

---

# Enhancing BLAST Performance by Using the Paracel Filtering Package

---

Cecilie Boysen and Marc A. Rieffel

Paracel, Inc.  
1055 East Colorado Blvd.  
Suite 410  
Pasadena, CA 91106  
USA

**PARACEL**<sup>®</sup>  
Applied High-Performance Computing

© Copyright 2004 Paracel Inc.

This document is the proprietary property of Paracel Inc., and is protected under federal copyright law, with all rights reserved. No part of this document may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior written consent of Paracel Inc.

Paracel BLAST is derived from NCBI BLAST which is in the public domain. Paracel owns all rights to the Paracel BLAST software product, as it results from additions, modifications and/or deletions to and from the original NCBI source code.

Paracel, Inc. is a wholly-owned subsidiary of Applera Corporation through its business unit, Celera Genomics Group.

Paracel Filtering Package™ is a trademark of Paracel Inc.

Paracel® is a registered trademark of Paracel Inc.

# Enhancing BLAST Performance by Using the Paracel Filtering Package

Cecilie Boysen and Marc A. Rieffel<sup>†</sup>

Using the Paracel Filtering Package (PFP) to mask genome-wide repeats and low-complexity regions, instead of the default NCBI BLAST masking of low-complexity regions, offers significant improvements to the BLAST search. PFP filters contaminants and repetitive regions from query sequences which eliminates undesired hits against these contaminants and repetitive regions from the results of the BLAST search. PFP also allows BLAST searches to succeed that would have otherwise failed if repetitive regions had not been removed. The integration of PFP into Paracel BLAST condenses the two-step process of using PFP to clean the query and then using the cleaned query file in the BLAST search into a one-step automatic process.

## 1. Introduction

The Basic Local Alignment Search Tool (BLAST) was developed by the National Center for Biotechnology Information (NCBI) to accelerate protein and DNA sequence similarity searches (Altschul *et al.*<sup>1,2</sup>). BLAST can detect weak but biologically significant sequence similarities and is considerably faster than other heuristic algorithms. NCBI BLAST includes the option of masking low-complexity regions in the query sequences before performing the search to eliminate undesired hits against such regions.

NCBI BLAST has its limitations, however, especially for very large query sequences and for very large databases: searches either take too long to complete or fail after exhausting available memory resources on conventional computers.

As the size of sequence-comparison BLAST searches grows, pharmaceutical, biotechnology and academic biologists and bioinformaticians have increasingly recognized the need to have access to large computational resources with suitable accompanying software. Paracel has addressed this need by developing *Paracel BLAST*.

Furthermore, NCBI BLAST's default masking of low-complexity regions does not remove all undesired repeats, such as vector sequences and genome-wide repeats, from the query sequences, which may produce many undesired hits. Paracel has solved this problem by developing the *Paracel Filtering Package* (PFP).

In this paper, we will demonstrate how the integration of PFP into Paracel BLAST provides an easy-to-use, one-step solution and results in significant improvements to the BLAST search.

## 2. Paracel BLAST

Paracel BLAST was developed to overcome the memory, sequence size, and efficiency problems of NCBI BLAST. Paracel BLAST software is an enhancement of NCBI BLAST that is capable of executing searches on multiple, non-shared-memory processors simultaneously. Paracel BLAST 1.6<sup>3</sup> is designed to run on SPARC systems running SunOS 5.6 (or higher) or on Linux clusters using Intel Xeon or AMD Opteron processors. Paracel BLAST delivers superior performance for large-scale BLAST searching by incorporating a number of optimizations that facilitate searches with large query sequences, large databases, and large numbers of small query sequences. The Paracel BLAST software, when deployed on a multi-processor cluster systems, can solve biologically relevant large-scale problems faster, more cost-effectively, and more conveniently than competitive software/hardware combinations. For further details on these optimizations, please refer to the *Paracel BLAST User Manual*.<sup>4</sup>

## 3. The Paracel Filtering Package (PFP)

Unfiltered repeats in query sequences can yield many undesired hits when searching against unfiltered databases. These repeats are an intrinsic part of the actual sequence. They create false-positive hits due to their abundance in the genome. Approximately 45-50% of the human genome, for example, consists of these repeats. Another problem that occurs in searches of unfiltered sequences is the presence of vector and other contaminants.

The presence of contaminants and repeats in searches can significantly skew results. Since repeat regions are often highly conserved, their high degree of similarity can obscure more distantly related sequences, making sensitive searches virtually

<sup>†</sup> Corresponding author: [marc@paracel.com](mailto:marc@paracel.com)

impossible. This manifests itself in the search results whereby the output will contain dozens of high-scoring alignments that the user would have to sift through to find the desired results. If the number of returned results is limited, sometimes the desired results may not even appear in the output. Similarly, matches to contaminant sequences are not desirable.

To address the problem of low-complexity, intrinsic repeats (e.g. poly-A, GTGTGT...), NCBI BLAST offers filtering of query sequences using DUST (for nucleotides) (Tatusov and Lipman<sup>5</sup>) and SEG (for proteins) (Wootton and Federhen<sup>6</sup>). Each query sequence is automatically screened with the appropriate algorithm before it is compared to the database. However, this approach is not always sufficient. While it eliminates hits between low-complexity regions, it does not filter complex repeats such as SINEs (Short Interspersed Elements) and LINEs (Long Interspersed Elements) that appear in nucleotide sequences. These are even more prevalent than low-complexity regions and, as we will see below, are also detrimental to search results. Moreover, DUST and SEG do not remove contaminants.

Paracel has solved the problem of screening all repeats with the Paracel Filtering Package (PFP), a full-featured application that automates the multistep screening process necessary for repeat-dense genomes. PFP takes a file of sequences and passes it through user-defined stages to remove repetitive or contaminant elements from each sequence in the file. Users may choose to either filter (remove entire sequences), mask, or trim contaminants, repeats, and low-complexity regions. For example, PFP masks repetitive elements from each sequence in the file by replacing the characters in the repetitive region with a user-specified character such as X. For nucleotide sequences, PFP typically first uses DUST to screen the low-complexity regions, and then compares the sequences to GIRI's RepBase repeats database<sup>7</sup> to find complex repeats. The parameters for this comparison have been highly optimized to detect even the most divergent repeat families. This minimizes the "noise level" in the sequences and enables sensitive downstream analysis. It also eliminates undesired hits from the results of the BLAST search.

#### 4. PFP Bundled with Paracel BLAST

PFP is bundled with Paracel BLAST, thus providing the user with the option of using PFP instead of the default NCBI BLAST filtering to clean query sequences. Paracel has integrated PFP into Paracel BLAST so that PFP can be automatically invoked as part of the Paracel BLAST search. This eliminates the need to first run PFP on the query file and then use the cleaned query file in the BLAST search. To invoke PFP as

part of a Paracel BLAST search, the user merely types the following on the pb command line:

```
pb ... --pfp=<configfile> ...
```

Where <configfile> is the name of the configuration or parameter file that specifies various parameters for PFP to use when performing its filtering. PFP includes several parameter files that have been configured for various organisms. These parameter files are set to find low-complexity regions using DUST for nucleotides and organism-specific repetitive elements using HASTE (Hash-Accelerated Search Tool). The following parameter files are included with PFP for masking high-throughput genomic data:

TABLE 1: Masking Files Included with PFP

Parameter File	Masked Organism	Trivial Name
pfp_aratha.prm	<i>Arabidopsis thaliana</i>	wild mustard plant
pfp_celeg.prm	<i>Caenorhabditis elegans</i>	nematode
pfp_danr.prm	<i>Danio rerio</i>	zebrafish
pfp_dros.prm	<i>Drosophila melanogaster</i>	fruit fly
pfp_human.prm	<i>Homo sapiens</i>	human
pfp_mouse.prm	<i>Mus musculus/domesticus</i>	mouse
pfp_rat.prm	<i>Rattus norvegicus</i>	rat

There is another set of parameter files for filtering EST data. These will mask possible vector contamination and exclude sequences with high similarity matches to mitochondria, rRNA or *E. coli*, in addition to masking repetitive elements. Users can also create their own parameter files to perform the desired filtering and masking.

For example, consider the pfp\_human.prm file, which is part of the parameter file collection that comes with PFP:

```
# PFP parameter file for human genomic data
-MaskChar      X
-Complement    Y
-Debris        /dev/null
-align
% DUST
-Alg           dust
-Threshold     22
-Action        mask
% REPEATS
-Alg           Haste
-Reference     reference/human.repeats
-Threshold     187
-Action        mask
-WordLen       8
-Matrix        matrix/dna.p9m12TT-30-5.mat
```

This parameter file instructs PFP to use the X character for masking contaminants and repeats. It tells PFP to perform two

cleaning stages. For the first stage, it instructs PFP to use the DUST algorithm to mask low-complexity regions. For the second stage, it instructs PFP to use the HASTE algorithm to mask repetitive regions which will be identified by searching against PFP's human.repeats database and using PFP's evolutionary matrix dna.p9m12TT-30-5.mat to calculate scores. For a comprehensive description of PFP parameters and configuration files and how to customize them, please refer to the *PFP User Manual*.<sup>8</sup>

## 5. Results and Discussion

As an example, we compared the results of a `blastn` search for a single query sequence from the Human Transcripts database from Baylor University against NCBI's Human Reference Sequence database (10,239 sequences and 24,300,774 bases). According to the NCBI web site, the NCBI Reference Sequence project (RefSeq) provides reference sequence standards for the naturally occurring molecules of the central dogma from chromosomes to mRNAs to proteins<sup>9</sup>. The query sequence that was chosen was 1153 bases long and is given below.

```
>gi|306454|gb|L19955|HUMARYTRAA Human aryl
sulfotransferase mRNA, complete cds
GGTAAGGGAACGGGCTGGCTCTGGCCCCTGACGCAGGAACATGG
AGCTGATCCAGGACACCTCCCGCCCGCCACTGGAGTACGTGAAGG
GGGTCCCGCTCATCAAGTACTTTGCAGAGGCACCTGGGGCCCCCTGC
AGAGCTTCCAGGCCCGGCCTGATGACCTGCTCATCAGCACCTACC
CCAAGTCCGGCACCACCTGGGTGAGCCAGATTCTGGACATGATCT
```

*bases 226 - 900 not shown for sake of brevity*

```
AGAATGAGCGCTTCGATGCGGACTATGCGGAGAAGATGGCAGGCT
GCAGCCTCAGCTTCCGCTCTGAGCTGTGAGAGGGGCTCCTGGGGT
CACTGCAGAGGGAGTGTGCGAATCAAACCTGACCAAGCGGCTCAA
GAATAAAATATGAATTGAGGGCCTGGGACGGTAGGTCATGCTGT
AATCCAGCAATTTGGAGGCTGAGGTGGGAGGATCATTTGAGCCC
AGGAGTTCGAGACCAACCTGGGCAACATAGTGAGATTCTGTTAAA
AAAATAAAATAAAATAAAACCAATTTT
```

This query has an Alu repeat region (Claverie and Makalowski<sup>10</sup>) at query locations 1010-1121 (the boldface region) followed by a low-complexity A-rich region at query locations 1122-1144 (the italicized region).

Using the default NCBI BLAST filtering as part of the Paracel BLAST search, we got a total of 494 hits with a total run time of 0.62 s. The default NCBI BLAST filtering masked only the low-complexity A-rich region but failed to mask the Alu repeat region. This resulted in 423 undesired hits (86% of the total hits) against the Alu repeat.

Using PFP with the `pfp_human.prm` parameter file described above, instead of the default NCBI BLAST filtering, as part of the Paracel BLAST search resulted in only 71 hits and a total run time of 0.55 sec (including the PFP time). In this case, PFP masked both the Alu repeat region and the A-rich low-complexity region, thus eliminating all undesired hits against both regions and resulting in a smaller overall run time since there were significantly fewer hits to process.

FIGURE 1: NCBI Filtering vs. PFP with Paracel BLAST

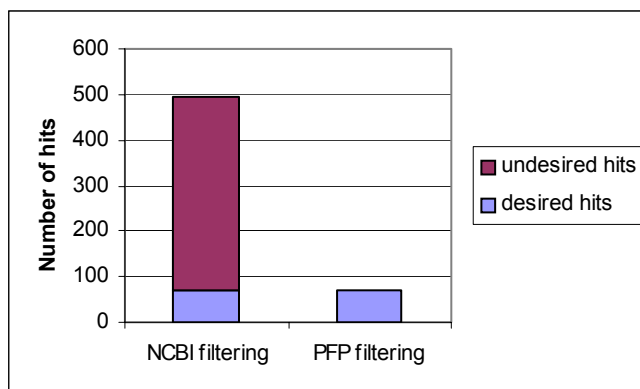


Table 2 presents the top 10 hits for the aforementioned search. Note that when the default NCBI filtering was used on the query, there were false-positive hits against the database sequences in the third through tenth rows. Those hits were against the Alu repeat region. When PFP was used to filter the query, there were no hits against these database sequences, since PFP masked the Alu repeat region in the query. Note also that the score for the first hit is higher when the default NCBI filtering was used than the corresponding score when PFP was used. This is because the Alu repeat region was included in the hit in the former case which increased the score, whereas the Alu repeat region was excluded from the hit in the latter case, which lowered the score.

TABLE 2: Paracel BLAST Score (bits)

Human Reference Database Sequences	NCBI Filtering	PFM Filtering
NM_001055	2222	2000
NM_001054	1681	1681
NM_018173	96	no hits
NM_020119	94	no hits
NM_017528	94	no hits
NM_003033	92	no hits
NM_001039	92	no hits
NM_004482	90	no hits
NM_017720	90	no hits
NM_015478	84	no hits

Furthermore, for very large query sequences with too many contaminants or repetitive regions, such as chromosomes, the BLAST search would usually fail as a result of running out of memory. Using the default NCBI BLAST filtering is usually not sufficient to solve this problem since it only masks low-complexity regions. However, using PFM to mask repetitive regions as well as low-complexity regions often allows the search to be completed. One such example is the *blastn* search for human chromosome 8 against the Human Reference Sequences database. The *blastn* search failed when the default NCBI BLAST filtering was used. However, when PFM was used to mask repetitive and low-complexity regions, the Paracel BLAST search was successfully completed.

## 6. Conclusion

It is evident from the previous examples that using PFM, instead of the default NCBI BLAST filtering, offers significant improvements to the BLAST search. PFM eliminates contaminants and repetitive regions from query sequences. This results in eliminating undesired hits against these contaminants and repetitive regions from the results of the BLAST search. It also allows searches to succeed that would have otherwise failed if repetitive regions were not removed. For cases where there are too many genome-wide repeats in the query and database, using PFM as part of Paracel BLAST even further enhances the significant speed gains as compared to NCBI BLAST searches. Furthermore, the integration of PFM into Paracel BLAST condenses the two-step process of first using PFM to clean the query and then using the cleaned query file in the BLAST search into a one-step process in which PFM will automatically be used to clean the query as part of the overall Paracel BLAST search. The result is an easy-to-use, one-step solution.

### Acknowledgments

We would like to thank Jon Murray for creating and running many of the benchmarks listed, and for numerous interesting discussions. For information about Paracel products and publications, contact Paracel, 1055 E. Colorado Blvd., Pasadena, CA 91106, USA. (626)-744-2000, <http://www.paracel.com>.

- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Meyers, and David J. Lipman, "Basic Local Alignment Search Tool", *J. Mol. Biol.* 215, 403-410 (1990).
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research*, 25(17): 3389-3402 (1997).
- Revision 1.6, April 2004.
- Paracel BLAST User Manual*, Paracel, Inc.
- Tatusov and Lipman, unpublished manuscript.
- Wootton J.C., and S. Federhen, "Analysis of compositionally biased regions in sequence databases", *Methods Enzymol* 266, 554-571 (1996).
- Distributed by the Genetic Information Research Institute (<http://www.girinst.org>).
- Paracel Filtering Package User Manual*, Paracel, Inc.
- For more information about the Human Reference Sequence project, please refer to the NCBI Internet web site at <http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html>.
- Claverie and Makalowski, "Alu Alert", *Nature* 371, 752 (1994).