and for each of the models the variability in the

[18]This was the methodology described in [18].

Table 5: Comparison of the Variability Reduction from the One-Parameter Model to that from the Two Parameter Model for the Correlated Measurement Error. "sd"=standard deviation, "fc"=fragment column.

Variability in the Adjusted Measured Lengths for the One-Parameter and the Two-Parameter Correlated Error Models					
	fc_1	fc_2	fc_3	fc_4	fc_5
Mean Measured Length	2912.6	2267.5	1762.4	670.0	297.3
Sd of Unadjusted Data	21.8	22.4	14.1	6.0	4.8
Sd of Adjusted Data Using 2-Param Model	14.2	11.8	8.2	3.3	2.5
Sd of Adjusted Data Using 1-Param Model	14.4	15.7	8.8	4.6	3.4

Table 6: Comparison of the Reduction in Measured Length Differences from the One-Parameter Model to that from the Two Parameter Model for the Correlated Measurement Error. "ml diff"=average difference in measured lengths, "fc"=fragment column.

Average Difference in Measured Lengths Between the True Length and the Adjusted Length for the One-Parameter and the Two-Parameter Correlated Error Models					
	fc_1	fc_2	fc_3	fc_4	fc_5
Mean Measured Length	2912.6	2267.5	1762.4	670.0	297.3
Ml Diff of Unadjusted Data	16.7	16.3	11.2	4.5	3.9
Ml Diff of Adjusted Data Using 2-Param Model	10.3	8.0	6.8	2.7	2.0
Ml Diff of Adjusted Data Using 1-Param Model	11.5	10.7	7.3	3.6	2.7

columns' adjusted measured lengths is shown. For comparison, the variability in the unadjusted (original) measured lengths for each column is also shown.

Second, for each fragment in each column the difference between the adjusted measured length and the estimated true length was calculated, and for each column the mean of these differences was determined. The models were compared on the basis of the magnitude of the reduction in these means. Table 6 shows the results.

Table 5 demonstrates that, for each of the columns, the one-parameter model substantially reduced the variability in the adjusted measured lengths, relative to the unadjusted measured lengths. Table 6 demonstrates that, for each column, the one-parameter model also substantially reduced the mean difference between the adjusted measured lengths and the true length, relative to the mean difference for the unadjusted measured lengths. As expected, both tables show that the two-parameter model can further reduce both the variability and the mean difference in the measured lengths, fairly substantially in particular instances (fc_2 in Table 5 and fc_2 in Table 6). However, the reduction by the one-parameter model is substantial enough to indicate that it is a reasonable method with which to model correlated error, in view of the mathematical difficulties which have precluded the use of the two-parameter model.

The one parameter model will be used as follows. Let r be the number of fragments in a clone, and X_j and Y_j be the true and measured length, respectively, of fragment j in the clone. Using the one parameter model, the correlated error associated with a lane can be *corrected* by multiplying each measured length from the clone by the same **shifting coefficient** c, so that $X_j = c \times Y_j$ for $1 \leq j \leq r$. **Shifting** a clone's measured fragment lengths will be defined as multiplication of every fragment length of the clone by the same value of the shifting coefficient, c. Although all fragments from one clone are associated with the same value of c, the value of c may be different for different clones. If $c = 1$, then no correlated error occurred. If $c < 1$, then the clone's fragments had moved slowly relative to other lanes resulting in measured lengths that were larger than the true length. If $c > 1$,

then the clone's fragments had moved quickly relative to other lanes resulting in measured lengths that were smaller than the true length.

Define the **percent correlated error** as the proportionality constant which determines the range in which the majority of the values for the shifting coefficient c will fall. Specifically, if there is a k% correlated error, then the majority of the values for c should fall within $1 \pm \frac{k}{100}$. This thesis assumes that the same percent correlated error applies to all of the lanes in all of the gels used for analysis. In other words, different lanes may have different values for the shifting coefficient, c, but the range in which the majority of values for c are expected to fall is constant.[19] For example, let a clone have fragments of true length 500, 1200, 3400, 7200, and let the percent correlated error be 2.5%, implying that c should fall in the range $1 \pm .025$. Let the clone be fingerprinted twice, running in lane_1 and lane_2, and let the measured lengths for lane_1 be 505, 1212, 3434, 7272, and for lane_2 be 490, 1176, 3332, 7056. Assume there is no random measurement error. If c is .99 for lane_1 and 1.0204 for lane_2, then the measured lengths from both lanes would be shifted back to the true lengths.

A smaller percent correlated error implies increased accuracy in the measured fragment lengths. The percent correlated error will vary between laboratories.

In this work, the correlated error is modelled using a probability density h, where

[19]The percent correlated error is a constant with respect to the subsequent probability expression derivations. Therefore, using a different value for particular gels or lanes will not invalidate subsequent derivations. Therefore, this assumption can be made without loss of generality, but with increased notational convenience.

the probability that the value of the shifting coefficient is between C_1 and C_2 is,

$$P(C_1 < c < C_2) \stackrel{\text{def}}{=} \int_{c=C_1}^{C_2} h(c)dc.$$

Further, h is assumed to be a normal distribution with mean 1 (μ_c) and standard deviation (σ_c),

$$h(C) \stackrel{\text{def}}{=} \mathcal{N}(C \mid \mu_c = 1, \sigma_c) = \frac{1}{\sqrt{2\pi}\sigma_c} e^{\frac{-1}{2\sigma_c^2}(C-1)^2}.$$

A normal distribution is used because a unimodal probability density is needed, and the normal distribution has certain mathematical properties which will allow subsequent approximations to be made (see Section 3.4.3). Defining the mean of h to be 1.0 is based on the assumption that it is most likely for the fragments in a gel lane to migrate at the typical speed.

For this thesis, σ_c is defined such that $1 \pm 2\sigma_c = 1 \pm \frac{k}{100}$ for a k% correlated error. In other words, $\sigma_c \stackrel{\text{def}}{=} \frac{k}{2\times 100}$. By defining σ_c in this way, 95% of all values of c will fall within the range defined by the percent correlated error.

In this model different lanes may be associated with different values of the shifting coefficient, c. It will be assumed that the value of c for a lane is independent of the value of c for any other lane. In other words, this model will *not* account for correlated error which affects the gel as a whole,[20] but only for correlated error which affects lanes in a gel individually. Recent work on spatial normalization of digitized gel images [8] will greatly diminish the correlated error which affects the

[20]Correlated error which affects the gel as a whole includes the smile and frown effects often seen in gels.

gel as a whole. Therefore, it is reasonable to exclude this source of error from the model.

2.2.3 Fragment Length Multiplicity

Fragments which do not actually overlap may nevertheless have the same true length. Because agarose gel electrophoresis can only differentiate between fragments of different lengths, this fingerprinting methodology will not be capable of distinguishing between overlapping fragments (which, by definition, must have the same true length) and non-overlapping fragments which simply happen to have the same true length. Therefore, even if there were no measurement error, inferences that two fragments with the same true length overlap may still be erroneous. This type of error is termed the error due to fragment length multiplicity.

The probability of erroneously inferring that fragments with the same true length overlap will depend on both the fragment length and on the frequency with which that fragment length occurs. If cutsites occur frequently along the genome, then there will be many small fragments and few large fragments. Thus, the probability of an erroneous overlap inference will be higher for two small fragments of the same true length than for two large fragments of the same true length. Conversely, if cutsites occur infrequently along the genome, then there will be many large fragments and few small fragments. Thus, the probability of an erroneous overlap inference will be higher for two large fragments of the same true length than for two small fragments of the same true length.

In this work, error due to multiplicity of fragment lengths is modelled using

a probability function f, where $f(X)$ is the probability that the true length of a fragment is X,

$$P(tl = X) \stackrel{\text{def}}{=} f(X).$$

Further, f is assumed to be a geometric distribution with p the probability of a cutsite occurrence,

$$f(X) \stackrel{\text{def}}{=} p(1-p)^{X-1}.$$

A geometric distribution is used because the probability that a fragment has a true length X is the probability that $X-1$ sequential positions along the genome are not cutsites, and that the X^{th} position is a cutsite. Recall that cutsites are contained within recognition sequences. This use of a geometric distribution does assume that the recognition sequences containing the cutsites which define the two ends of a fragment do not overlap. Therefore, this reasoning is not valid for fragment lengths less than twice the sitesize. In practice, measured fragment lengths are greater than four hundred base pairs [18] while sitesizes are between four and six base pairs long. Thus, this boundary condition is not of practical concern. Additionally, the use of a geometric distribution assumes that the cutsites are randomly distributed along the genome.

To calculate p, the base content of the genome and the recognition sequence of the enzyme are needed. Let ν_A, ν_T, ν_G, ν_C be the frequency with which the bases occur in the genome, and let $\#_A$, $\#_T$, $\#_G$, $\#_C$ be the number of times the bases occur in the recognition sequence. Assuming that the bases are randomly

distributed along the genome and that there is no strand bias, $p = \nu_A^{\#_A}\nu_T^{\#_T}\nu_G^{\#_G}\nu_C^{\#_C}$ [22].[21] For example, if the genome has an A-T content of 60%, a G-C content of 40%,[22] and the recognition sequence of the enzyme is AATGC, then $p = .3^2.3^1.2^1.2^1$.

In this model, the probability of obtaining a fragment of a particular true length X, monotonically decreases as X increases.

2.3 Basic Probabilistic Approach

Using the previously defined probability models for the sources of error in the laboratory methodology, probability expressions can be derived for each of the four overlap types described in Section 2.1.

The probability of interest is the conditional probability $P(Overlap \mid Data)$, which is the probability that a postulated overlap is correct given that the measured fragment length data have been obtained. Recall that a postulated overlap consists of a set of constituent f-f, f-fc, or fc-fc overlaps. Therefore, $P(Overlap \mid Data)$ is the joint probability that the postulated constituent f-f (f-fc, or fc-fc) overlaps are correct given the measured length data. Unfortunately, the probabilities for each individual f-f (f-fc, fc-fc) overlap are *not* independent. For example, if a postulated overlap consists of two constituent f-f overlaps, f-f_1 and f-f_2, then the probability that f-f_2 is correct should be much higher if it is known that f-f_1 is correct than if it

[21]This definition can be extended to the case of multiple enzymes which are treated as a single unit, as discussed in Section 1.3.1. Let n be the number of enzymes used, and $p_i = \nu_A^{\#_{A,i}}\nu_T^{\#_{T,i}}\nu_G^{\#_{G,i}}\nu_C^{\#_{C,i}}$ where $\#_{A,i}$, $\#_{T,i}$, $\#_{G,i}$, $\#_{C,i}$ are defined as the number of times the bases occur in the recognition sequence of enzyme i. Then, $p = \sum_{i=1}^{n} p_i$.

[22]The frequency of A equals the frequency of T, and the frequency of G equals the frequency of C because of the base pairing complementation restrictions. Therefore, $\nu_A = \nu_T = .3$, and $\nu_G = \nu_C = .2$.

is not known that f-f_1 is correct. Because of this dependence, the joint probability $P(Overlap \mid Data)$ cannot be expressed as a product of the probabilities for the simpler f-f (f-fc, or fc-fc) cases. As such, it has proven very difficult to develop expressions directly for $P(Overlap \mid Data)$.

However, this is not the situation for the "reverse" conditional probability $P(Data \mid Overlap)$, which is the probability of obtaining the measured fragment length data given that the postulated overlap is correct. It can be shown that for any constituent postulated f-f (f-fc, fc-fc) overlap, the probability of obtaining the measured length data given that the fragments do overlap *is* independent of the probabilities for the remaining constituent f-f (f-fc, fc-fc) overlaps (see Section 3.1). Thus, $P(Data \mid Overlap)$ can be expressed as a product of probabilities for the simpler f-f (f-fc, fc-fc) cases. As such, it proven fairly straightforward to derive expressions for $P(Data \mid Overlap)$.

In addition, it has been possible to derive expressions for the probability $P(Overlap)$, which is the prior probability that a postulated overlap is correct with no knowledge of any measured length data.

To obtain the conditional probability of interest, $P(Overlap \mid Data)$, from the probability expressions for $P(Data \mid Overlap)$ and $P(Overlap)$, Bayes' formula [6] can be used. Let n be the number of overlaps that can be postulated and $overlap_p$ be the particular overlap of concern. Using Bayes' formula,

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

Theoretically, the denominator sum includes *every* possible overlap between the two clones (or the clone and the map, or the two maps) which does not violate any partial order restrictions. This includes overlaps for which paired fragments have very different measured lengths and, for any postulated overlap, every suboverlap down to, and including, the null overlap. However, in practice all of these overlaps do *not* have to be included in the sum. Those overlaps for which the probability $P(Data \mid Overlap) \times P(Overlap)$ is very small can be excluded without significantly affecting the resultant probability $P(Overlap \mid Data)$. Specifically, overlaps which include *any* constituent f-f, f-fc, or fc-fc overlaps in which fragments are paired which do not have the same measured lengths, within experimental error, can be excluded.

The definition of "the same measured length, within experimental error," depends on both the percent random and percent correlated errors. Let Y_j be the measured length of fragment j, and k_r and k_c be the percent random and correlated errors, respectively. For the implementations done in this work, Y_1 and Y_2 were considered the "same" if $|Y_1 - Y_2| \leq \frac{1}{100}(k_c \times Y_1 + k_c \times Y_2 + k_r \times Y_1 + k_r \times Y_2)$. The term **postulatable overlap** will refer to overlaps for which (a) there are no violations on pre-existing partial orderings, and (b) each constituent f-f, f-fc, or fc-fc overlap contains only fragments of the "same" measured lengths, as defined above.

For notational convenience, subsequent discussions will refer to the probability $P(Overlap \mid Data)$ as $P_{o|d}$, the probability $P(Data \mid Overlap)$ as $P_{d|o}$, and the probability $P(Overlap)$ as P_o. However, when expressions are based on particular overlaps and particular measured length data, they will be written as $P(\text{“}Overlap$

A" | Data for "Fragments B")[23], *P(Data for "Fragments B" | "Overlap A")*[24], and *P("Overlap A")*[25].

To reiterate, the probability expression for $P_{o|d}$ is based on *postulated* overlaps. The approach is to enumerate every postulatable overlap between the two clones (or the clone and the map, or the two maps), and for each overlap to determine the probability $P_{d|o}$ and the probability P_o. Using Bayes' formula, the probability $P_{o|d}$ for any of the postulatable overlaps can then be calculated. All subsequent discussions of reverse conditional and prior probabilities of overlaps are related *solely* to postulated overlaps unless otherwise stated.

---

[23]This is the probability that the postulated "overlap A" is correct given the measured length data for "fragments B."

[24]This is the probability that measured length data for "fragments B" will be obtained given that the postulated "overlap A" is correct.

[25]This is the prior probability that the postulated "overlap A" is correct.

CHAPTER 3

Probability Expression Derivations

In this chapter, the probability expressions are first derived using a simplified model in which correlated error is assumed not to occur. The expressions derived from this model are then extended to a more general model, in which correlated error is assumed to occur. Two different formats for the probability expressions are then discussed – the absolute probability that a postulated overlap is correct, and the relative odds for which of two particular postulated overlaps is correct. Finally, several approximations are derived which greatly improve the computational feasibility of the probability expressions.

3.1 Simplified Model: No Correlated Error

The simplified model probability expressions are derived using only the probability models for the random measurement error and the error due to the multiplicity of fragment lengths, because this model assumes that correlated error does not occur. The expressions are first derived for the simplest type of overlap, the fragment-fragment overlap, and then are progressively extended to the clone-clone, clone-map, and map-map overlaps. Derivations will be given for the probabilities $P_{d|o}$ and P_o. Subsequent application of Bayes' formula to calculate $P_{o|d}$ will be left implicit.

The probability $P_{d|o}$ will be derived for each overlap type using the functions g and f. The probability P_o will be derived for each overlap type using four component probabilities: **alignment probability** P_a, **grouping probability** P_g, **ordering**

probability P_r, and **orientation probability** P_n, such that

$$P_o = P_a \times P_g \times P_r \times P_n.$$

Define a **fragment position**[26] on a DNA molecule as the interval between two consecutive cutsites, and a **genome position** as a fragment position on a genome.[27] The genome position for a clone (or a map) is defined by the genome position of its leftmost fragment (or fragment column). The alignment probability, P_a, is the probability that the clones (or maps) originated from genome positions which are consistent with the postulated overlap. The grouping probability, P_g, is the probability that particular subsets of fragments (or fragment columns) within the clone (or map) are located in contiguous positions. The subsets of fragments (or columns) which should occupy contiguous positions are determined by the partial order that the postulated overlap would impose if it were inferred to be correct. The ordering probability, P_r, is the probability that the sequential arrangement of the fragments (or fragment columns) within these subsets are consistent with the postulated overlap. The orientation probability, P_n, is the probability of occurrence for the particular left-right orientation of the clone(s) and/or map(s) with respect to the genome.

---

[26] In subsequent sections when the meaning is clear, this may be referred to as the **position**.

[27] If t is the number of genome positions and the genome is linear, then $t = n + 1$. If the genome is circular, then $t = n$.

3.1.1 Fragment-Fragment Overlap

For the fragment-fragment overlap type, the probabilities are derived for two fragments randomly chosen out of a pool of fragments resulting from complete digestion of the genomic DNA. The concepts developed for this overlap type will be used in the derivations for the subsequent overlap types.

Probability $P_{d|o}$ The probability that a fragment has a measured length Y is

$$\begin{aligned} P(ml = Y) &= \sum_{X=1}^{\infty} P(ml = Y \mid tl = X) P(tl = X) \\ &= \sum_{X=1}^{\infty} g(Y \mid X) f(X). \end{aligned} \tag{2}$$

The product of g and f in Expression 2 yields the joint probability $P(ml = Y, tl = X)$. Summing this joint probability over all possible true lengths X gives the probability that the measured length is Y.

Using Expression 2 for $P(ml = Y)$, the probability $P_{d|o}$ that two fragments have measured lengths Y and Z if the fragments do not overlap[28] is the product of the probabilities $P(ml_1 = Y)$ and $P(ml_2 = Z)$, because both the random measurement errors are independent and the true lengths of non-overlapping fragments are independent. Therefore,

$$\begin{aligned} P_{d|o} &= P(ml_1 = Y, ml_2 = Z \mid frag_1 \textit{ does not overlap } frag_2) \\ &= P(ml_1 = Y) \times P(ml_2 = Z) \\ &= \sum_{X=1}^{\infty} P(ml_1 = Y \mid tl_1 = X) P(tl_1 = X) \times \sum_{X=1}^{\infty} P(ml_2 = Z \mid tl_2 = X) P(tl_2 = X) \\ &= \sum_{X=1}^{\infty} g(Y \mid X) f(X) \times \sum_{X=1}^{\infty} g(Z \mid X) f(X). \end{aligned} \tag{3}$$

The probability $P_{d|o}$ that two fragments have measured lengths Y and Z if the

[28] This is the probability $P_{d|o}$ for the fragment-fragment null overlap.

fragments do overlap is essentially the probability that two separate measurements of one DNA fragment yield Y and Z, because overlapping fragments represent the *same* piece of original genomic DNA. Therefore,

$$\begin{aligned} P_{d|o} &= P(ml_1 = Y, ml_2 = Z \mid frag_1 \text{ does overlap } frag_2) \\ &= \sum_{X=1}^{\infty} P(ml_1 = Y \mid tl = X) P(ml_2 = Z \mid tl = X) P(tl = X) \\ &= \sum_{X=1}^{\infty} g(Y \mid X) g(Z \mid X) f(X). \end{aligned} \tag{4}$$

There is one f factor in Expression 4 because there is essentially one DNA fragment, and there are two g factors because there are two measurements.[29]

Probability P_o Let t be the number of genome positions. The prior probability P_o that two fragments do or do not overlap is $\frac{1}{t}$ or $1 - \frac{1}{t}$, respectively.

By definition, fragments which overlap have originated from the same position on the genome. Given the genome position of one fragment, the probability P_a that the other fragment originated from the same genome position is $\frac{1}{t}$. There are no grouping, ordering, or orientation considerations, thus P_g, P_r, and P_n are each 1.0 for this case.

By definition, fragments which do not overlap have originated from different positions on the genome. Given the genome position of one fragment, the probability P_a that the other fragment originated from a different genome position is $1 - \frac{1}{t}$. Again, P_g, P_r, and P_n are each 1.0 for this case.

[29]The probability that the first measured length is Y is conditionally independent of the probability that the second measured length is Z, given the true length X. This conditional independence occurs because this model assumes there is no correlated error. Because of the conditional independence, the product of g, g, and f yields the joint probability $P(ml_1 = Y, ml_2 = Z, tl = X)$.

3.1.2 Clone-Clone Overlap

Let clone$_1$ have r fragments, clone$_2$ have s fragments, and let there be m constituent f-f overlaps in the postulated overlap between clone$_1$ and clone$_2$.

Probability $P_{d|o}$ Without loss of generality, number the fragments so that $1 \ldots m$ are the overlapping fragments from each clone. Let Y_j and Z_j be the measured lengths of fragment j from clone$_1$ and clone$_2$, respectively. Then, the probability $P_{d|o}$ for a clone-clone overlap is

$$P_1 \times P_2 \times P_3,$$

where

$$P_k = \begin{cases} \prod_{j=1}^{m} \sum_{X=1}^{\infty} g(Y_j \mid X) g(Z_j \mid X) f(X) & \text{if k=1.} \\ \prod_{j=m+1}^{r} \sum_{X=1}^{\infty} g(Y_j \mid X) f(X) & \text{if k=2.} \\ \prod_{j=m+1}^{s} \sum_{X=1}^{\infty} g(Z_j \mid X) f(X) & \text{if k=3.} \end{cases} \quad (5)$$

P_1 is the probability $P_{d|o}$ for the m constituent f-f overlaps. Expression 4 is the probability $P_{d|o}$ for one f-f overlap. Because this model assumes there is no correlated error, the probability for the m f-f overlaps is the product of the m individual f-f overlap probabilities.

P_2 is the probability $P_{d|o}$ for the $r - m$ fragments of clone$_1$ which do not overlap clone$_2$, and P_3 is the probability $P_{d|o}$ for the $s - m$ fragments of clone$_2$ which do not overlap clone$_1$. Expression 3 is the probability $P_{d|o}$ for two fragments which do not overlap. Because this model assumes there is no correlated error, this expression can be extended to the $r - m$ and $s - m$ fragments, as the product of $P(ml_j = Y_j)$ and $P(ml_j = Z_j)$ for $m + 1 \leq j \leq r$ and $m + 1 \leq j \leq s$, respectively.

The null overlap for the clone-clone case occurs when $m = 0$. Therefore, $P_{d|o}$ for the null overlap is $P_2 \times P_3$.

Probability P_o Let t be the number of genome positions. The clone-clone prior probability P_o is

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

There is only one possible left-right orientation of a clone with respect to the genome, therefore P_n is 1.0 for each of the cases.

For $m = 0$ and $t \geq r+s-m$, it is postulated that the two clones have no fragments in common. Given the genome positions of the fragments of clone$_1$, the leftmost fragment of clone$_2$ cannot have originated from any of the r positions of clone$_1$ or any of the $s - 1$ positions preceding clone$_1$. The probability of this occurring, P_a, is $1 - \frac{r+s-1}{t}$. Any fragment grouping and ordering is consistent with the null overlap, thus both P_r and P_g are 1.0.

For $0 < m < min(r,s)$ and $t \geq r + s - m$, it is postulated that each clone contains some fragments which do not overlap the other clone, and some which do. Given the genome positions of the fragments of clone$_1$, the leftmost fragment of clone$_2$ must have originated from either position $r - m + 1$ of clone$_1$ or from the $(s - m)^{th}$ position preceding clone$_1$. The probability of this occurring, P_a, is $\frac{2}{t}$.

For the positions of the clones' fragments to be consistent with the partial order which would be imposed if the overlap were correct, the fragments involved in the m f-f overlaps must be located in the first m positions of one clone and the last m positions of the other. There are $\binom{r}{m}\binom{s}{m}$ different ways in which fragments could fill these positions, thus P_g is $\frac{1}{\binom{r}{m}\binom{s}{m}}$. If the overlap is correct, then the fragments which are paired must have originated from the same genome positions. Therefore, the order of the m fragments in clone$_2$ must be the same as the order of the m paired fragments in clone$_1$. There are $m!$ different ways in which the m fragments from clone$_2$ could be ordered, thus P_r is $\frac{1}{m!}$.

For $m = min(r, s)$ and $t \geq r + s - m$, the smaller clone is a subclone of the larger. Given the genome positions of the fragments of the larger clone, the leftmost fragment of the smaller clone must have originated from one of the first $|r - s| + 1$ positions of the larger. Therefore, P_a is $\frac{|r-s|+1}{t}$. Let the leftmost fragment of the smaller clone originate from position j of the larger. To be consistent with the partial order that would be imposed if the overlap were correct, the m fragments of the larger clone which are involved in the f-f overlaps must occupy positions $j \ldots (j + m - 1)$. There are $\binom{max(r,s)}{m}$ different ways in which fragments could fill these positions. The postulated overlap imposes no constraints on the grouping of the subclone's fragments within positions $j \ldots (j+m-1)$. Therefore, P_g is $\frac{1}{\binom{max(r,s)}{m}}$. Only one of the $m!$ possible orderings of the smaller clone's fragments is the same as the ordering of the m paired fragments of the larger clone, and so P_r is $\frac{1}{m!}$.

Appendix A.1 proves that these expressions for P_o do sum to 1.0 over the probability space.

3.1.3 Clone-Map Overlap

The probabilities $P_{d|o}$ and P_o for an overlap between a clone and a map will be derived by assuming that the map is correct. Specifically, $P_{d|o}$ will be *P(data | overlap, map is correct)*, P_o will be *P(overlap | map is correct)*, and $P_{o|d}$ will be *P(overlap | data, map is correct)*[30]. A method for determining the probability that a map is correct has not been found, and so comparisons of $P_{o|d}$ for different clone-map overlaps can only be done if both overlaps involve the same map. For ease of notation, the "map is correct" term will be implicit in subsequent discussions. Let r be the number of fragments in the clone, s be the number of fragments in the map, and m be the number of constituent f-fc overlaps.

Probability $P_{d|o}$ Just as the "data" for the clone-clone overlap includes the measured lengths of all the fragments of both clones, the "data" for the clone-map overlap includes the measured lengths of all the fragments in both the clone and the map.

Let n_j be the number of fragments in column j of the map, $Y_{i,j}$ be the measured length of fragment i in column j, and Z_j be the measured length of fragment j of the clone. Without loss of generality, number the fragments and fragment columns so that $1 \ldots m$ overlap.

The probability $P_{d|o}$ that the fragments in column j have measured lengths

[30]If *env* is the background (unchanging) environment, then

$$P(A_p \mid B, env) = \frac{P(B \mid A_p, env) \times P(A_p \mid env)}{\sum_i P(B \mid A_i, env) \times P(A_i \mid env)}.$$

$Y_{1,j} \ldots Y_{n_j,j}$ is essentially the probability that n_j independent measurements of one piece of DNA yield $Y_{1,j} \ldots Y_{n_j,j}$. Extending Expression 4, which is the probability $P_{d|o}$ for two fragments postulated to overlap, the probability $P_{d|o}$ for obtaining the measured lengths of the n_j fragments in column j of the map is

$$\begin{aligned} P_{d|o} &= P(ml_1 = Y_{1,j}, \ldots, ml_{n_j} = Y_{n_j,j} \mid frag_{1,j} \ldots frag_{n_j,j} \; overlap) \\ &= \sum_{X=1}^{\infty} f(X) \prod_{i=1}^{n_j} g(Y_{i,j} \mid X). \end{aligned} \tag{6}$$

Similarly, the probability $P_{d|o}$ of obtaining the measured length data in an f-fc overlap betwen fragment j of the clone and column j of the map is

$$\begin{aligned} P_{d|o} &= P(ml_{clone,j} = Z_j, ml_1 = Y_{1,j}, \ldots, ml_{n_j} = Y_{n_j,j} \mid frag_{clone,j} \; overlaps \; column_j) \\ &= \sum_{X=1}^{\infty} f(X) g(Z_j \mid X) \prod_{i=1}^{n_j} g(Y_{i,j} \mid X), \end{aligned} \tag{7}$$

where $ml_{clone,j}$ is the measured length of fragment j of the clone, $frag_{clone,j}$.

Using Expressions 6 and 7, the probability $P_{d|o}$ for the clone-map overlap is

$$P_1 \times P_2 \times P_3,$$

where

$$P_k = \begin{cases} \prod_{j=1}^{m} \sum_{X=1}^{\infty} f(X) g(Z_j \mid X) \prod_{i=1}^{n_j} g(Y_{i,j} \mid X) & \text{if k=1.} \\ \prod_{j=m+1}^{r} \sum_{X=1}^{\infty} g(Z_j \mid X) f(X) & \text{if k=2.} \\ \prod_{j=m+1}^{s} \sum_{X=1}^{\infty} f(X) \prod_{i=1}^{n_j} g(Y_{i,j} \mid X) & \text{if k=3.} \end{cases} \tag{8}$$

P_1 is the probability $P_{d|o}$ for the m constituent f-fc overlaps. Because this model assumes there is no correlated error, this probability is the product of the m individual f-fc overlap probabilities, as defined in Expression 7.

P_2 is the probability $P_{d|o}$ for the $r - m$ fragments of the clone which do not overlap the map. It is the same as P_2 and P_3 of Expression 5.

P_3 is the probability $P_{d|o}$ for the $s - m$ fragment columns of the map which do not overlap the clone. Because this model assumes there is no correlated error, P_3 is the product of the $s - m$ individual fragment column probabilities, as defined in Expression 6.

The clone-map null overlap occurs when $m = 0$, and so $P_{d|o}$ for the null overlap is $P_2 \times P_3$.

Probability P_o Let t be the number of genome positions, a and b be two sets of fragment columns in the map, and $|a|$ and $|b|$ be the number of fragment columns in sets a and b, respectively. Define the **clone end** as either the first or the last fragment in the clone. Note that if a set in the map overlaps a clone end, then some of the columns in the set may be involved in f-fc overlaps but some may not. If a (or b) is postulated to overlap a clone end, then let a_1 (or b_1) be the subsets of the columns in a (or b) which are involved in f-fc overlaps. Figure 7 illustrates this terminology. In the figure, an overlap between a clone and a map is shown in which set a of the map overlaps the left clone end and set b of the map overlaps the right clone end.

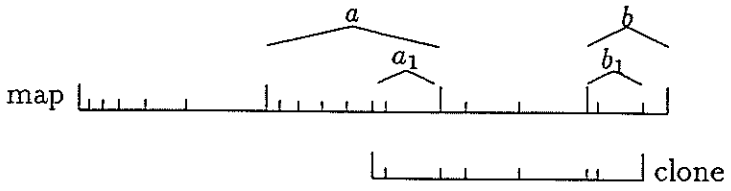


Figure 7: Clone Ends Overlapping Sets in a Map. A postulated overlap between a clone and a map is shown, in which the left clone end overlaps set a of the map, and the right clone end overlaps set b of the map. The subsets of sets a and b which contain the columns involved in the f-fc overlaps are marked a_1 and b_1. The paired fragments and columns involved in f-fc overlaps are vertically aligned. For notational convenience, except when otherwise required by the vertical alignment of paired fragments and columns, the unordered fragments and columns (which are delimited by short vertical lines) are written in order of ascending length.

The clone-map prior probability P_o is

$$P_o = \begin{cases} 0 & \text{if } t < r+s-m. \\ 1 - \frac{r+s-1}{t} & \text{if } m = 0 \text{ and } t \geq r+s-m. \\ \frac{1}{t\binom{r}{m}\binom{|a|}{|a_1|}m!} & \text{if } 0 < m < min(r,s), \text{ set } a \text{ overlaps clone end, and } t \geq r+s-m. \\ \frac{r-s+1}{t\binom{r}{s}s!} & \text{if } m = s \text{ and } t \geq r+s-m. \\ \frac{1}{t\binom{|a|}{|a_1|}\binom{|b|}{|b_1|}r!} & \text{if } m = r, \text{ clone ends overlap sets } a \text{ and } b, \text{ and } t \geq r+s-m. \\ \frac{|a|-r+1}{t\binom{|a|}{|a_1|}r!} & \text{if } m = r, \text{ clone overlaps only set } a, \text{ and } t \geq r+s-m. \end{cases}$$

Although for a clone there is only one possible orientation with respect to the genome, there are two possible orientations for a map (when the map contains more than one set). Therefore, for each of the five cases described below, P_n is

$1 \times \frac{1}{2} = \frac{1}{2}$. In addition, the probability P_o for each case is the sum of P_o for each of the two map orientations. P_a, P_g, and P_r are derived below only for the map orientation shown in Figure 8. However, the expressions derived below can also be derived for the alternate map orientation. Therefore, for each of the five cases, $P_o = 2 \times P_a \times P_g \times P_r \times \frac{1}{2} = P_a \times P_g \times P_r$, where P_a, P_g, and P_r are as derived below.

Figure 8(a) shows, for one particular map orientation, a clone-map overlap when $m = 0$ and $t \geq r + s - m$. In this case, it is postulated that the clone and the map have no fragments in common. For similar reasoning as used in the clone-clone case, P_a is $1 - \frac{r+s-1}{t}$. Both P_g and P_r are 1.0.

Figure 8(b) shows, for one particular map orientation, a clone-map overlap when $0 < m < min(r, s)$ and $t \geq r + s - m$. Here it is postulated that the clone contains some fragments which overlap the map and some which do not, and that the map contains some columns which overlap the clone and some which do not. Let set a in the map overlap the clone end. (Only one clone end is overlapped by the map.) Although an overlap containing m constituent f-fc overlaps is consistent with the leftmost fragment of the clone originating from either position $s - m + 1$ in the map or the $(r - m)^{th}$ position preceding the map, these two overlaps are *different* because of the partial order already established in the map.[31] Only one of these overlaps can be consistent with the f-fc pairing in the postulated overlap. Therefore, P_a is $\frac{1}{t}$. Without loss of generality, let the leftmost fragment of the clone originate from

[31] Figure 8(b) shows the case where the leftmost fragment of the clone originated from position $s - m + 1$ in the map.

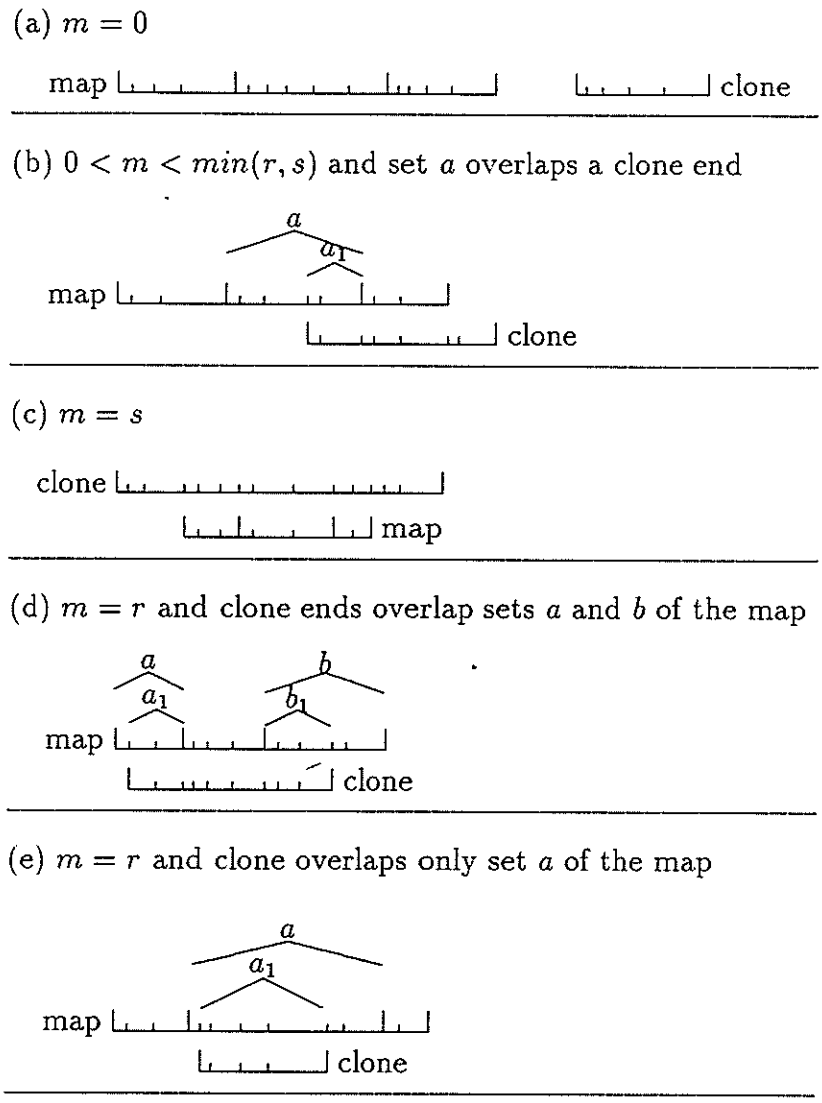


Figure 8: Examples of Overlaps for Each of the Cases used in Deriving the Clone-Map Prior Probability P_o. The fragments within the map columns are not shown. The paired clone fragments and map columns involved in f-fc overlaps are vertically aligned. For notational convenience, except when otherwise required by the vertical alignment of paired fragments and columns, the unordered fragments and columns (delimited by short vertical lines) are written in order of ascending length. Only one of the two possible orientations for the map is shown.

position $s - m + 1$ in the map. To maintain consistency with the partial order which would be imposed by the overlap, the m fragments of the clone which are postulated to overlap the map must be located in the first m positions of the clone, and the $|a_1|$ fragment columns of the map which are postulated to overlap the clone end must be located in the last positions of set a. There are no *new* constraints on the positions of the other columns in the map. Because the map is assumed to be correct, the partial ordering which was already present in the map does not need to be accounted for in P_g. Therefore, P_g is $\frac{1}{\binom{r}{m}\binom{|a|}{|a_1|}}$. The order of the m fragments in the clone which overlap the map must be the same as the order of the m paired map columns, thus P_r is $\frac{1}{m!}$.

Figure 8(c) shows, for a particular map orientation, a clone-map overlap for which $m = s$ and $t \geq r + s - m$. The map is contained entirely within the clone. Given the genome positions of the clone's fragments, the leftmost column of the map may have originated from any of the first $r - s + 1$ positions of the clone, thus P_a is $\frac{r-s+1}{t}$. For similar reasoning as in the clone-clone case when $m = min(r, s)$, P_g is $\frac{1}{\binom{r}{s}}$, and for similar reasoning as in the clone-map case when $0 < m < min(r, s)$, P_r is $\frac{1}{s!}$.

For $m = r$, the clone is contained entirely within the map. There are two possible situations. Either the clone overlaps only one set of the map, or it overlaps two or more sets of the map.

Figure 8(d) shows, for a particular map orientation, a clone-map overlap for which $m = r$, $t \geq r+s-m$, and the clone overlaps two or more of the map's sets. Let sets a and b overlap the clone ends, and let a precede b. Given the genome positions of

the fragment columns of the map, the leftmost fragment of the clone could have only originated from the $(|a| - |a_1| + 1)^{th}$ position in set a, and so P_a is $\frac{1}{t}$. The fragment columns in a_1 must be located in the last positions of a, and the fragment columns in b_1 must be located in the first positions of b to maintain consistency with the partial order which would be imposed by the overlap. There are $\binom{|a|}{|a_1|}\binom{|b|}{|b_1|}$ different ways in which fragment columns could fill these positions. Therefore, P_g is $\frac{1}{\binom{|a|}{|a_1|}\binom{|b|}{|b_1|}}$. Because no grouping constraints were placed on the clone, the ordering of the clone's fragments, as a whole, can be considered to determine P_r. The probability that the clone's fragments will be in the same order as the paired columns in the map, P_r, is $\frac{1}{r!}$.

Figure 8(e) shows, for a particular map orientation, a clone-map overlap for which $m = r$, $t \geq r + s - m$, and the clone is contained entirely within one set of the map. Let set a of the map contain the clone. The probability for this case is essentially the same as the probability for the clone-clone case where $m = min(r, s)$, in which set a of the map replaces the larger clone of the clone-clone case. Given the genome positions of the fragment columns in set a, the leftmost fragment of the clone could have originated at any of the first $|a|-m+1$ positions in set a. Therefore, P_a is $\frac{|a|-m+1}{t}$. For similar reasoning as in the clone-clone case when $m = min(r, s)$, P_g is $\frac{1}{\binom{|a|}{|a_1|}}$. ($|a_1| = m$.) For similar reasoning as in the previous clone-map $m = r$ case, P_r is $\frac{1}{r!}$.

Appendix A.1. proves that these expressions for P_o sum to 1.0 over the probability space.

3.1.4 Map-Map Overlap

As for the clone-map overlap, the probability expressions for the map-map overlap are derived assuming that both maps are correct. Let r be the number of columns in map_1, s be the number of columns in map_2, and m be the number of constituent fc-fc overlaps.

Probability $P_{d|o}$ The "data" for a map-map overlap consists of all of the measured fragment lengths in both maps. Let $n_{j,1}$ and $n_{j,2}$ be the number of fragments in column j of map_1 and map_2, respectively, and $Y_{i,j}$ and $Z_{i,j}$ be the measured lengths of fragment i in column j of map_1 and map_2, respectively. Without loss of generality, number the fragment columns of the maps so that $1 \ldots m$ overlap.

Using Expression 6 for the probability of obtaining the measured lengths of the fragments in one column, the probability $P_{d|o}$ for the map-map overlap is

$$P_1 \times P_2 \times P_3,$$

where

$$P_k = \begin{cases} \prod_{j=1}^{m} \sum_{X=1}^{\infty} f(X) \prod_{i=1}^{n_{j,2}} g(Z_{i,j} \mid X) \prod_{i=1}^{n_{j,1}} g(Y_{i,j} \mid X) & \text{if k=1.} \\ \prod_{j=m+1}^{r} \sum_{X=1}^{\infty} f(X) \prod_{i=1}^{n_{j,2}} g(Z_{i,j} \mid X) & \text{if k=2.} \\ \prod_{j=m+1}^{s} \sum_{X=1}^{\infty} f(X) \prod_{i=1}^{n_{j,1}} g(Y_{i,j} \mid X) & \text{if k=3.} \end{cases} \quad (9)$$

P_1 is the probability $P_{d|o}$ for the m constituent fc-fc overlaps. It is an extension of P_1 in $P_{d|o}$ for the clone-map overlap (Expression 8) in which the probability $P_{d|o}$ for each column in map_2 is substituted for the probability $P_{d|o}$ for each single fragment in the clone.

P_2 is the probability $P_{d|o}$ for the $r - m$ non-overlapping columns of map_1, and P_3 is the probability $P_{d|o}$ for the $s - m$ non-overlapping columns of map_2. Therefore, these expressions are the same as for P_3 in $P_{d|o}$ for the clone-map overlap (Expression 8).

The probability $P_{d|o}$ for the null overlap, where $m = 0$, is $P_2 \times P_3$.

Probability P_o Let t be the number of fragment positions in the genome, a_j be set j of map_1, b_j be set j of map_2, d_1 be the total number of sets of map_1 which overlap map_2, and d_2 be the total number of sets of map_2 which overlap map_1. Without loss of generality, number the sets in the maps so that $a_1 \ldots a_{d_1}$ and $b_1 \ldots b_{d_2}$ are the sets involved in fc-fc overlaps. Let k_j be the number of sets of map_2 which overlap set a_j of map_1, and let $a_{1,j} \ldots a_{k_j,j}$ be subsets of a_j such that each subset contains all of the columns which overlap one set of map_2 and the intersection of any two subsets is null. Let l_j be the number of sets of map_1 which overlap set b_j of map_2, and let $b_{1,j} \ldots b_{l_j,j}$ be subsets of b_j such that each subset contains all of the columns which overlap one set of map_1 and the intersection of any two subsets is null. Figure 9 illustrates this terminology. The figure contains a postulated overlap between map_1 and map_2 for which each a_j, b_j, $a_{i,j}$, and $b_{i,j}$ are shown, and the values of d_1, d_2, k_j, and l_j are given.

Define a **map end** as the leftmost or rightmost column of a map. Note that if set a_j of map_1 overlaps an end of map_2, then $\sum_{i=1}^{k_j} |a_{i,j}|$ may not equal a_j because there may be columns in a which do not overlap map_2. For example, for set a_2 of map_1 in Figure 9, $\sum_{i=1}^{3} |a_{i,2}| \neq |a_2|$. Similar comments can be made for a set b_j of

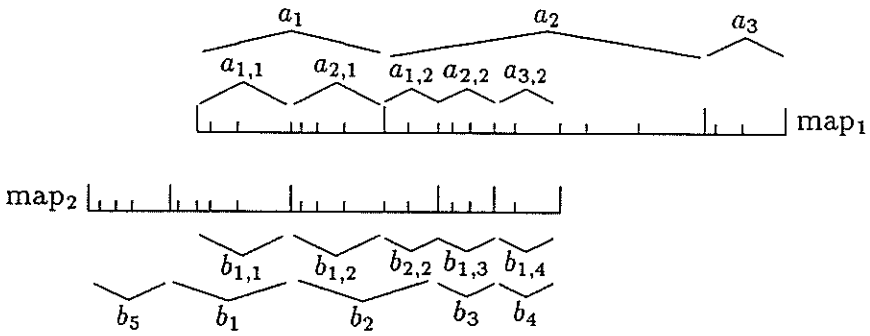


Figure 9: A Postulated Map-Map Overlap. The notation for a_i, $a_{i,j}$, b_i, $b_{i,j}$ is as described in the text. The values of d_i, k_i, and l_i, as defined in the text, are: $d_1 = 2$, $d_2 = 4$, $k_1 = 2$, $k_2 = 3$, $l_1 = 1$, $l_2 = 2$, $l_3 = 1$, $l_4 = 1$. The fragments within the columns of the maps are not shown. The paired columns involved in fc-fc overlaps are vertically aligned. For notational convenience, except when otherwise required by the vertical alignment of paired columns, the unordered columns (which are delimited by short vertical lines) are written in order of ascending length. The ordered sets are delimited by tall vertical lines.

map$_2$ which overlaps an end of map$_1$. This situation will need to be accounted for in the derivations for P_o.

The probability P_o is[32]

$$P_o = \begin{cases} 0 & \text{if } t < r+s-m. \\ 1 - \frac{r+s-1}{t} & \text{if } m = 0 \text{ and } t \geq r+s-m. \\ \frac{1}{2} \times \frac{1}{t \prod_{j=1}^{d_1} \binom{|a_j|}{|a_{1,j}|,\ldots,|a_{k_j,j}|} \prod_{j=1}^{d_2} \binom{|b_j|}{|b_{1,j}|,\ldots,|b_{l_j,j}|} \prod_{j=1}^{d_1} \prod_{i=1}^{k_j} |a_{i,j}|!} & \text{if } 0 < m < min(r,s) \text{ and } t \geq r+s-m. \\ \frac{1}{2} \times \frac{1}{t \prod_{j=1}^{d_1} \binom{|a_j|}{|a_{1,j}|,\ldots,|a_{k_j,j}|} \prod_{j=1}^{d_2} \binom{|b_j|}{|b_{1,j}|,\ldots,|b_{l_j,j}|} \prod_{j=1}^{d_1} \prod_{i=1}^{k_j} |a_{i,j}|!} & \text{if } m = min(r,s), \text{ small map overlaps} \geq 2 \text{ sets of large map, and } t \geq r+s-m. \\ \frac{|a_j|-s+1}{t\binom{|a_j|}{s} s!} & \text{if } m = s, \text{ both ends of map}_2 \text{ overlap set } a_j \text{ of map}_1, \text{ and } t \geq r+s-m. \\ \frac{|b_j|-r+1}{t\binom{|b_j|}{r} r!} & \text{if } m = r, \text{ both ends of map}_1 \text{ overlap set } b_j \text{ of map}_2, \text{ and } t \geq r+s-m. \end{cases}$$

Each map has two possible orientations with respect to the genome, thus there are four possible combinations of the maps' orientations. The probability of occurrence for a particular combination of orientations of the two maps, P_n, is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

---

[32]For the third and fourth cases, $\binom{|a_j|}{|a_{0,j}|,\ldots,|a_{k_j,j}|}$ replaces $\binom{|a_j|}{|a_{1,j}|,\ldots,|a_{k_j,j}|}$ in the expression, where $a_{0,j}$ is defined as the subset of a_j which contains the columns which do not overlap map$_2$. Similarly for b_j, $\binom{|b_j|}{|b_{0,j}|,\ldots,|b_{l_j,j}|}$ replaces $\binom{|b_j|}{|b_{1,j}|,\ldots,|b_{l_j,j}|}$ in the expression.