Sequence Assembly Validation by Restriction Digest Fingerprint Comparison

Eric C. Rouchka and David J. States

WUCS-97-40

September 1997

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

Institute for Biomedical Computing
Washington University
700 S. Euclid Avenue
Saint Louis, MO 63110

Ecr@ibc.wustl.edu
States@ibc.wustl.edu

Abstract

DNA sequence analysis depends on the accurate assembly of fragment reads for the determination of a consensus sequence. Genomic sequences frequently contain repeat elements that may confound the fragment assembly process, and errors in fragment assembly may seriously impact the biological interpretation of the sequence data. Validating the fidelity of sequence assembly by experimental means is desirable. This report examines the use of restriction digest analysis as a method for testing the fidelity of sequence assembly. Restriction digest fingerprint matching is an established technology for high resolution physical map construction, but the requirements for assembly validation differ from those of fingerprint mapping. Fingerprint matching is a statistical process that is robust to the presence of errors in the data and independent of absolute fragment mass determination. Assembly validation depends on the recognition of a small number of discrepant fragments and is very sensitive to both false positive and false negative errors in the data. Assembly validation relies on the comparison of absolute masses derived from sequence with masses that are experimentally determined, making absolute accuracy as well as experimental precision important. As the size of a sequencing project increases, the difficulties in assembly validation by restriction fingerprinting become more severe. Simulation studies are used to demonstrate that large-scale errors in sequence assembly can escape detection in fingerprint pattern comparison. Alternative technologies for sequence assembly validation are discussed.

Sequence Assembly Validation by Restriction Digest Fingerprint Comparison

Eric Rouchka and David J. States

Institute for Biomedical Computing
Washington University
St. Louis, Missouri 63110

ecr@ibc.wustl.edu states@ibc.wustl.edu

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 a 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. Because 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 well established so it is easy electronically generate a set of predicted fragment masses from the finished sequence, and the predicted fragment mobilities agree well with experimental data (also ±1%). 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. Because 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 digest pattern are missed or a number of false positives are scored.

In this report, we examine the use of restriction digest fingerprints for assembly validation, and we compare the requirements for fingerprint mapping with the requirements for assembly validation.

Methods

Simulated restriction digest patterns were derived by adding random perturbations to the computationally predicted mobilities. These test fingerprints were compared with reference fingerprint patterns derived from either the correct sequence or sequences rearranged by introducing a segmental inversion between two randomly chosen points in the sequence. Fingerprint patterns were matched using a dynamic programming algorithm to generate the optimal pairing of bands in the test fingerprint to bands in the reference fingerprint. Pattern alignments were scored using a log odds system based on the likelihood of deriving the observed fragment mobilities from the predicted digest masses relative to the odds of observing the 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

$$P(m_{obs} \mid m\ ) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(m_{obs} - m\ )^2}{2\sigma^2})$$

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]. For the purposes of this work, the mobility, m, of a fragment was

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

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 0.1mm corresponds to a relative accuracy of mass determination of 0.5%.

The probability, P_{random}, of matching a band at random given a gel precision of X and N bands is

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

This scoring system penalizes either matching a band with an error in the mobility or failing to match a band altogether. The optimal 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 the ln(2).

Restriction digest patterns were generated computationally using palindromic 6 base sites. 200 trials were performed using different mobility perturbations and inversion sites for each restriction site. The distribution of scores is plotted as a histogram. False positive and false negative error rates between 0.1% and 5% were examined. For the examples shown below, the sequence was a 220 kb interval derived from the human X chromosome [Chen et al 1996]. Digest patterns were generated using the sites GAATTC (EcoRI), GGATCC (BamH1), GCCGGC (NaeI) and GGCGCC (NarI) cleavage sites. Data are shown for EcoRI. Similar results were obtained for all cleavage sites.

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 of 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 experience with fingerprint analysis, performance has improved substantially. For the purpose of this analysis, false positive and false negative error rates of 0.1% to 1% were considered.

As is shown in figure 1, deviation in band mobilities from their reference values degrades the match score significantly (from 148 log units to a mean of 88 log units). Even for the worst case of this self-comparison, the match is still highly informative (less than 1 in 10^{29} chance of occurring at random). If the purpose of this test was to identify clones sharing overlapping regions (e.g. physical mapping), such a match would be highly significant in screening a full human genome library.

As is shown in figure 2 and table 1, it is not possible to reliably distinguish the digest of the parental clone from the digest of a rearranged clone (segmental inversion) on the basis of state-of-the-art experimental data. The distribution of match scores for simulated digests derived from a rearranged sequence overlap extensively the scores derived from matches to fingerprints derived from correct sequence.

Table 1).

	Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
Correct Sequence	69.48	75.57	87.43	88.25	97.82	124.3
Segmental Inversion	17.96	58.7	70.86	69.49	82.56	105.8

The table summarizes the distribution of match scores for comparison of simulated digest patterns to predictions derived from the correct sequence or from a sequence which includes a random segmental inversion. No significant difference in the distributions is observable. Data are based on an EcoRI digest of the 220 kb region around G6PD [Chen et al, 1996] with an accuracy of mass determination of 0.5% and false positive and false negative band calling error rates of 1% each.

As figure 3 demonstrates, it is difficult to reliably distinguish band predictions based on faithful sequence from predictions based on rearranged sequence even when test data of very high quality are used (gel resolution of 1,000:1, 1 in 1,000 false positive and false negative error rates).

This analysis assumes that the stoichiometry of band intensity is known, and the dynamic programming alogrithm requires a one to one matching of bands between the reference and test digest patterns. If stoichiometry is not known or can not be established reliably from the experimental data, additional degrees of freedom are introduced into the problem, and it will be even more difficult to distinguish fingerprints of rearranged sequence from the fingerprint of the correct sequence.

Discussion

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 physical 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 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 in 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 increases. 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.

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 errors in the physical map may confound analysis.

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 with 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, every base in the assembled sequence will likely be covered by a uniquely assigned fragment for at least one enzyme digest. 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 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. To be useful, scale up in processing capacity and reliability need to be achieved. At present, optical mapping is being practiced at only a single laboratory and sustained high-throughput analysis of large insert clones has not been 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 represent 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.

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 independent verify the validation.

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. For the future, alternative methods for assembly need to be explored.

Acknowledgements

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 David Politte, David Maffitt, and Thomas Blackwell for many hours of thoughtful discussion and critique.

Bibliography

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.

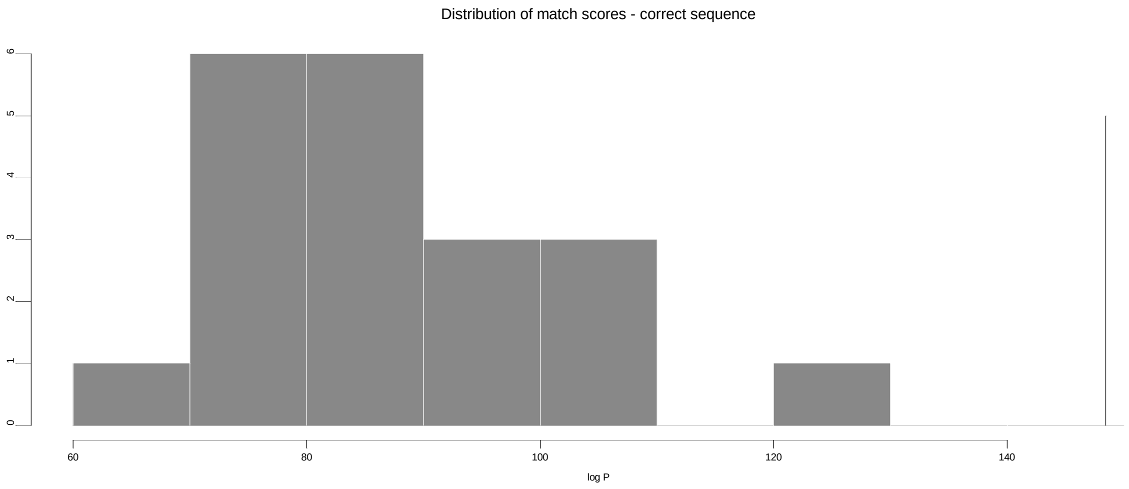
Peacock, A.C., Bunting, S.L., Cole, S.P., and Seidman, M., (1985) “Two-dimensional electrophoretic display of restriction fragments from genomic DNA.” Analytical Biochemistry, 149(1):177-182.

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.

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.

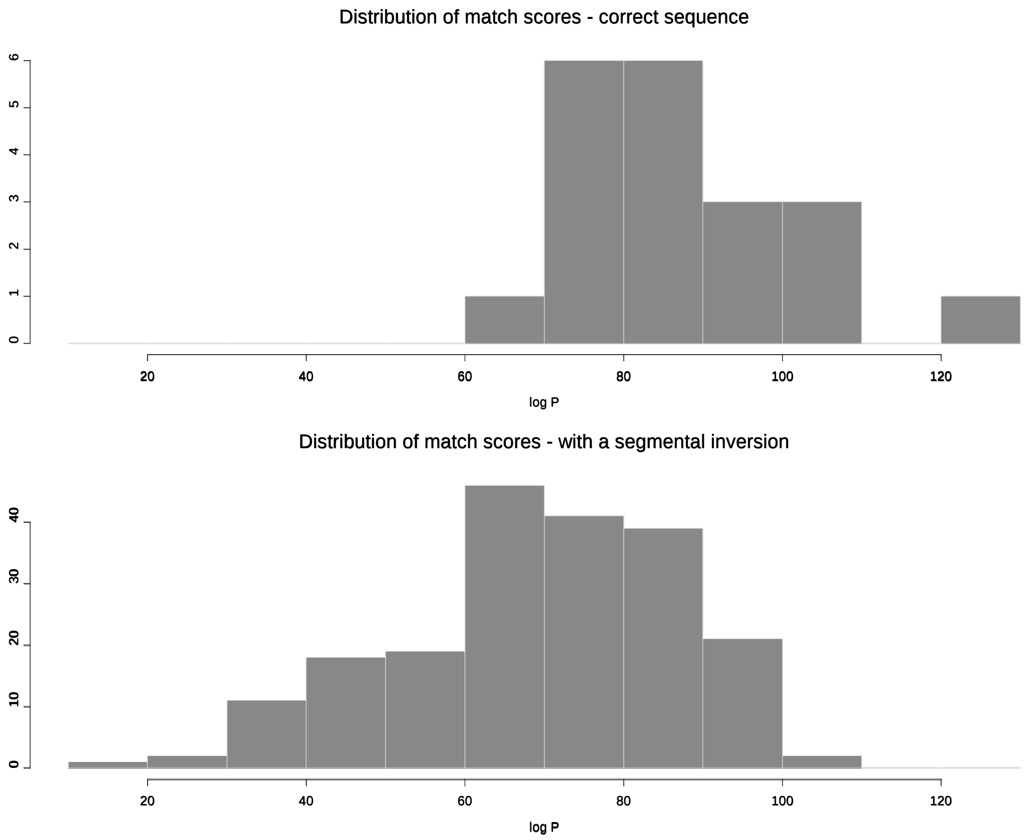
Figures

Figure 1).



Shown in the figure is the distribution of match scores for simulated digests of a 220 kb interval of human around the G6PD locus assuming a mobility range from 0 to 20 cm and an absolute accuracy of 0.1 mm in fragment position. The line at the right indicates the score that would be achieved with perfect fragment mass determination.

Figure 2).

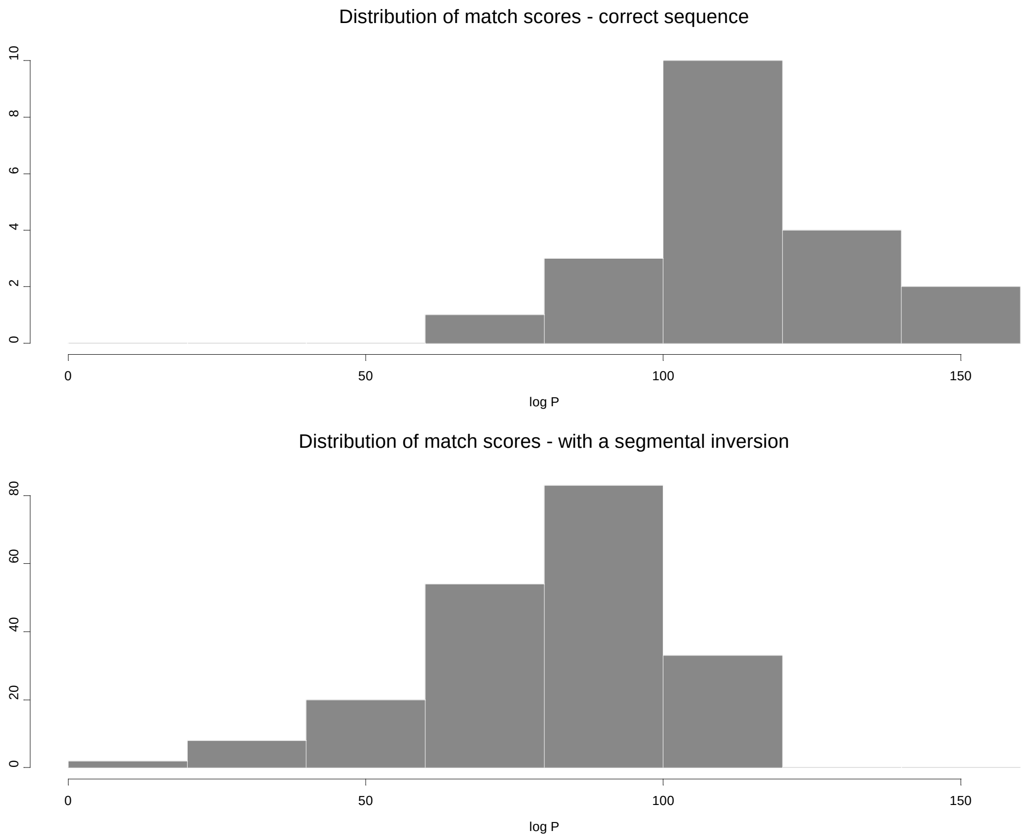


Distribution of match scores - correct sequence

Distribution of match scores - with a segmental inversion

Shown in the figure is a comparison of the match scores for simulated EcoRI digests derived from the correct sequence (top panel) and from sequences in which a segmental inversion between two randomly chosen points has been applied to the sequence. For this simulation, the absolute accuracy of fragment position was assumed to be 0.1 mm (200:1 resolution, 0.5% mass accuracy) and the frequency of false negative and false band calls was 1%.

Figure 3)



Distribution of match scores - correct sequence

Distribution of match scores - with a segmental inversion

Shown in the figure is a comparison of the match scores for simulated EcoRI digests derived from the correct sequence (top panel) and from sequences in which a segmental inversion between two randomly chosen points has been applied to the sequence. For this simulation, the absolute accuracy of fragment position was assumed to be 0.05 mm (400:1 resolution, 0.25% mass accuracy) and the frequency of false negative and false band calls was 1 per thousand.