Sequence Assembly Validation by Multiple Restriction Digest Fragment Coverage Analysis

Eric C. Rouchka and David J. States

Institute for Biomedical Computing
Washington University
700 South Euclid Avenue
St. Louis, MO 63110-1012, USA
Email: ecr@ibc.wustl.edu

Abstract

DNA sequence analysis depends on the accurate assembly of fragment reads for the determination of a consensus sequence. This report examines the possibility of analyzing multiple, independent restriction digests as a method for testing the fidelity of sequence assembly. A dynamic programming algorithm to determine the maximum likelihood alignment of error prone electrophoretic mobility data to the expected fragment mobilities given the consensus sequence and restriction enzymes is derived and used to assess the likelihood of detecting rearrangements in genomic sequencing projects. The method is shown to reliably detect errors in sequence fragment assembly without the necessity of making reference to an overlying physical map. An html form-based interface is available at http://www.ibc.wustl.edu/services/validate.html

Keywords: Sequence assembly validation, dynamic programming, restriction fingerprint analysis

Introduction

Genomic sequence analysis depends on the accurate assembly of short (400 to 1,000 base pair) sequence reads into contigs that cover extended regions as a necessary step in deriving finished sequence. Errors at the fragment layout assembly stage may be difficult or impossible to detect later in the editing process, and fragment assembly errors may have a serious impact on the biological interpretation of the data. For example, entire regions of the genome could be inverted or swapped as a result of assembly errors. Such errors could impact the biological interpretation of the sequence data, potentially leaving groups of exons out, swapping exons or control elements onto the anti-sense strand, breaking genes into pieces, or dissociating genes from their control elements. Since assembly errors are difficult to detect and can impact the utility of the finished sequence, experimental validation of the fragment assembly is highly desirable.

Comparison of predicted and experimental restriction digests has been proposed as a means for validating fragment assembly. The pattern of fragment masses resulting from a restriction digest of the source DNA can be readily determined with a precision of ±1%. This pattern of restriction fragment masses is commonly referred to as a restriction fingerprint. The cleavage sites for restriction enzymes are specific so it is easy to electronically generate a set of predicted fragment masses from the finished sequence. Similarly, the location of each of the predicted fragments on the finished sequence is known. Errors in sequence assembly will either change fragment masses directly or rearrange the position of restriction sites resulting in new fragments with altered masses.

Restriction fragment matching has been extensively used as the basis for physical map assembly (Riles et al. 1993; Waterston et al. 1993). Similarities in fingerprint are used to infer clone overlap. Since most clones overlap over only a fraction of their length and because restriction digest sites may be polymorphic, software has been developed to recognize common features of fingerprint patterns while ignoring the disparities. Most of the information in a fingerprint is accessible even if several bands in the digestion pattern are missed or a number of false positives are scored.

In this report, we examine the use of multiple restriction digest fingerprints for assembly validation. Both simulated and experimental results will be discussed as well as a specific application to clone mapping. We also compare the requirements for fingerprint mapping with the requirements for assembly validation.

Methods

Dynamic programming algorithms were first used in the context of computational biology for the purpose of finding the best alignment between two DNA or protein sequences (Needleman and Wunsch 1970; Sellers 1974; Smith and

Waterman 1981). We have developed a similar dynamic programming algorithm to determine the maximum alignment of error prone electrophoretic mobility data to predicted fragment mobilities. The expected fragment mobility information can be calculated when the sequence to validate and the restriction enzyme patterns used in creating the experimental data are known. String matching functions are used to find the exact location of a particular cutting site in the sequence. Predicted fragments are generated according to these locations. The mobility, *m*, for each of these expected fragments is calculated using the same formula from which the experimental data is derived:

$$m_{fragment} = 2Log\left(\frac{L_{tot}}{L_{fragment}}\right)$$

where L_{tot} is the total length of the sequencing project. The factor of 2 is applied to give mobilities in the range typical of current experimental protocols, 0 to 20 cm. In these units, a standard deviation in determination of band position of 0.1 cm corresponds to a relative accuracy of mass determination of 0.5%.

Within the dynamic programming algorithm, fingerprint pattern alignments were scored using a log odds system based on the likelihood of deriving the observed fragment mobilities from the predicted digest mobilities relative to the odds of observing the fingerprint pattern at random.

Relationship	**Score**
Band match	$Log(P_{match}/P_{random})$
False positive	$Log(P_{false\ positive})$
False negative	$Log(P_{false\ negative})$

The probability, P_{match}, of a fragment having an observed mobility, m_{obs}, given a true mobility, *m*, and normally distributed errors in mobility determination (Drury et al. 1990, 1992), is

$$P_{match}(m_{obs} \mid m) = \frac{1}{\sqrt{2\pi\sigma}} e^{(-\frac{(m_{obs}-m)^2}{2\sigma^2})}$$

Assuming that the fragment mobilities scale as the log of the molecular weight of the fragment (Maniatis et al. 1975), this formulation results in a constant fractional error in mass determination and agrees with empirical observations based on current data (M. Marra, personal communication).

The probability, P_{random}, of matching a band at random given a maximum mobility of *X* and *N* bands is:

$$P_{random} = \frac{N}{X}$$

The values of $P_{false\ positive}$ (false positive "added" band probability), $P_{false\ negative}$ (false negative "missing" band probability), and σ (standard deviation from true mobility) are calculated based on the precision with which the experimental data can be extracted.

This scoring system penalizes either matching a band with an error in the mobility or failing to match a band altogether. The false positive score represents the case where a band in the experimental data does not match up with a band in the expected data. The false negative score represents the case where a band in the expected data does not match up with any experimental bands. The maximum score is the log likelihood that the query fingerprint was derived from the target pattern under the assumptions of our model relative to the likelihood of assuming the same match at random. Scores are reported in units of the natural logarithm of the likelihood ratio (nats). They may be converted to bits by dividing ln(2).

Coverage

Since the sequence to be validated is known, a map of the restriction enzyme cut sites can be created for each of the restriction enzymes used in the experiments. As a result, the location of each of the expected fragments within the sequence is known. Figure 1 shows an example of the known cutting sites for the restriction enzymes BamHI, EcoRI, HindIII, and KpnI within an example sequence.

For each of the four restriction enzymes, an experimental digest has been performed independent of the other 3 enzymes. The experimental fragments are compared to the expected fragments using the previously described dynamic programming algorithm. The purpose of the algorithm is to tell which of the expected fragments are matched with an experimental fragment. A region

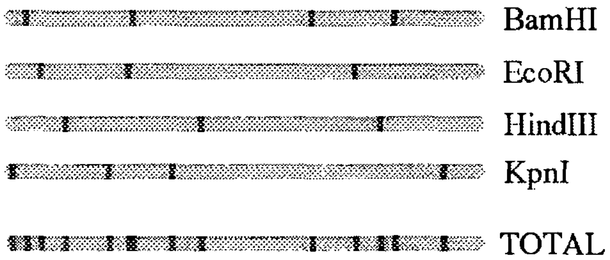


Figure 1. The sequences labeled BamHI, EcoRI, HindIII and KpnI show the location of the respective restriction enzyme recognition sites within an example sequence. The sequence labeled TOTAL indicates the location of all of the enzyme restriction sites within the sequence.

between two restriction sites in the sequence to be validated is said to be *covered* when it is matched with an experimental fragment. The results of the coverage analysis for each individual restriction enzyme can be combined to produce a total coverage map where the coverage for any particular fragment can range from 0% to 100%. When four enzymes are used, the coverage for any fragment between two restriction sites can be 0% (not covered by any individual restriction enzyme coverage map), 25% (covered by one), 50% (covered by two), 75% (covered by three), or 100% (covered by all four restriction enzyme coverage maps).

Analysis of coverage maps can indicate possible sequence assembly errors. For instance, suppose that one segment within the clone has been reversed in the sequence assembly. In such a case, we would expect two predicted restriction fragments from each digest not to be matched, resulting in a low coverage for the regions containing these fragments. The regions of low coverage contain within them the endpoints of the reversed segment.

Setting up the Simulations

Simulated restriction digest patterns were created by adding random perturbations to the computationally predicted mobilities.. The predicted mobilities were created using a subset of the palindromic 6 base restriction sites EcoRI (GAATTC), BamHI (GGATCC), HindIII (AAGCTT), BalI (TGGCCA), HpaI (GTTAAC), PstI (CTGCAG), SalI (GTCGAC), KpnI (GGTACC), NaeI (GCCGGC), and NarI (GGCGCC). The test fingerprints were compared with reference fingerprint patterns derived from sequences rearranged by introducing a segmental inversion between two randomly chosen points in the sequence. For each of the patterns, we find which target bands get matched up with an experimental band. Using this information, a coverage plot can be generated for the target sequence. By comparing the digest patterns of more than one restriction enzyme and overlapping their coverage results, it is proposed that errors in sequence assembly can be differentiated from false positive and false negative experimental bands. We ran simulations to test the effects of false positive and false negative band rates (ranging from .5% - 2%), band mobility resolution (ranging from .1% - 1%; 0.02mm - 0.2mm), and the number of restriction enzymes used. We looked at false negative rates (the percentage of time that one of the ends in the inversion is not detected by coverage analysis) and false positive rates (the percentage of time that an incorrect inversion location is detected by coverage analysis). The data presented is based on the simulations using a 219.4 kb interval derived from the human X chromosome (GenBank accession no. L44140) (Chen et al. 1996). We will focus on the results using 4 restriction enzymes for a more detailed discussion.

Experimental results have also been achieved using a HindIII digest on the bWXD718 sequencing project at the Washington University Center for Genetics in Medicine. These results are discussed as well.

Results

The Washington University Center for Genetics in Medicine and Genome Sequencing Center have been collaborating in construction of sequence ready maps and reagents for the human X chromosome, and over 1,000 clones have now been fingerprinted. The precision of fragment mass determination was 1% (M. Marra personal communication). In the early phases of this work 30 clones were sent for repeat analysis making it possible to estimate the reliability of the fingerprint data. In this preliminary data set, one discrepancy in 25 bands was observed between identical clones implying a combined false positive and false negative rate of roughly 4%. As the lab has become more experienced with fingerprint analysis, performance has improved substantially.

Increasing the Number of Restriction Enzymes

Figure 2 illustrates the use of a single restriction enzyme. Fingerprint analysis is sensitive to false positive and false negative bands. As a result, it can be impossible to differentiate between false negative bands and regions of incorrect sequence assembly. A restriction site is expected every 4^6 = 4096 bases in random sequence since 6 base restriction enzymes are used. It is well known that genomes are not randomly distributed. Thus, some restriction sites might be rare in a particular region. Two problems can result. The first is that an inversion can be missed because it has a greater likelihood of occurring between two sites where it cannot be detected. The second is that even though a region of low coverage might be detectable, a greater area might have to be considered as a possible location for the inversion.

A second enzyme can help alleviate the problem of differentiating false negatives and areas of concern. However, if the restriction enzymes are not chosen carefully, relatively long stretches where there is not a restriction site for either enzyme can still exist. Figure 3 illustrates the results using a second restriction enzyme.

Coverage analysis of our simulations suggests that the use of four or more enzymes should produce the desired results (compare Figures 2, 3, and 4). Two enzymes still present the difficulty of an inversion occurring in between two restriction sites. Experimental errors will also have some effect when only two enzymes are used. We have analyzed the results using an even number of enzymes. This is done to balance the number of A+T restriction patterns with the number of G+C restriction patterns, so as to avoid compositional biases. Figure 4 illustrates the results using four restriction enzymes. If the restriction digests are repeated when a potential region of difficulty is observed, experimental gel errors can be filtered out and differentiated from sequence assembly errors. Figure 5 illustrates this point. Note that if a single enzyme is used (as in Figure 2), the digests would have to be repeated quite often due to false negative bands.

Table I and Figure 6 examine the effects on the percentage of time that a region of faithful sequence is found to have low coverage by restriction digest fragment mapping. Figure 7 shows the percentage of time that a region that is involved in a segmental inversion is found to have high coverage. This corresponds to the fraction of the time that the rearrangement would be missed by our analysis.

Figure 2: Coverage graph using one restriction enzyme.

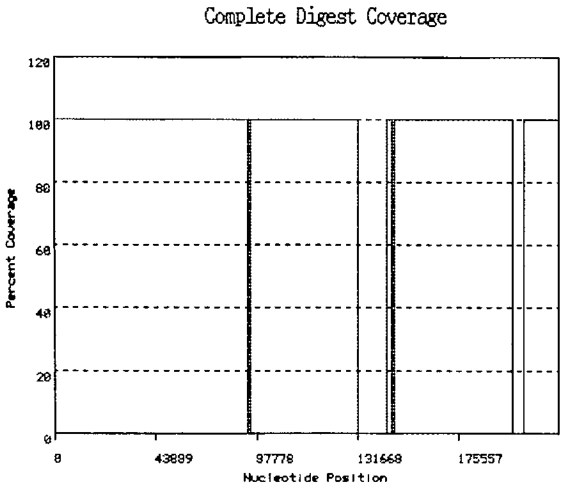


Figure 3: Coverage graph using two restriction enzymes.

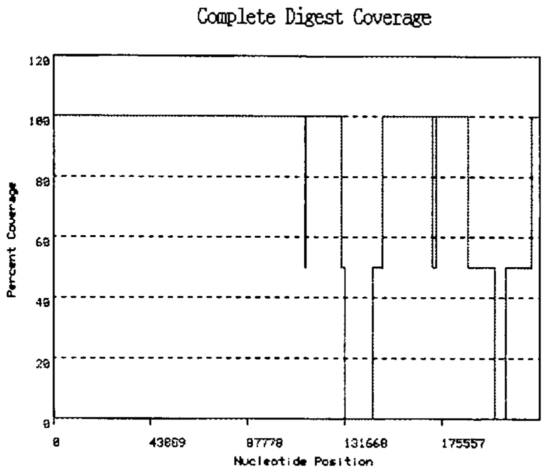


Figure 4: Coverage graph using 4 enzymes.

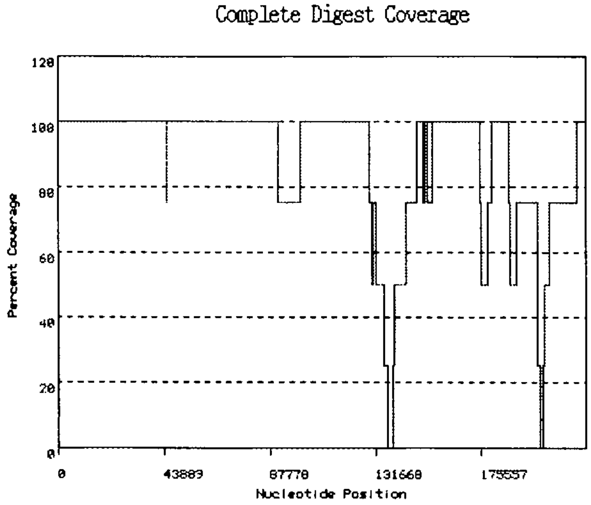
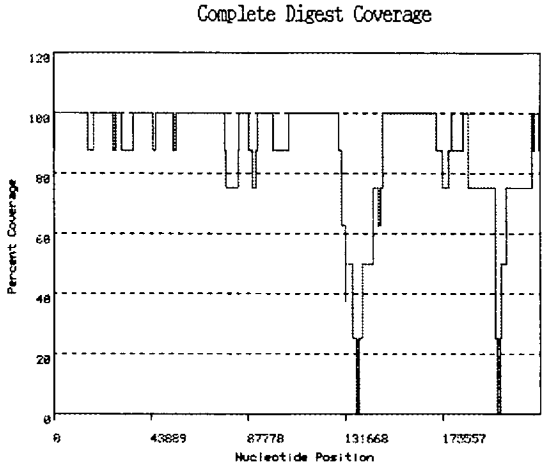


Figure 5: Coverage graph using 4 enzymes and repeating the digest analysis.



Figures 2-5: Coverage graphs. Indicated in all four figures is the coverage for the 219.4 kb region with a segmental inversion between nucleotides 136,796 and 201,014. A single restriction enzyme is used in figure 2, resulting in four regions of zero coverage. Two of these are due to experimental false negative rates, suggesting that a single enzyme is not sufficient for sequence assembly validation. When two restriction enzymes are used as in figure 3, only the two regions where the inversion occurs have zero coverage, indicating that using a second restriction enzyme improves the analysis. Figures 4 and 5 show the results using four enzymes. In figure 4, the band around the segmental inversion endpoints has shrunk to 2175 nucleotides for the left end and 1161 nucleotides for the right end. Figure 5 repeats the restriction digest. Some bands begin to have better coverage and the area surrounding the left end has shrunk from 2175 to 1286 nucleotides.

Gel Res.	False Positive Result			False Negative Result		
	.5%	1%	2%	.5%	1%	2%
0.001	4.2%	6.8%	9.9%	6.2%	3.8%	3.9%
0.0025	5.5%	7.5%	11.9%	6.1%	4.2%	5.8%
0.004	5.9%	7.2%	11.2%	2.8%	3.8%	6.3%
0.0055	4.9%	8.2%	12.6%	3%	4.9%	3%
0.007	7.5%	7.7%	13.2%	3.9%	4.6%	3.3%
0.0085	5.5%	7.2%	13.5%	5%	3.5%	5%
0.01	5.2%	8.5%	11.4%	4.3%	3.6%	6%

Table I: Empirical error rates for band assignment. The table presents the error rates for the assignment of segmental inversions to their corresponding segment of genomic sequence. The column on the far left represents experimental gel resolution values. False positives are the percentage of time that a region not involved in a segmental inversion is found to have low coverage. False negatives are the percentage of time that a region that is involved in a segmental inversion is not found. Within each section results are presented for simulations conducted with false negative and false positive band calling rates of 0.5%, 1% and 2%, and these results are presented separately. These results are based on 4 enzyme digests, each performed once, and a coverage cutoff of 50%.

Analysis of Experimental Data

One of the sequencing projects that the Washington University Center for Genetics in Medicine and Genome Sequencing Center is working on involves a region of the human X chromosome labeled bWXD718. In a preliminary assembly, the sequence appears to be 79,612 nucleotides long. The experimental HindIII digest of this clone indicates a total fragment size of 169,699 nucleotides, indicating the preliminary assembly contains errors.

All but two of the expected fragments match up with experimental fragments. The two fragments that do not match up are 558 and 145 nucleotides long. It is possible that some of the smaller fragments travel through the gel more rapidly, and thus there are greater errors, so the 558 nucleotide segment might actually map to an expected segment that is 520 nucleotides long. Also, the 145 nucleotide segment might have gone undetected in the gels. Thus, the validation program cannot discern where the problem is located, but rather alerts the biologists that there is an existing assembly problem or a molecular biological rearrangement that occurred between the fingerprint and sequence analysis stages.

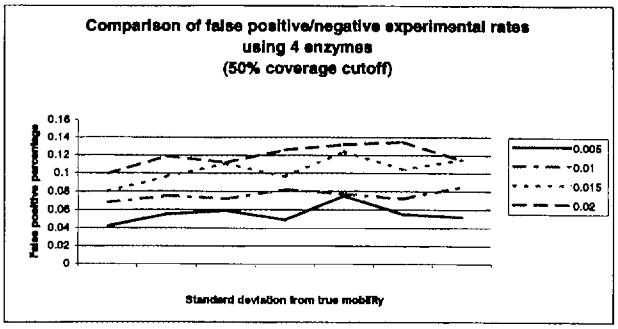


Figure 6: False positive rates. This figure corresponds to the data from Table I. The x-axis represents the standard deviation from true mobility and the y-axis represents the false positive rates. By examining this graph, we can see that the experimental false positive and false negative rates have an effect on false positives. In particular, as the experimental rates increase, so does the percentage of time that a region that is not involved in a segmental inversion is found to have low coverage. At the same time, the standard deviation from true mobility does not seem to affect the false positive percentage.

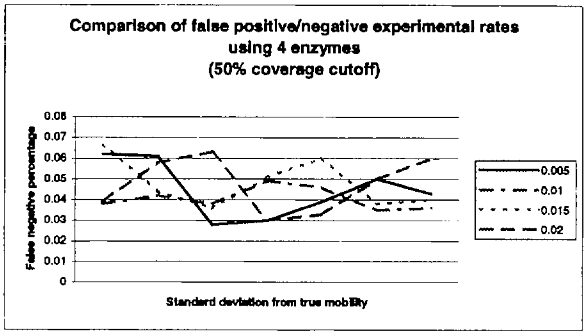


Figure 7: False negative rates. This figure corresponds to the data from Table I. The x-axis represents the standard deviation from true mobility and the y-axis represents the false negative rates. By examining this graph, we can see that the experimental false positive and false negative rates do not have much of an effect on the rate of missing a rearrangement.

Discussion

The results presented here demonstrate that it is possible to detect most sequence fragment assembly errors using a set of four restriction digests and without reference to an overlying physical map. The confidence of sequence validation can be further improved by independently repeating the digests or by using additional enzymes (data not shown). The confidence of sequence validation improves with both the resolution of the electrophoretic fragment sizing and the accuracy of band calling.

False Negatives

There are four reasons why the simulated segment inversion sites may not be determined correctly. One reason is that the inversion could occur in a segment such that it does not overlap any restriction sites. Another explanation is that the inversion occurs in such a way that the restriction sites are located near the middle of the inverted segment, resulting in similar fragment mobilities. Thirdly, an inversion occurs in such a way that the modified segments are similar to other existing segments, so coverage is preserved, albeit at a lower percentage than normal. Finally, the inversion could occur within a long repeat segment, resulting in no change with an inversion.

Application to Clone Mapping

We are currently collaborating with the Washington University Center for Genetics in Medicine and Genome Sequencing Center to use these assembly validation techniques to map locations of BAC and YAC clones within the human genome. For the purposes of our analysis, we are given both the end sequences of the clones and a set of restriction digest fragments for the enzymes BamHI, EcoRI, HindIII, and KpnI. Once we have the experimental data, the process begins by searching GenBank for homologies with the end sequences using a local sequence alignment technique. We find which, if any, of the sequences in the database have stretches of matching nucleotides longer than 30 nucleotides. We take the longest stretches and try to find a contiguous sequence connecting the two ends. If such a contiguous sequence exists, we can compare an expected digest covering this region with the experimental digests. A coverage graph of the results can then be analyzed. Such a study can be helpful because it places the clones within existing sequences, helping to determine whether or not the whole clone should be sequenced. This might help to bridge the gap between two segments. We have gathered data for the bWXD1034 and bWXD1035 sequencing projects and are in the process of assimilating the results.

Differences Between Physical Mapping and Assembly Validation

Restriction digest fingerprinting has been an effective and useful tool in physical map assembly (Riles et al. 1993; Waterston et al. 1993), but there are several critical differences between genome mapping and sequence assembly validation. In physical mapping, the problem is to identify overlapping clones by similarity in their digest patterns. The presence of one or more discrepant bands in comparing fingerprints in overlapping clones is expected. Clones are rarely the same length, rarely overlap over their full extent, and may be derived from different haplotypes in a heterogeneous population. Fingerprint matching algorithms have been developed that recognize the common features of an overlapping pair and ignore the discrepancies. False positives and false negatives in scoring the bands on a gel are readily tolerated. In physical mapping, all comparisons are made between experimental data so the precision of electrophoretic analysis is important but the absolute accuracy is not. Fragments exhibiting anomalous migration behavior in gel electrophoresis (Chastain et al. 1995) match reliably as long as their anomalous behavior is reproducible.

The goal in sequence assembly validation is to recognize the possible presence of a small number of disparities between the experimentally observed fingerprint and the pattern inferred from the sequence. Many rearrangements, such as a segmental inversion, will alter only two or three of the fragments in a digest that may contain 50 or more bands. Comparisons must be made between experimental data and theoretically derived predicted patterns so the absolute accuracy as well as the precision of mass determination are important. False positive and false negative band calls are potentially confounding and could be mistaken for fingerprint disparities resulting from an incorrect sequence assembly.

The difficulty of sequence assembly validation by fingerprint comparison increases with the size of the project being analyzed. There are several reasons for this dependence. As the size of the clone increases, the number of bands in the restriction pattern will also increase. This makes it more likely that matches will occur at random, decreasing the information content of a match. As the number of bands in the pattern increases, the number that are expected to deviate from their predicted migration behavior also increases. In a digest with 50 bands, 2 or 3 are expected to deviate from the predicted position by $P<0.05$. The number of disparities arising from a sequence rearrangement is constant while the number of uninformative bands increases. For all of these reasons, the task of assembly validation by fingerprint matching becomes more difficult as the size of the project increases. Trends in high-throughput sequencing are moving toward the use of very large insert clones (200kb BACs and YACs). It is important to be aware that experience in assembly validation based on previous generations of small (10 kb lambda) to moderate (35 kb cosmid) insert vector systems may not be applicable to the case of current BAC or YAC scale projects.

Alternative Sequence Assembly Validation Techniques

High coverage clone maps. To address the problem of experimental sequence assembly validation, several methods appear worth exploring. The first is the use of high coverage clone maps assembled from restriction fingerprint data to bin the fingerprint markers by clone content. For a map with a 5X mean clone coverage, there will, on average, be 5 clone ends and 5 clone beginnings in the interval spanned by the sequencing project of interest. These endpoints will define 10 intervals. By comparing the fingerprint content of the overlapping clones, it should be possible to assign most fragments to a unique interval. Comparing this binned set of fingerprint markers to the digest predicted from the assembled sequence will provide a more powerful test of sequence integrity. This strategy is particularly attractive because the necessary data are likely to be available as a result of clone retrieval and mapping work done prior to the initiation of sequence analysis. The strategy needs to be tested in a production setting. Phenomena such as restriction site polymorphisms in the clone libraries, errors in fingerprint band calling, and uncertainty in the physical map may confound analysis.

Multiple complete digest (MCD) mapping. Multiple complete digest (MCD) mapping (Gillett 1992; Gillett et al. 1996) is a more demanding physical map assembly process that utilizes multiple restriction enzyme digests and complete fragment accounting in the physical map assembly. MCD data should provide a powerful test of sequence assembly. Compared to single digest analysis with complete fragment accounting, MCD offers two advantages. Even if it is not possible to uniquely assign all fragments of each enzyme digest to unique intervals in an MCD map, a uniquely assigned fragment will likely cover every base in the assembled sequence for at least one enzyme digest (as we show above). A single restriction fragment map may be insensitive to some rearrangements if the fragment mass pattern for the rearranged sequence fortuitously matches the original pattern, but it is very unlikely that this will be the case for all of the enzymes in an MCD data set. MCD mapping requires the analysis of multiple enzyme digests for each clone increasing the necessary experimental work by several fold. Experimental and analytical studies are needed to determine if the additional work of multiple complete digest analysis is warranted.

Optical restriction mapping. Optical restriction mapping determines both fragment mass and order through the use of advanced microscopy technology to visualize the digest patterns for individual DNA molecules. In principle, the technique is ideally suited to the problem of assembly validation. Optical mapping is capable of determining accurate fragment masses and orders even for large insert clones (Cai et al. 1995) and requires very little input DNA, but production scale throughput remains to be demonstrated. A second alternative is the use of 2-dimensional gels (Peacock et al. 1985) in which the first dimension is a rare cutting enzyme and the second dimension is a frequent cutting (4-cutter) digest. The resulting data set is a two-dimensional fingerprint for the clone in which each column represents 4-cutter fragments derived from a rare-cutter fragment. Comparing the experimental fingerprint with a pattern predicted from the sequence would provide a powerful test of assembly validity. While only the sequenced clones need be analyzed, 2-D gel analysis is labor intensive, difficult to standardize, and difficult to run reproducibly.

Ordered Shotgun Sequencing (OSS). Finally, some sequencing strategies, notably Ordered Shotgun Sequencing (OSS) (Chen et al. 1993), incorporate high coverage intermediate length clone end sequences into the sequence assembly. The map built from these end pair overlaps serves as an intrinsic verification of assembly fidelity and can be used for assembly validation as long as this information has not already been used in assembling the project. Given the high clone coverage (typically 10X) used in OSS framework map generation, it should be possible to choose an initial tiling set of lambda clones from the framework map and to reserve the remaining lambda end pair relationships for assembly validation. Bootstrap procedures could be used to independently verify the validation.

Summary

In summary, comparison of experimental restriction digest fingerprints with inferred patterns derived from finished sequence data may identify some errors in sequence assembly, but high-resolution electrophoretic analysis and accurate scoring of bands are necessary. The problem of assembly validation by fingerprint comparison becomes more difficult as the size of the sequencing project increases. Even with state-of-the-art experimental technology, it is difficult to exclude the possibility of an undetected assembly error such as a large segmental inversion in a BAC-scale sequencing project. In the work presented here, we demonstrate that reliable validation of assembly integrity is possible using multiple restriction digests without the necessity of constructing a full MCD physical map.

Acknowledgments

We wish to thank Marco Marra and Bernard Brownstein for making experimental fingerprint data available for our use in the course of this study, and Volker Nowotny, David Politte, David Maffitt, and Thomas Blackwell for many hours of thoughtful discussion and critique.

This work was supported in part by a grant from the National Human Genome Research Institute, (HG-01391, D. States PI).

References

Cai, W., Aburatani, H., Stanton, V.P., Jr, Housman, D.E., Wang, Y.K., and Schwartz, D.C., (1995) "Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces." *Proceedings of the National Academy of Sciences USA*, **92**(11):5164-5168.

Chastain, P.D. 2nd, Eichler, E.E., Kang, S., Nelson, D.L., Levene, S.D., and Sinden, R.R., (1995) "Anomalous rapid electrophoretic mobility of DNA containing triplet repeats associated with human disease genes." *Biochemistry*, **34**(49):16125-16131.

Chen, E., Zollo, M., Mazzarella, R., Ciccodicola, A., Chen, C-N., Zuo, L., Heiner, C., Burough, F., Ripetto, M., Schlessinger, D. and D'Urso, M. (1996). "Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD." *Human Molecular Genetics*, **5**, 659-668.

Chen, E.Y., Schlessinger, D., Kere, J., (1993) "Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones." *Genomics* **17**(3):651-656.

Drury, H. A., Green, P., McCauley, B. K., Olson, M. V., Politte, D. G., and Thomas, Jr., L. J., (1990) "Spatial Normalization of One-Dimensional Electrophoretic Gel Images." *Genomics*, **8**:119-126.

Drury, H.A., Clark, K.W., Hermes, R.E., Feser, J.M., Thomas, Jr., L.J., and Donis-Keller, H., (1992) "A Graphical User Interface for Quantitative Imaging and Analysis of Electrophoretic Gels and Autoradiograms." *BioTechniques*, **12**:892-901.

Gillett, W., (1992) "DNA Mapping Algorithms: Strategies for Single Restriction Enzyme and Multiple Restriction Enzyme Mapping." Technical Report, Washington University, Department of Computer Science, WUCS-92-29.

Gillett, W., Hanks, L., Wong, G.K.S., Yu, J., Lim, R., and Olson, M.V., (1996) "Assembly of high-resolution restriction maps based on multiple complete digests of a redundant set of overlapping clones." *Genomics*, 33(3):389-408.

Maniatis, T., Jeffrey, A., and van deSande, H., (1975) "Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis." *Biochemistry*, **14**(17):3787-3794.

Marra M.A., Kucaba T.A., Dietrich N.L., Green E.D., Brownstein B., Wilson R.K., McDonald K.M., Hillier L.W., McPherson J.D., Waterston R.H. (1997) "High throughput fingerprint analysis of large-insert clones." *Genome Research,* **7**(11)**:**1072-1084.

Needleman S.B., and Wunsch, C.D. (1970) "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular Biology*, **48**:443-453.

Peacock, A.C., Bunting, S.L., Cole, S.P., and Seidman, M., (1985) "Two-dimensional electrophoretic display of restriction fragments from genomic DNA." *Analytical Biochemistry*, **149**(1):177-182.

Riles, L., Dutchik, J.E., Baktha, A., McCauley, B.K., Thayer, E.C., Leckie, M.P., Braden, V.V., Depke, J.E., and Olson, M.V., (1993) "Physical maps of the six smallest chromosomes of Saccharomyces cerevisiae at a resolution of 2.6 kilobase pairs." *Genetics*, **134**(1):81-150.

Sellers, P.H., (1974) "On the theory of computation of evolutionary distances." SIAM Journal of Applied Mathematics, **26**:787-793.

Smith, T.F., and Waterman M.S., (1981) "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* **147**:195-197.

Waterston, R.H., Ainscough, R., Anderson, K., Berks, M., Blair, D., Connell, M., Cooper, J., Coulson, A., Craxton, M., Dear, S., et al. (1993) "The genome of the nematode Caenorhabditis elegans." *Cold Spring Harbor Symposium on Quantitative Biology*, **58**:367-376.