BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 3 2008, pages 319–324
doi:10.1093/bioinformatics/btm585

Sequence analysis

Choosing BLAST options for better detection of orthologs as reciprocal best hits

Gabriel Moreno-Hagelsieb* and Kristen Latimer

Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada, N2L 3C5

Received on August 29, 2007; revised on October 21, 2007; accepted on November 19, 2007

Advance Access publication November 26, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The analyses of the increasing number of genome sequences requires shortcuts for the detection of orthologs, such as Reciprocal Best Hits (RBH), where orthologs are assumed if two genes each in a different genome find each other as the best hit in the other genome. Two BLAST options seem to affect alignment scores the most, and thus the choice of a best hit: the filtering of low information sequence segments and the algorithm used to produce the final alignment. Thus, we decided to test whether such options would help better detect orthologs.

Results: Using *Escherichia coli* K12 as an example, we compared the number and quality of orthologs detected as RBH. We tested four different conditions derived from two options: filtering of low-information segments, hard (default) versus soft; and alignment algorithm, default (based on matching words) versus Smith–Waterman. All options resulted in significant differences in the number of orthologs detected, with the highest numbers obtained with the combination of soft filtering with Smith–Waterman alignments. We compared these results with those of Reciprocal Shortest Distances (RSD), supposed to be superior to RBH because it uses an evolutionary measure of distance, rather than BLAST statistics, to rank homologs and thus detect orthologs. RSD barely increased the number of orthologs detected over those found with RBH. Error estimates, based on analyses of conservation of gene order, found small differences in the quality of orthologs detected using RBH. However, RSD showed the highest error rates. Thus, RSD have no advantages over RBH.

Availability: Orthologs detected as Reciprocal Best Hits using soft masking and Smith–Waterman alignments can be downloaded from http://popolvuh.wlu.ca/Orthologs.

Contact: gmoreno@wlu.ca

1 INTRODUCTION

The main purpose of this work is to evaluate different option sets to run BLASTP (Altschul *et al.*, 1997) that might help improve the proper detection of orthologs as Reciprocal Best Hits (RBH—for definitions read introductory paragraphs below). The main justification being that, despite important efforts at producing and making available ortholog mappings by various means (Alexeyenko *et al.*, 2006; Deluca *et al.*, 2006; Fulton *et al.*, 2006; Tatusov *et al.*, 2003; von Mering *et al.*, 2007), computational biologists often need their own orthologous sets for variable reasons such as: (1) a newly sequenced genome that needs annotation; (2) a need for updated mappings not available in published ortholog databases; (3) the lack of agreement about the genome annotations to use, for instance, those provided by the authors of a genome, corrections such as those within the RefSeq database (Maglott *et al.*, 2000; Pruitt *et al.*, 2005), the Genome Reviews (http://www.ebi.ac.uk/GenomeReviews/), the HAMAP project (Boeckmann *et al.*, 2003; Gattiker *et al.*, 2003) or even those re-annotations produced by particular research groups (Besemer *et al.*, 2001).

Orthologs are defined as genes that have diverged after a speciation event (Fitch, 2000). Another way to define them might be as the 'same genes' in different organisms. This evolutionary relationship implies that products of orthologous genes should tend to keep their original functions. Paralogs, on the other hand, are defined as genes that have diverged after a duplication event (Fitch, 2000). These have been proposed as a source of functional innovation (Francino, 2005; Ohno, 1970) and are less expected to have similar functions. It is therefore very important to be able to differentiate between orthologs and extra-paralogs, paralogous genes residing in different organisms (Janga and Moreno-Hagelsieb, 2004).

The definitions above are based on the event separating the histories of the homologous genes in question. In practice, one has to rely on sequence similarity and suitable statistics to detect homologs. Once putative homologs have been detected, evolutionary models such as phylogenetic trees, would be too computationally intensive to run for orthology detection, especially given the growth rate of genome sequence databases. Thus, most research in comparative genomics relies on some sort of shortcut, or working definition, to detect orthology. The most common working definition of orthology seems to be RBH (Bork *et al.*, 1998; Tatusov *et al.*, 1997), whereby two genes residing in two different genomes are deemed orthologs if their protein products find each other as the best hit in the opposite genome.

The task of finding homologs to a sequence of interest (the *query*) in a database containing many other sequences (the *targets*) can be conceptualized as getting the best possible alignment of the query against all the targets, scoring each of these alignments, and choosing those whose scores surpass a given threshold, or that comply with some alignment statistic (such as a maximum *E*-value). The process can be so time

*To whom correspondence should be addressed.

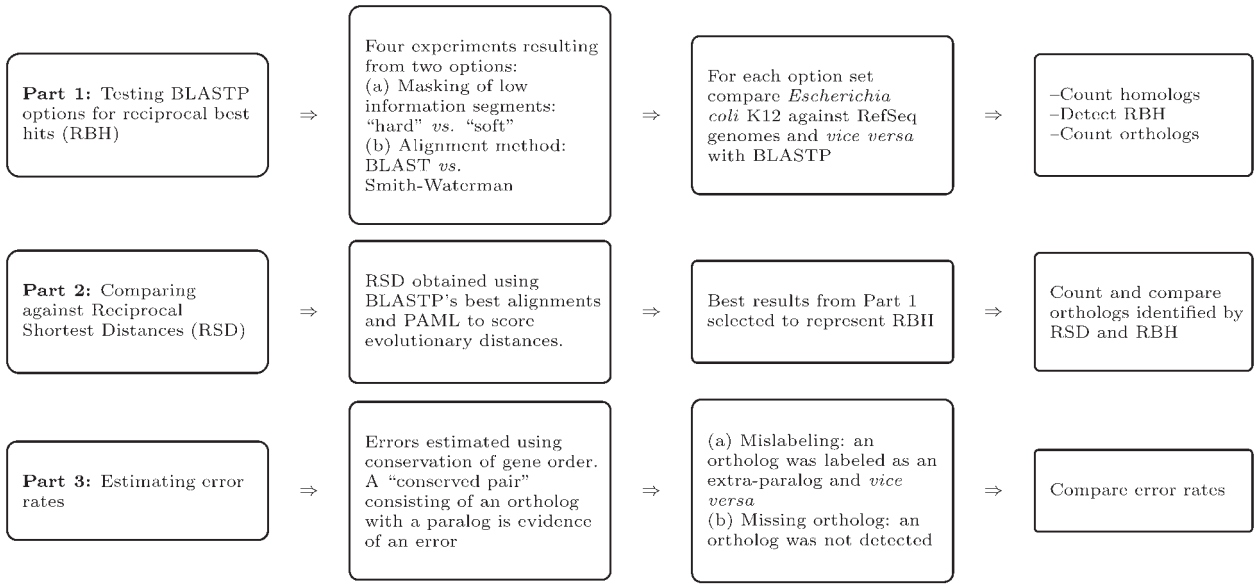

Fig. 1. Overview of the work.

consuming that researchers developed heuristic algorithms where the process is divided in two steps: the database search phase, where a shortcut is used to quickly choose those targets most likely to produce a meaningful alignment and the alignment phase, where those most promising targets are aligned to the query and scored (Pertsemlidis and Fondon 2001).

The Basic Local Alignment Search Tool (BLAST) (Altschul *et al*., 1990, 1997) is probably the most common heuristic algorithm used to find homologs. During the database phase, BLAST decomposes the query into small *words* and compares them to words in the database to find the most promising target sequences for alignment (Altschul *et al*., 1990). The final scores produced by BLAST might mainly be affected by two options: Masking of low-information segments and the method to produce the final alignments. Low information segments of sequence are commonly masked (marked to be ignored), because these segments might initiate too many spurious alignments of a query with unrelated targets. The default option is to use a *hard* filter, that applies the masking technique at both the database phase and the alignment phase. *Soft* filtering (chosen in the standalone NCBI version of BLAST with the *-F "m S"* option) masks low information segments only during the search phase, thus taking advantage of the time-saving purpose of masking, while also allowing for a more proper scoring of the aligned sequences (Schaffer *et al*., 2001).

The default alignment produced by BLAST is mainly based on overlapping and extending little matching 'words' (the ones used during the search phase) between the query and the target. BLASTP, the BLAST program that aligns protein sequences, also offers the choice of using a Smith–Waterman algorithm to produce the final alignments and scores (with the *-s T* option) (Schaffer *et al*., 2001). The Smith–Waterman algorithm aligns sequences using dynamic programming (Smith and Waterman, 1981) to produce the mathematically optimal local alignment (Brenner *et al*., 1998; Eddy, 2004).

Table 1. BLASTP options tested

Options	BLAST alignment	Smith–Waterman alignment
Hard Filter	*-F T -s F*	*-F T -s T*
Soft Filter	*-F "m S" -s F*	*-F "m S" -s T*

BLASTP alignments are based on small matching words, while the Smith–Waterman alignments are based on dynamic programing. A hard filter would mask low-information sequences during both the database search phase, and the alignment phase, while soft filtering masks low-information sequences only during the database search phase.

Here we used the genome of *Escherichia coli* K12 (Blattner *et al*., 1997) as a query genome to compare the differences in the number of orthologs found as RBH with BLASTP run with different options. We tested four option sets resulting from choosing hard versus soft filtering of low information segments; and default versus Smith–Waterman alignments (Table 1). We also compared these results against those obtained with another working definition of orthology, named Reciprocal Shortest Distances (RSD) (Wall *et al*., 2003), that uses evolutionary distances to rank the protein homologs found by BLASTP (Fig. 1).

2 DATA AND METHODS

We used the 587 prokaryotic genomes available at the RefSeq database (Maglott *et al*., 2000; Pruitt *et al*., 2005) (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) by October 2007. We calculated Genomic Similarity Scores (GSS) as described previously (Moreno-Hagelsieb and Collado-Vides, 2002; Janga and Moreno-Hagelsieb, 2004).

We ran NCBI's BLASTP comparisons of all the proteins encoded by the annotated genes of *E.coli* K12 against all the proteins encoded by the genes annotated in any other genome, and vice versa, with a maximum *E*-value threshold of 1×10^{-6}; a database size fixed at 5×10^{8} (*-z 5e8*), and any of the four option sets displayed in Table 1. For any

Table 2. Statistical analyses of detected orthologs (paired *t*-tests)

Option set	RSD	*-F "m S" -s T*	*-F "m S" -s F*	*-F T -s T*	*-F T -s F*
Number of genes finding orthologs					
RSD	–	2.64 (8.5×10^{-03})	5.08 (5.0×10^{-07})	23.80 ($<2.2 \times 10^{-16}$)	26.36 ($<2.2 \times 10^{-16}$)
-F "m S" -s T	5.75 (1.4×10^{-08})	–	27.28 ($<2.2 \times 10^{-16}$)	55.83 ($<2.2 \times 10^{-16}$)	58.82 ($<2.2 \times 10^{-16}$)
-F "m S" -s F	3.74 (2.0×10^{-04})	−11.44 ($<2.2 \times 10^{-16}$)	–	51.45 ($<2.2 \times 10^{-16}$)	56.41 ($<2.2 \times 10^{-16}$)
-F T -s T	−12.60 ($<2.2 \times 10^{-16}$)	−34.87 ($<2.2 \times 10^{-16}$)	−31.51 ($<2.2 \times 10^{-16}$)	–	29.47 ($<2.2 \times 10^{-16}$)
-F T -s F	−15.20 ($<2.2 \times 10^{-16}$)	−36.97 ($<2.2 \times 10^{-16}$)	−36.26 ($<2.2 \times 10^{-16}$)	−16.32 ($<2.2 \times 10^{-16}$)	–
Number of unique orthologous pairs					
RSD	–	−0.24 (0.81)	1.43 (0.15)	15.27 ($<2.2 \times 10^{-16}$)	17.23 ($<2.2 \times 10^{-16}$)
-F "m S" -s T	0.27 (0.79)	–	26.44 ($<2.2 \times 10^{-16}$)	51.50 ($<2.2 \times 10^{-16}$)	54.95 ($<2.2 \times 10^{-16}$)
-F "m S" -s F	−8.85 ($<2.2 \times 10^{-16}$)	−20.11 ($<2.2 \times 10^{-16}$)	–	47.03 ($<2.2 \times 10^{-16}$)	52.17 ($<2.2 \times 10^{-16}$)
-F T -s T	−45.62 ($<2.2 \times 10^{-16}$)	−57.38 ($<2.2 \times 10^{-16}$)	−57.65 ($<2.2 \times 10^{-16}$)	–	28.90 ($<2.2 \times 10^{-16}$)
-F T -s F	−44.38 ($<2.2 \times 10^{-16}$)	−52.35 ($<2.2 \times 10^{-16}$)	−57.63 ($<2.2 \times 10^{-16}$)	−20.71 ($<2.2 \times 10^{-16}$)	–

The results produced with all BLASTP options sets were significantly different to each other. RSD increased the number of genes finding orthologs over those detected with RBH with the *-F "m S" -s T* option set. However, the increase was even smaller than that between the *-F "m S" -s T* and *-F "m S" -s F* options sets. The triangles to the right show comparisons of the overall numbers of genes finding orthologs (or unique orthologous pairs) (*t* and probability), while the triangles to the left show results normalized against the corresponding numbers of homologs (or homologous pairs). The statistics were calculated in the order shown (RSD to *-F T -s F*). While the numbers of orthologs decreased in that order, with RSD and *-F "m S" -s T* producing the highest numbers, the ratios of orthologs per homolog increased.

further consideration, we also required coverage of at least 50% of any of the protein sequences in the alignments.

To find orthologs as RBH we sorted the BLASTP hits from highest to lowest bit-score, then, if the bit-scores were identical, from smallest to highest *E*-values (sorting by either bit-score first, or by *E*-value first, works out very close to the same end results). The first hit within these sortings would be the best hit. If the next hit had the very same bit-scores and *E*-values, there would be more than one best hit (multiple orthologs can occur).

To find orthologs by the RSD definition we used the alignments obtained from the BLASTP run with the *-F "m S" -s T* options (no masking of low-information sequences during the alignment phase, with Smith–Waterman alignment). Wall *et al.* (2003) used CLUSTAL W (Thompson *et al.*, 1994) to re-align the homologous sequences found by BLASTP before calculating RSD. However, the global alignments obtained using CLUSTALW would miss fused and re-arranged genes. The Smith–Waterman algorithm guarantees a mathematically optimal local alignment (Smith and Waterman, 1981) making re-alignments unnecessary (see further arguments below). We used PAML (Yang, 1997, 2007) to calculate evolutionary distances as maximum likelihood of amino acid substitutions as described by Wall *et al.* (2003).

3 ORTHOLOG DETECTION INCREASES WITH *SOFT FILTERING* AND *SMITH–WATERMAN* ALIGNMENTS

The number of genes finding orthologs as RBH increased significantly with any of the tested options departing from the default ones (hard filter with default alignment algorithm; Table 2). The option sets can be ordered from the one giving the highest to the one giving the lowest numbers of orthologs as follows: *-F "m S" -s T*>*-F "m S" -s F*>*-F T -s T*>*-F T -s F* (Fig. 2a). Thus, the highest increases occurred when selecting for both soft filtering (*-F "m S"*) and Smith–Waterman alignments (*-s T*), with soft filtering accounting for most of the difference (Smith–Waterman alignments added about one tenth of the number of orthologs added by soft filtering alone).

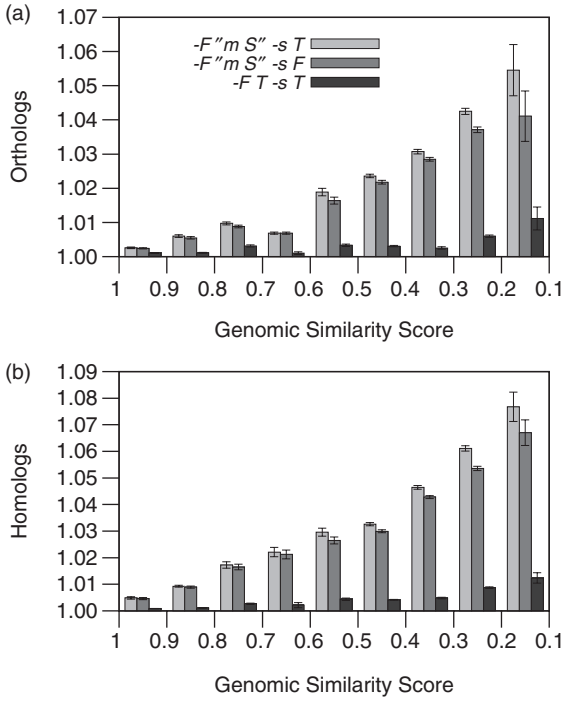

Fig. 2. Differences in the number of genes finding orthologs as RBH. The numbers were normalized against the corresponding numbers of genes finding orthologs/homologs in the default options set (*-F T -s F*). The maximum increases in the number of orthologs and homologs were found with the *-F "m S" -s T* options set. However, most of the increase was attained with the soft masking (*-F "m S"*) option. (**a**) The increments tend to be higher in evolutionarily distant genomes (low Genomic Similarity Score—GSS). (**b**) The increase in the number of genes finding homologs might explain the increased number of orthlogs. The bars represent averages for genomes separated from *E.coli* K12 at intervals of 0.1 GSS. The number of genomes at each GSS interval is not the same.

This is an important result, given that the option that slowed down the BLASTP runs the most was the Smith–Watermann alignment, while the extra time required to run BLASTP with soft filtering was negligible. The maximum increase in number of orthologs from the default options was of 1.13 times as many as those obtained with the default option set, with an average increase of 1.03. While the average is rather low, using these options would be more valuable for the detection of orthologs between evolutionarily distant organisms.

A total of 1675 genes displayed different orthology results, in at least one of the genomes compared, using the *-F "m S" -s T* options instead of the *-F T -s F* options. The proteins encoded by these genes tended to be slightly longer than other proteins finding orthologs ($t = 11.16$, $p < 2.2e-16$). As it would be expected, 89% of these 1675 proteins contained low-information segments that get masked by the *-F T* option.

Since the number of homologs detected might also be affected by the choice of options we also accounted for the number of genes finding homologs (Fig. 2b). These showed the same tendencies as the number of orthologs above suggesting that the increase in the number of orthologs might be more related to a corresponding increase in the number of homologs to start with, than to better scoring for deciding on RBH. If we normalize the numbers of orthologs by the corresponding numbers of homologs, we find that the differences among the normalized numbers of orthologs are still significant (Table 2). However, the tendency changes direction with the *-F T -s F* option set producing the highest numbers of orthologs per homolog detected and the *-F "m S" -s T* option set producing the lowest numbers.

4 RECIPROCAL SHORTEST DISTANCES BARELY DETECT MORE ORTHOLOGS THAN RECIPROCAL BEST HITS

Since we wanted to have a control, a ceiling, of the expected number of orthologs, we tested the RSD definition of orthology, because it was reported to produce a much higher number of orthologs than RBHs (Wall *et al*., 2003). As reported, the RSD differs from RBH in two ways (Deluca *et al*., 2006; Wall *et al*., 2003): (1) use of a measure of evolutionary distance, namely maximum likelihood of amino acid changes, instead of BLASTP scores or statistics; (2) use of global alignments, rather than local alignments, before calculating these distances. As a first approach, we only used evolutionary distances calculated with local alignments produced with the Smith–Waterman option without masking of low-information regions (*-F "m S" -s T*). This setup ensured that the only difference between RBH and RSD was the use of evolutionary distances. We made sure to run RSD and RBH with the very same homologs detected with BLASTP. However, RSD did not appreciably increase the number of genes finding orthologs compared with those found with the *-F "m S" -s T* RBH (Table 2). We further tested whether the RSD would be successful at finding more orthologs than RBH if the RBH compared were those obtained with default options (*-F T -s F*). In this case, we calculated amino acid change distances using the alignments produced with the *-F "m S" -s T* options for homologs detected with the *-F T -s F* options, and then found RSD orthologs. This test failed to give any advantage to RSD (Fig. 3). It is plausible then, that the results presented in the RSD article (Wall *et al*., 2003), were either more related to the use of global alignments than to the use of a more evolutionarily meaningful scoring system, limited to the model organisms used in that study (*Saccharomyces cerevisiae*), or the RBH were not setup properly (the method to detect RBH was not described with enough detail in Wall *et al*., 2003).

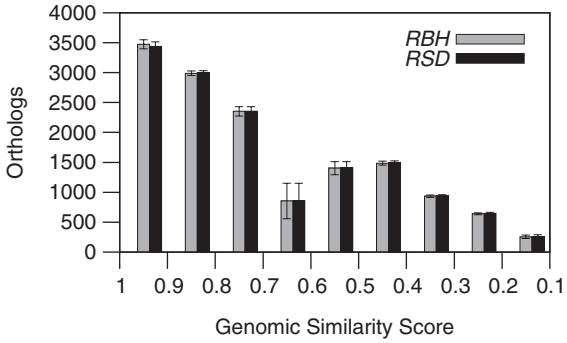

Fig. 3. Differences in the number of genes finding orthologs as RBH and by RSD. Even if RSD are calculated only for homologs detected with default BLASTP options (*-F T -s F*) the number of genes detecting orthologs by RSD is not that different to the number detected as RBH. The bars represent averages for genomes separated from *E.coli* K12 at intervals of 0.1 GSS. The number of genomes at each GSS interval is not the same.

We decided not to further pursue this problem for the following reasons: Our review of the literature did not reveal a convincing reason why evolutionary comparisons should use global alignments. Discussions with colleagues yielded mostly textbook statements were the use of global alignments for evolutionary studies is advised, but neither justified, nor explained. The most convincing argument we could come up with was that perhaps the local alignment algorithms drop the alignments too soon, thus ignoring parts of the sequences that the global alignments will try and align anyway, and that those missing pieces of the alignment might provide better, or more meaningful, scoring by revealing evolutionary events missed by the local alignments. However, global alignments would require a high sequence coverage within the alignments, such as the requirement for the alignable region of the two sequences to exceed 80% of the global alignments total length used by the authors of RSD (Wall *et al*., 2003). Besides such requirement would preclude the finding of gene fusions, testing global alignments and finding a fair corresponding setup for RBH (for instance: What would be the equivalent to the 80% of the alignable regions above? Or, since BLAST can produce more than one high-scoring alignment for the same query-target pair of sequences—so called High-scoring Segment Pairs or HSPs, what would happen if we compound the BLASTP statistics for more than the first HSP?), would require much more work, might constitute another research report by itself, and thus escapes the scope of this article. Finally, the only improvement provided in the RSD article was an increased number of orthologs, with no evidence about their quality. As we stand, RSD barely detected more orthologs than RBH.

5 THE ERROR RATES ARE HIGHEST WITH RSD

The number of orthologs detected does not constitute a test of quality. Thus, we also estimated error rates. To estimate error rates, we analyzed conservation of adjacency of homologous genes. Conservation of gene order has been previously suggested to be of limited use for the assignment of orthology in Prokaryotes due to the high divergence of gene order prevailing in these organisms (Bork *et al*., 1998). However, conservation, even if low, should be useful to estimate error rates assuming that conserved pairs of target genes are both orthologs to their corresponding query genes. Evidence of mistakes came in the form of conserved homologous pairs where one of the genes was found to be an ortholog, while the other was found to be an extra-paralog. If the query gene finding an extra-paralog found an ortholog elsewhere, this mistake was deemed a mislabeling mistake, whereby the true ortholog was labeled an extra-paralog, while the true extra-paralog was labeled as an ortholog. If no orthlog was found elsewhere the error was deemed a missing ortholog, whereby the algorithm did not find an existing ortholog for the query gene. Error rates (*ER*) were calculated as:

$$ER = \frac{N_P}{N_P + N_O}$$

where N_P is the number of errors (number of paralogs found next to an ortholog) and N_O is the number of correct orthologs (orthologs found next to corresponding orthologs, or the ortholog found next to a paralog).

The error rate increased with the evolutionary distance as measured using GSS (Fig. 4). These error rates were approximately the same among all the RBH orthologs (Table 3). However, the RSD orthologs had the highest rates of both kinds of errors, especially of those related to failure to detect an ortholog (Table 3). Note that reasons for these errors might be related to biologically meaningful events, such as gene conversion (recombination between homologous genes leading to divergences and convergences in sequence related to events other than single point mutations), or the genes have diverged so much that their status as orthologs or extra-paralogs are barely discernible. However, the comparisons still hold, since the background biological events are the very same, yet we found different error rates.

The main caveat of this analysis is that a method detecting every homologs in other genomes as orthologs will have an error rate of zero. There is no reason why a gene in one organisms should have only one ortholog in another. Such a situation would happen if a duplication event occurs after a speciation event (Fitch, 2000). However, considering every homolog an ortholog would defy the purpose of finding the most probable functional equivalents. Thus, there is a need for further work to discern such possibilities (more than one proper ortholog), at the same time that we obtain those genes most likely to share their overall functions.

6 CONCLUDING REMARKS

The results presented show that the number of orthologs detected varies according to the choices made when searching

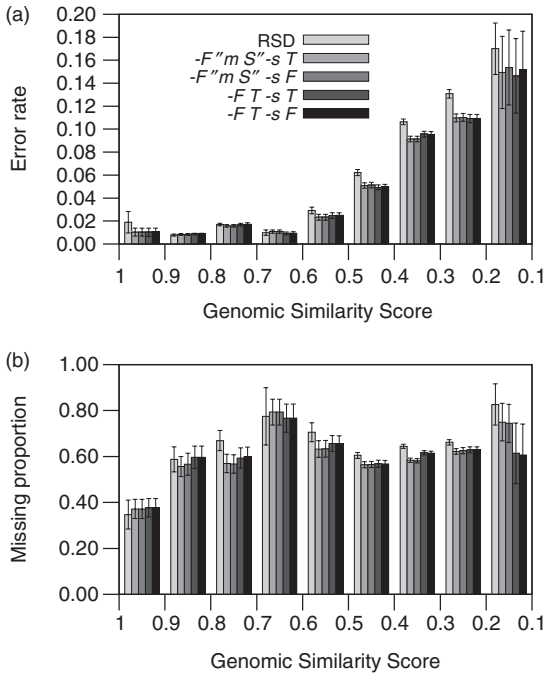

Fig. 4. Error rate estimates. These estimates were calculated using conservation of gene order. The estimate consists on erroneous genes (an extra-paralog found next to an ortholog) divided by the sum of erroneous + correct genes (correct genes being those where the conserved pair is an ortholog adjacent to an ortholog, or the ortholog found next to the extra-paralog). (**a**) Orthologs detected with RSD showed the highest error rates. (**b**) A higher proportion of the errors were missing orthologs, except at high GSS where the prevalent error was mislabeling (an ortholog for an extra-paralog, and vice versa). The bars represent averages for genomes separated from *E.coli* K12 at intervals of 0.1 GSS. The number of genomes at each GSS interval is not the same.

for homologs in the genome databases. Both options tested, soft filtering and Smith–Waterman final alignment, proved important for finding homologs and for detecting orthologs as RBH. Surprisingly, orthologs detected with RSD contained higher error rates than RBH. RSD might have been instinctively perceived as a better definition of orthology because they are based on an evolutionary model, compounded perhaps with a report showing that the best BLAST hits do not always correspond to the closest evolutionary neighbors (Koski and Golding, 2001). However, the comparison of BLAST hits with evolutionary neighbors was based on a default BLAST run, with evolutionary neighbors detected using phylogenetic trees (Koski and Golding, 2001), not just distances calculated from amino acid substitutions (as in Wall *et al*., 2003). We must also bear in mind that, despite BLAST was not designed to produce evolutionary distances, but rather to quickly find similar sequences, the scoring matrices used by the program have an underline of evolution. Thus, evolutionary distances have to be reflected, even if not perfectly, by these scores. The most important lesson from this part of the study might be that we should not just assume that an intuitively perceived better strategy (such as evolutionary distances versus BLAST statistics and scores), is really better unless properly and fairly tested.

Table 3. Statistical analyses of error estimates (paired *t*-tests)

Option set	RSD	*-F "m S" -s T*	*-F "m S" -s F*	*-F T-s T*
RSD	–	–	–	–
-F "m S" -s T	15.72 ($<2.2 \times 10^{-16}$)	–	–	–
-F "m S" -s F	15.13 ($<2.2 \times 10^{-16}$)	−1.70 (0.09)	–	–
-F T -s T	13.57 ($<2.2 \times 10^{-16}$)	−2.07 (0.04)	−1.64 (0.10)	–
-F T -s F	13.41 ($<2.2 \times 10^{-16}$)	−2.21 (0.03)	−1.83 (0.07)	−0.57 (0.57)

The statistics were calculated in the order shown (RSD to *-F T -s F*). While other comparisons. might be considered significant, with the *-F "m S" -s T* options set making the lowest number of errors, RSD had the highest error rates.

Using *E.coli* K12 as a model allowed us to present results at many levels of evolutionary divergence (measured with Genomic Similarity Scores). The biases in the genome data would not allow for such a clean presentation using most other organisms. An organism with no close relatives would yield higher average differences in the numbers of orthologs because the effect of the BLASTP options tested increases against the GSS. We speculate that eukaryotic organisms might also yield higher differences because the longer proteins of these organisms tend to contain a higher proportion of low-information segments than prokaryotic proteins. However, a more comprehensive study might be needed to confirm this hypothesis.

Based on our results, the recommended parameters for the best detection of orthologs as reciprocal best hits is the combination of soft filtering with a Smith–Waterman final alignment (the *-F "m S" -s T* options in NCBI's BLASTP). These options resulted in both the highest number of orthologs and the minimal error rates. However, most of the improvement can be achieved using soft filtering (*-F "m S"*) alone.

ACKNOWLEDGEMENTS

G.M.H. acknowledges the Shared Hierarchical Academic Research Computing Network (SHARCNET) for computer cluster usage. This work was supported with funds from WLU and a discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Conflict of Interest: none declared

REFERENCES

Alexeyenko,A. *et al.* (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, **22**, e9–e15.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Besemer,J. *et al.* (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.

Blattner,F.R. *et al.* (1997) The complete genome sequence of*Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bork,P. *et al.* (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.

Brenner,S.E. *et al.* (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, **95**, 6073–6078.

Deluca,T.F. *et al.* (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22**, 2044–2046.

Eddy,S.R. (2004) What is dynamic programming? *Nat. biotechnol.*, **22**, 909–910.

Fitch,W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.

Francino,M.P. (2005) An adaptive radiation model for the origin of new gene functions. *Nat. Genet.*, **37**, 573–577.

Fulton,D.L. *et al.* (2006) Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, **7**, 270.

Gattiker,A. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.

Janga,S.C. and Moreno-Hagelsieb,G. (2004) Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Res.*, **32**, 5392–5397.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

Maglott,D.R. *et al.* (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128.

Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18** (Suppl. 1), S329–S336.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Pertsemlidis,A. and Fondon,J.W.3rd (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol.*, **2** REVIEWS2002.

Pruitt,K.D. *et al.* (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33** (Database Issue), D501–D504.

Schaffer,A.A. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov,R.L. *et al.* (2003) The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

von Mering,C. *et al.* (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35** (Database issue), D358–D362.

Wall,D.P. *et al.* (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

Yang,Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.