W108–W112 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh450

A web server for performing electronic PCR

Kirill Rotmistrovsky, Wonhee Jang and Gregory D. Schuler*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20984, USA

Received February 27, 2004; Revised and Accepted April 21, 2004

ABSTRACT

'Electronic PCR' (e-PCR) refers to a computational procedure that is used to search DNA sequences for sequence tagged sites (STSs), each of which is defined by a pair of primer sequences and an expected PCR product size. To gain speed, our implementation extracts short 'words' from the 3′ end of each primer and stores them in a sorted hash table that can be accessed efficiently during the search. One recent improvement is the use of overlapping discontinuous words to allow matches to be found despite the presence of a mismatch. Moreover, it is possible to allow gaps in the alignment between the primer and the sequence. The effect of these changes is to improve sensitivity without significantly affecting specificity. The new software provides a search mode using a query STS against a sequence database to augment the previously available mode using a query sequence against an STS database. Finally, e-PCR may now be used through a web service, with search results linked to other web resources such as the UniSTS database and the MapViewer genome browser. The e-PCR web server may be found at www.ncbi.nlm.nih.gov/sutils/e-pcr.

INTRODUCTION

A major milestone in the history of genome map construction was the notion of a sequence tagged site (STS), which is defined by a pair of oligonucleotide primers that can be used in a PCR to amplify a unique site within the genome (1). STS markers have formed the basis for virtually all physical and genetic maps constructed over the last decade, rapidly replacing the earlier generation of cloned DNA segment markers. PCR primer pairs can also be used to probe the transcriptome, yielding large-scale profiles of gene expression. In an era when the large-scale sequencing of genomes and transcriptomes is routinely undertaken, there is significant utility in being able to cross-reference large collections of PCR primer pairs and sequences.

We have previously described the concept of 'electronic PCR' (e-PCR) as a computational procedure for finding sequence tagged sites within DNA sequences and provided an efficient implementation of this procedure (2). To gain speed we employed the commonly used strategy of hashing, in which the bases from a window of size *W* (a 'word') are used as an index into a hash table (for an overview of program parameters, see Table 1). Each time a matching word is found, a portion of the sequence is checked for an alignment to the corresponding primer. Finally, a match is reported if both primers are found in the correct orientation and imply a product size that is within *M* bases of the expected size (Figure 1). Increasing the value of *W* accelerates the search by reducing the background of word matches that must be investigated. In the original implementation of the program, only one word was hashed per primer and the requirement that its *W* contiguous bases match exactly led to a loss of sensitivity. Even though mismatches were allowed in the primer alignment step, their presence within the hashed word was sufficient to deny a match.

In previous reports, we have described applications of e-PCR for binding genomic and transcribed sequences to map positions (2) and for using STS maps to assess the quality and completeness of a genomic sequence (3). Here we describe algorithmic changes that result in improved sensitivity, a search mode in which a query STS can be compared to a sequence database, and a web server for performing e-PCR.

Table 1. Electronic PCR program parameters

Parameter	Meaning
W	Number of bases used as a word for hashing
F	Number of discontiguous words hashed (*F* = 0 for contiguous)
N	Number of mismatches allowed in primer alignment
G	Number of gaps allowed in primer alignment
M	Number of bases the STS size may differ from expected size

Additional options may be added in the future. Invoking the program with -h as the argument produces a list of all available options.

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: schuler@ncbi.nlm.nih.gov

SEARCH SENSITIVITY

The original e-PCR implementation was fairly rigid in its match criteria, but there are a number of reasons why a more 'fuzzy' matching strategy might be desired. For example, when searching against single-pass (low-quality) sequences, such as expressed sequence tags and clone end sequences, a certain rate of sequencing error is expected. Alternatively, the sequence may be free of errors and the goal to find near-matches that may cause confounding signals. In any event, depending on the particular bases involved, the PCR biochemical reaction may tolerate some mispairing within the primers. To improve search sensitivity, we have modified both the hashing and the primer alignment steps of e-PCR.

To reduce the likelihood that a true STS will be missed due to mismatches, we have changed the way in which hash table values are generated. Instead of using a single exact word, multiple, discontiguous words are used, each of which has groups of significant positions separated by 'wildcard' positions that are not required to match. They have been variously called 'templates', 'patterns' and 'motifs' and have been used previously in DNA database searching (4–6) and multiple protein sequence alignment (7). In e-PCR, the *F* parameter specifies the number of words generated as well as the spacing of the wildcard positions. For example, using $F = 3$ (as in Figure 1b), the wildcards occur every third position. By having this template successively shifted by one position in each of the three words, every base corresponds to a wildcard in some word. Thus, a word match is guaranteed for any case where there is just one mismatch. Two or more mismatches will still pose a problem (except in the fortuitous case where their spacing is a multiple of *F*). However, as will be shown later, allowing more than one mismatch greatly increases the number of false positives.

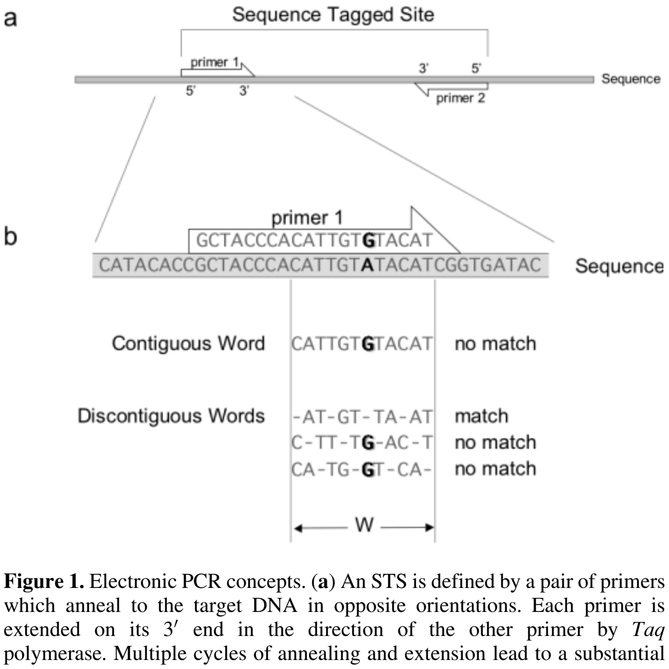

Figure 1. Electronic PCR concepts. (**a**) An STS is defined by a pair of primers which anneal to the target DNA in opposite orientations. Each primer is extended on its 3′ end in the direction of the other primer by *Taq* polymerase. Multiple cycles of annealing and extension lead to a substantial amplification of the STS sequence, also known as the 'amplicon'. (**b**) By default, a single contiguous word of *W* bases ($W = 9$ in this example) is extracted from the 3′ end of each primer. Any mismatch or gap within this region is sufficient to eliminate a match. With discontiguous words enabled ($F = 3$), three words are indexed, each with a 'wildcard' position every third base. Staggering of the wildcard positions ensures that no single mismatch in this region will deny a match.

The primer alignment step that is invoked following each word hit has been modified to allow gaps (insertions or deletions) in the alignment. This feature is enabled by the *G* parameter, whose value specifies the maximum number of gaps allowed in each primer. Although the algorithm does not place any constraints on where the gap may be, it is important to note that gaps within the *W* bases used for hashing will generally prevent getting a word hit. Allowing gaps is very useful when searching low-quality sequences or when using primers designed from low-quality data. However, this option will also increase the running time and may generate false positives.

Given these modifications for improving sensitivity, it is of interest to see how effective they are and to what extent they affect specificity. To test this, we chose a set of 584 microsatellite STSs that were used as reference markers for the human transcript map (8,9). They have all been mapped with high confidence, and in a consistent order, by meiotic linkage mapping (10) and radiation hybrid (RH) mapping using two different RH panels, the GeneBridge 4 panel (11) and the Stanford G3 panel (12). Thus, we may be fairly certain that they represent unique sites in the genome. However, one caveat is that if a site appears multiple times within a window smaller than the map resolution, it would have appeared as a unique site in these mapping studies. Of the three mapping resources, Stanford G3 is the most precise, with an average resolution of ∼500 kb (12). In fact, we found two instances of STSs reacting with multiple sites that were <500 kb apart and we treated each pair as a single site. If an STS was found only once in the genome, and on the expected chromosome, it was assumed to be a true positive. For those that hit multiple times, one was counted as a true positive (if on the correct chromosome) and all the rest were counted as false positives. Of course, any STS not found was regarded a false negative.

We compared the test STS set to the complete sequence of the human genome using various e-PCR parameters to see how specificity and sensitivity would be affected. Table 2 provides the numbers of true and false matches, the number of STSs not found and calculations of sensitivity (fraction of STSs reporting a match) and specificity (fraction of matches that are true). Not surprisingly, a specificity of 1.0 is obtained when primers are required to match exactly ($N = 0$, $G = 0$), but 80 of the markers were not found, yielding a sensitivity of only 0.863. Although the human genomic sequence is known to contain gaps, a recent analysis suggests that it includes 99% of the euchromatin (The International Human Genome Sequencing Consortium, manuscript submitted), suggesting an upper bound of ∼0.990 on sensitivity. Using discontiguous words and allowing one mismatch and one gap per primer gives the best balance of sensitivity and specificity (0.983 and 0.991). Allowing two gaps resulted in only one additional STS being found but more than doubled the number of false positives. The most drastic loss of specificity is seen when two mismatches are allowed. The five markers that were not found under any conditions were investigated further and found to have alignment gaps within the hashed *W* bases. In these comparisons, the word size was kept constant ($W = 12$). It should be noted that for tests using discontiguous words, the

Table 2. e-PCR sensitivity and specificity with different search options

Search parameters Word type	Mismatches allowed	Gaps allowed	Search results True positives	False positives	False negatives	Sensitivity	Specificity
Contiguous	0	0	504	0	80	0.863	1.000
Contiguous	1	0	543	0	41	0.930	1.000
Contiguous	1	1	556	3	28	0.952	0.995
Discontiguous	1	0	560	0	24	0.959	1.000
Discontiguous	1	1	574	5	10	0.983	0.991
Discontiguous	1	2	575	14	9	0.985	0.976
Discontiguous	2	1	578	172	6	0.990	0.771
Discontiguous	2	2	579	874	5	0.991	0.398

Tests were conducted with a set of 584 microsatellite STSs with a consistent order among three maps. They were compared to the complete human genome (build 34; July 2003) using $W = 12$ and $M = 200$. Discontiguous words were activated with $F = 3$ and the number of mismatches and gaps allowed were varied using the N and G parameters (Table 1). Sensitivity (Sn) is defined as $Sn = TP/(TP + FN)$, where TP is true positives and FN is false negatives. Specificity (Sp) is defined as $Sp = TP/(TP + FP)$, where FP is false positives.

'effective word size' is eight (excluding four wildcard positions), which causes the program to run more slowly but does not significantly change the results shown in Table 2. It should be noted that these tests were performed using a set of well-mapped markers, which will surely have different properties from those chosen at random.

REVERSE SEARCHING

Although the original e-PCR program constructed a hash table from the STS database, there are situations in which it is more desirable to hash the sequence database. We have implemented this strategy and refer to it as 'reverse e-PCR'. Conversely, the previous use of a sequence query against an STS database would be 'forward e-PCR'. The main motivation for implementing reverse e-PCR was to make it feasible to search the human genome sequence (and other large genomes) in an interactive web service. Before performing a reverse e-PCR search, the sequence database must be processed using a specific word size and discontiguous word option. Sequences are scanned, examining each word in turn, and ultimately a data structure is created in which each possible word has an associated list of all sequence coordinates (pairs of sequence identifier and base position) at which it is found. This step is time-consuming, but only needs to be done once (unless the underlying sequence changes). Thereafter, an STS can be compared against the genome by extracting a few words from each primer, retrieving lists of positions to examine and reading only the necessary portions of the sequence into memory to test for primer alignments. However, the index—actually a set of several files organized for efficient memory mapping—requires storage that is ∼10–15× the size of the original sequence database. In other words, space is traded to gain time.

It should be noted that speed will degrade significantly when a primer contains a highly repetitive word. This is due to the fact that its list of sequence coordinates will be large, and following up on each one requires reading a segment of sequence data from the storage device. In the initial indexing of the database, it is possible to identify words that occur too frequently and simply mark them as repetitive rather than storing all of their positions. This seems reasonable because, as a general rule, users will only want to know that a candidate marker is repetitive (so that it can be eliminated from further consideration) and not see a full account of its positions. Eliminating repetitive word coordinates from the index both increases speed and decreases storage requirements. However, it also may result in occasionally missing a true STS because it is possible that the W bases used for the lookup are repetitive, even if the primer as a whole is not.

Table 3. Relative running times and storage requirements for forward and reverse e-PCR

Datasets	Forward e-PCR		Reverse e-PCR	
	Time (s)	Space (MB)	Time (s)	Space (MB)
Small sequence (270 kb)				
Versus single STS	4	<1	3	6
Versus small STS set	4	<1	16	6
Versus large STS set	11	12	78	17
Large sequence (2865 kb)				
Versus single STS	1178	2906	38	35 837
Versus small STS set	1161	2906	155	35 837
Versus large STS set	54 540	2917	n.d.	35 844

The large sequence dataset is the human genome and the small sequence is GenBank entry AB026898. The large STS database consists of all 132 648 non-repetitive human markers from UniSTS and the small set is a group of 13 markers found within AB026898. All tests were conducted using discontiguous words of size 12. n.d.: the time for searching a large sequence against a large STS set using reverse e-PCR was not determined exactly, but is estimated to take ∼10 days.

To provide a better sense of which search strategy is most appropriate for certain situations, we have devised a series of test cases using sets of sequences and STSs of different sizes. The large sequence database consists of all of the sequence contigs from the 2.86 Gb human genome sequence. For the small sequence, we chose a single 270 kb sequence entry (AB026898) corresponding to a region of human chromosome 3. This represents ∼1/10 000 of the genome and falls at the high end of a typical size distribution for large-insert clones. A set of 132 648 non-repetitive human markers from UniSTS constitutes the large STS database. Of these, a set of 13 (again, 1/10 000 of the large set) markers (all of which fall within AB026898) was chosen as the small STS database. In addition, the program was run using just a single STS (marker D3S3333). The time and disk space required for each situation are shown in Table 3. Overall, reverse e-PCR is faster when using small numbers of STSs, while forward e-PCR is better with larger STS sets. Comparing a single STS to the human genome required 38 s with reverse e-PCR, about a 40-fold increase in speed compared to the equivalent search with forward e-PCR. However, with the small (13 marker) set, the advantage is closer to 10-fold, and with the large set, the reverse search takes too long to be feasible. The basis for any performance benefit with reverse e-PCR lies in avoiding a scan of the entire genome. As the number of STSs increases, we rapidly approach the situation where most of the database must be examined. Furthermore, there are more data involved and they are retrieved in a random-access fashion rather than sequentially. Consequently, reverse e-PCR is best suited for its intended use in interactive searching of large sequences. It should be appreciated that the actual search times that may be expected with the web service may vary significantly due to system load, network latency and properties of the data.

THE e-PCR WEB SERVER

Although users may download the e-PCR software and apply it to any dataset they wish, there are a number of advantages to having a centralized web server dedicated to this task. Though not particularly difficult, downloading, installation and maintenance of the software are chores that occasional users will probably want to avoid. A more significant issue is the maintenance of the STS and sequence databases, which, as we have seen, may require substantial disk resources. We have developed an e-PCR web server on the NCBI site, which provides a comprehensive STS database, UniSTS (www.ncbi.nlm.nih.gov/genome/unists), and DNA sequence datasets for the genomes and transcriptomes of several well-studied organisms. Another advantage of a web-based implementation is that results can be linked to related resources and more sophisticated views can be easily provided. As described in more detail below, e-PCR results may be linked to UniSTS, the NCBI MapViewer [(13) available from http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books] and UniGene [(14) available from http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books]. Entry through the e-PCR home page (www.ncbi.nlm.nih.gov/sutils/e-pcr) provides an overview of the resource and links to forward e-PCR and reverse e-PCR search forms.

In preparation for a forward e-PCR search, the user must specify one or more query sequences, either by pasting the actual sequence data (FASTA format) or by entering GenBank accession numbers. The only STS database provided is UniSTS, which is a comprehensive collection covering all species, with data inputs from both GenBank STS sequence entries and published STS maps (as of May 2004, the database contained 265 380 distinct primer pairs). There are provisions for changing any of the parameters shown in Table 1 as well as an option to exclude STSs, from the database that have been flagged as too repetitive. Once the search is complete, results are presented in tabular form. For each STSs found, the position within the query sequence, the marker name, the chromosome (if known) and the species of origin are given. Each marker has a hypertext link to the corresponding UniSTS entry, which provides the primer sequences and expected product sizes, alternate names by which the marker may be known, mapping results and additional pre-computed e-PCR results.

Setting up a reverse e-PCR search requires entering one or more STSs and selecting a sequence database. It is mandatory that a species be selected, together with a choice of either genome or transcriptome. It is possible to change the values for some of the parameters, but the choices are limited for *W* and *F* because they are fixed at the time the sequence database is hashed. Several interfaces are provided for entering STS information, using either separate input fields or a single text area into which formatted information can be pasted. Once the search is complete, the results are summarized in a tabular format, giving the number of hits for each marker. The software also performs a lookup of the primers in UniSTS to determine if any of them correspond to markers that have already been developed. When searching a genome sequence, each hit has a link to a graphical display in the NCBI Map Viewer, where it is possible to see where the STS is found relative to other annotated features. When the transcriptome option is used, each hit is linked to a Gene or UniGene database entry.

DISCUSSION

With the genomic sequence in hand, a major application of e-PCR is the integration of legacy maps with the sequence. By doing so, all STSs—whether they come from a high-resolution clone-based map of a disease susceptibility locus or radiation hybrid map of the whole genome—can be placed in a common coordinate system. Indeed, the STS track presented in the NCBI Map Viewer is generated using e-PCR to compare all UniSTS entries to the genomic sequence. Furthermore, comparison of this computationally generated STS map with experimentally determined genetic and physical maps provides a certain level of validation of the genome assembly. However, it should be noted that only gross rearrangements are likely to be found given the resolution of these maps.

Integration of genetic linkage maps with the genomic sequence has the added benefit of allowing regions of particularly high and low rates of meiotic recombination to be identified. This is of interest because recombination can have a profound effect on the evolution of chromosomal segments. In a previous study (15), e-PCR was used to localize polymorphic STSs from a human linkage map (16) within an older ('working draft') version of the human genome (17). By looking at the ratio between genetic distances measured in centiMorgans (cM, defined as 1% recombination) and physical base-pair distances, several recombination 'deserts' (low) and 'jungles' (high) were identified. It was noted that regions of linkage disequilibrium extended for greater distances in the deserts than in the jungles.

The e-PCR program has increasingly important applications to the process of designing new PCR primer pairs. Primers are usually chosen using software, such as Primer3 (18), that selects DNA oligos with a desired melting temperature and applies various heuristics to avoid problems such as low-complexity sequences and self-annealing primer pairs. A useful adjunct to this process is to use e-PCR to compare the chosen primers to the genomic sequence. Primers that match multiple locations in the genome can be discarded before expending resources synthesizing the oligos and using them in an experiment. This is particularly important given the trend toward construction of large arrays containing tens of thousands of PCR products, which are commonly used to study gene expression patterns or to identify DNA-binding factors.

SOFTWARE AVAILABILITY

The e-PCR software is in the public domain and source code is freely available by FTP from ftp://ftp.ncbi.nlm.nih.gov/repository/e-PCR/. The code is compatible with, but does not require, the NCBI C++ Software Toolkit (19), available from http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books]. Entry to the web services is through http://www.ncbi.nlm.nih.gov/sutils/e-pcr/.

ACKNOWLEDGEMENTS

We would like to thank the reviewers and the many users of e-PCR who have reported bugs and have made helpful suggestions for improvements.

REFERENCES

1. Olson,M., Hood,L., Cantor,C. and Botstein,D. (1989) A common language for physical mapping of the human genome. *Science*, **245**, 1434–1435.
2. Schuler,G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.*, **7**, 541–550.
3. Schuler,G.D. (1998) Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.*, **16**, 456–459.
4. Califano,A. and Rigoutsos,I. (1993) FLASH: a fast look-up algorithm for string homology. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1**, 56–64.
5. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
6. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D., Miller,W., Ma,B., Tromp,J. *et al.* (2003) Human–mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
7. Posfai,J., Bhagwat,A.S., Posfai,G. and Roberts,R.J. (1989) Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res.*, **17**, 2421–2435.
8. Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tomé,P., Aggarwal,A., Bajorek,E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.
9. Deloukas,P., Schuler,G.D., Gyapay,G., Beasley,E.M., Soderlund,C., Rodriguez-Tome,P., Hui,L., Matise,T.C., McKusick,K.B., Beckmann,J.S. *et al.* (1998) A physical map of 30,000 human genes. *Science*, **282**, 744–746.
10. Dib,C., Faure,S., Fizames,C., Samson,D., Drouot,N., Vignal,A., Millasseau,P., Marc,S., Hazan,J., Seboun,E. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.
11. Gyapay,G., Schmitt,K., Fizames,C., Jones,H., Vega-Czarny,N., Spillett,D., Muselet,D., Prud'homme,J.F., Dib,C., Auffray,C. *et al.* (1996) A radiation hybrid map of the human genome. *Hum. Mol. Genet.*, **5**, 339–346.
12. Stewart,E.A., McKusick,K.B., Aggarwal,A., Bajorek,E., Brady,S., Chu,A., Fang,N., Hadley,D., Harris,M., Hussain,S. *et al.* (1997) An STS-based radiation hybrid map of the human genome. *Genome Res.*, **7**, 422–433.
13. Dombrowski,S.M. and Maglott,D. (2002) Using the Map Viewer to explore genomes. In McEntyre,J. (ed.), *The NCBI handbook* [Internet]. National Library of Medicine (US), National Center for Biotechnology Information, Bethesda (MD).
14. Pontius,J.U., Wagner,L. and Schuler,G.D. (2002) UniGene: a unified view of the transcriptome. In McEntyre,J. (ed.), *The NCBI handbook* [Internet]. National Library of Medicine (US), Bethesda (MD).
15. Yu,A., Zhao,C., Fan,Y., Jang,W., Mungall,A.J., Deloukas,P., Olsen,A., Doggett,N.A., Ghebranious,N., Broman,K.W. *et al.* (2001) Comparison of human genetic and sequence-based physical maps. *Nature*, **409**, 951–953.
16. Broman,K.W., Murray,J.C., Sheffield,V.C., White,R.L. and Weber,J.L. (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, **63**, 861–869.
17. The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
18. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
19. NCBI (2003) *The NCBI C++ Toolkit* [Internet]. National Library of Medicine (NLM), Bethesda, MD.