BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2141–2143
doi:10.1093/bioinformatics/btr342

Genome analysis

Advance Access publication on June 8, 2011

genBlastG: using BLAST searches to build homologous gene models

Rong She[1,†], Jeffrey Shih-Chieh Chu[2,†], Bora Uyar[2,3], Jun Wang[2], Ke Wang[1] and Nansheng Chen[1,2,3,*]

[1]School of Computing Science, [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada V5A 0A1 and [3]CIHR/MSFHR Strategic Training Program in Bioinformatics at the University of British Columbia, Simon Fraser University, and British Columbia Cancer Agency, Canada

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: BLAST users frequently expect to obtain homologous genes with certain similarity to their query genes. But what they get from BLAST searches are often collections of local alignments called high-scoring segment pairs (HSPs). On the other hand, most homology-based gene finders have been built using computation-intensive algorithms, without taking full advantage of BLAST searches that have been perfected over the last decades.

Results: Here we report an efficient algorithm, genBlastG that directly uses the HSPs reported by BLAST to define high-quality gene models.

Availability: http://genome.sfu.ca/genblast/download.html

Contact: chenn@sfu.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 22, 2010; revised on May 24, 2011; accepted on May 29, 2011

1 INTRODUCTION

Basic Local Alignment Search Tool (BLAST; Altschul *et al*., 1990) is one of the most popular and efficient bioinformatics tools ever developed. Frequently, BLAST users expect to identify homologous genes for comparative analysis. For example, following the discovery of a previously unknown gene in a human genome, a biologist will typically perform a BLAST search of the publicly accessible databases such as the mouse genome database to see if another species carries a similar gene, hoping to gain insights into the function and regulatory signals of the newly found gene. BLAST, however, presents the user a (usually large) collection of local alignments called high-scoring segment pairs (HSPs). Such HSPs provide no indication how they are structured into the gene model because they are usually only isolated regions of similarity with some being simply noises.

In the last decade, many homology-based gene predictors have been developed including GeneWise (Birney *et al*., 2004), Projector (Meyer and Durbin, 2004), TwinScan (Korf *et al*., 2001) and Exonerate (Slater and Birney, 2005). These algorithms either are independent of BLAST or use BLAST only as a preprocessing tool to narrow down gene search space (Cui *et al*., 2007). Our work is motivated by the following question: can we construct gene structures directly from the HSPs found by BLAST, which represent high-quality local alignments? The rationale is to delegate the expensive local alignment search to the well developed BLAST and focus on extracting and defining the best gene structure that such HSPs represent. Here we present a novel homology-based algorithm, genBlastG, which takes local similarity alignments (HSPs) identified by homologous searches as input and defines gene models by examining alignments and neighboring genomic regions for start/stop codons and splicing signals. Compared to previous homology-based gene finding algorithms, genBlastG is able to leverage the vast improvement in speed and search quality of BLAST made in last 20 years since its first publication and benefit from the wide acceptance and availability of the program. This hypothesis has been evaluated in nematode, plant and human genomes. Our tests show that genBlastG is extremely fast while providing better performance than previous algorithms in terms of specificity and sensitivity (Burset and Guigo, 1996).

2 DESCRIPTION

Each gene is composed of one or more exons separated by introns, which are flanked by splicing signals (Breathnach and Chambon, 1981) and has a start codon and a stop codon. In a recent project, we have developed a program genBlastA to define homologous genomic regions based on HSPs retrieved by BLAST (She *et al*., 2009). These homologous genomic regions potentially contain candidate genes, but they do not manifest the exact gene structures. Here, we take advantage of the homologous genomic regions returned by genBlastA to further define homologous gene models. Given a query gene (e.g. a protein), genBlastG finds the gene models in a target genome in two steps: (1) parsing the output of BLAST searches into groups of HSPs with each group representing a genomic region that contains a candidate homologous gene (She *et al*., 2009), and (2) examining the HSPs within each group and exploits sequence signals in the homologous genomic region to define introns and gene start and stop signals. The Step 1 is done by running genBlastA, while the Step 2 must address the following challenges.

First, there is no simple one-to-one correspondence between HSPs and exons: one HSP can correspond to multiple exons and multiple HSPs can correspond to one exon. Secondly, some exons may not be represented by any HSP at all, whereas some HSPs are simply

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

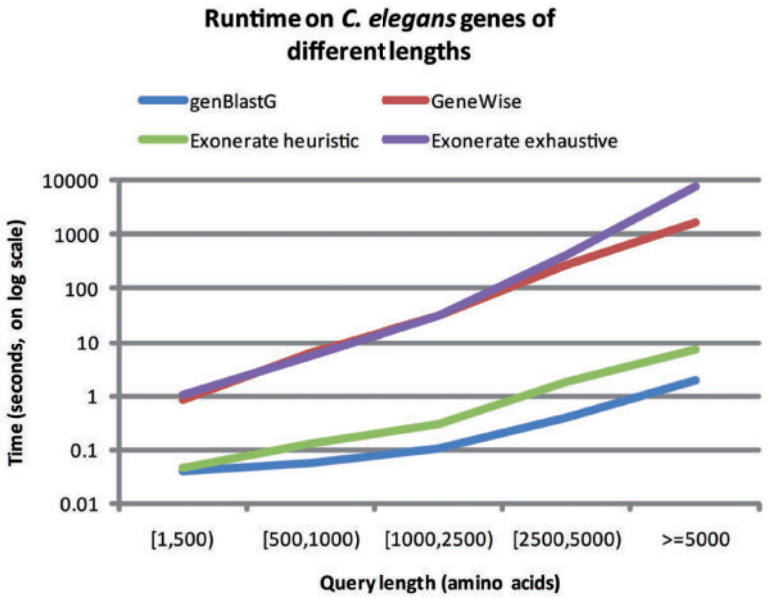

Fig. 1. Runtime comparison between genBlastG and GeneWise.

noises and do not correspond to any exon. Thirdly, even when HSPs correspond to an exon region, the structure of introns and exons has to be resolved because HSPs carry no information of such structures. Numerous possible combinations of splicing sites surrounding the HSPs could delimit introns and exons. Determining the best combination from these alternatives is not an arbitrary decision and is challenging. These issues are not considered in genBlastA.

genBlastG builds gene models by maximizing the similarity to their corresponding queries. First, putative intronic regions are detected by examining gaps between adjacent HSPs, as well as mismatches within HSPs. Secondly, candidate splice sites are detected. Consider an intron region *I*, which is associated with its own set of predicted splice donors (*d1, …, dn*) and acceptors (*a1, …, an*). For a pair of donor and acceptor to be considered a valid pairing, the donor and acceptor must be in-frame with each other and there is no in-frame stop codon in the corresponding spliced sequence S, which is formed by connecting the subject segment of the HSP at the upstream side of a donor site (called 'donor-side subject segment') with the subject segment of the HSP at the downstream side of an acceptor site (called 'acceptor-side subject segment'). The best donor–acceptor pair is the one that results in the highest alignment identity between the spliced sequence S and its corresponding query segment (at the amino acid level). It is possible that there exists no valid pair of donor and acceptor in an intron region, in which case no intron will be predicted and consequently the genomic region will be accepted as an exon.

After selecting the best pair of donor and acceptor for each intron region, an initial gene structure is obtained. However, gene models could still be incomplete for missing one or more exons because BLAST may fail to pick up weak and short similarities. genBlastG retrieves missing alignments to maximize the similarity between the predicted gene model and corresponding query gene. For each region that is subject to repair, a local alignment algorithm (Smith and Waterman, 1981) is used to find the possible missing alignment in the genomic region between two adjacent exons, or the upstream region before the first exon or the downstream region after the last exon.

3 RESULTS

We first tested the performance of genBlastG in predicting genes in the genome of the model organism *Caenorhabditis elegans* (C.elegans Sequencing Consortium, 1998) using *C.elegans* proteins as queries. For each gene prediction run for genBlastG, we used genBlastA-defined genomic region together all contained HSPs as input, while for GeneWise and exonerate, we used a protein as query and the corresponding homologous genomic region returned by genBlastA as the genomic search space. Thus for all three programs, we used genBlastA-defined genomic regions as gene search spaces. We found that genBlastG runs considerably faster than GeneWise, the arguably most widely used homology-based gene prediction program, and Exonerate (Slater and Birney, 2005) especially when it runs in the exhaustive mode (Fig. 1). genBlastG runs faster than Exonerate even when it runs at the less accurate heuristic mode.

This high speed of genBlastG is not achieved by sacrificing its performance. In fact, when genBlastG is used to remap *C.elegans* proteins to the *C.elegans* genomes, it outperforms both GeneWise and Exonerate in generating genes with high sensitivity and specificity at full-length transcript level (Table 1; Supplementary information). It performs similarly to GeneWise and Exonerate at the exon and nucleotide levels (Table 1) (Burset and Guigo, 1996). The better performance comes from the high quality of HSPs returned by BLAST and genBlastG's effort of maximizing similarity to the query gene in defining exons.

genBlastG also performs favorably in predicting genes in different including distantly related genomes. We evaluated its performance against GeneWise and exonerate in predicting genes in *Arabidopsis thaliana* genome using rice (*Oryza sativa* L. ssp. *japonica*) proteins that are fully supported by cDNAs as queries. These two species

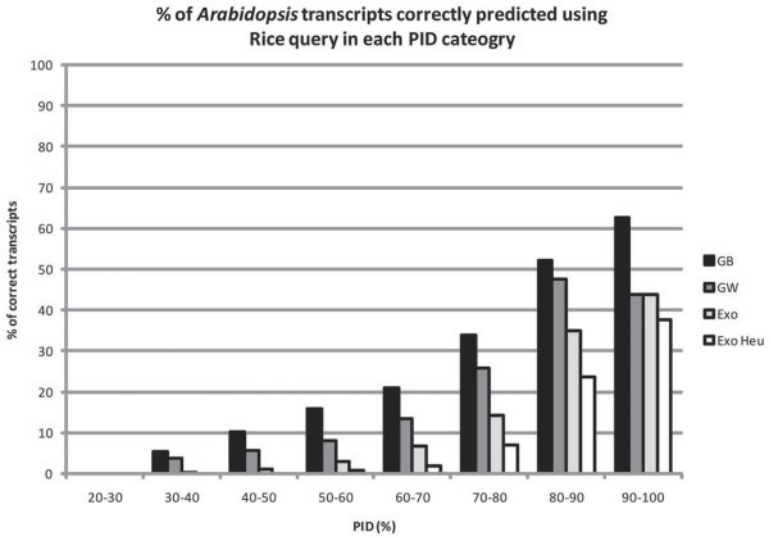

Fig. 2. Dependence of correctness of predicted genes on PID. GB, genBlastG; GW, GeneWise; Exo, exonerate; Exo Heu, heuristic exonerate.

Table 1. Remapping *C.elegans* proteome to the *C.elegans* genomes ($n = 6844$ genes)

	Transcript		Exon		Nucleotide	
	Sp. (%)	Sn. (%)	Sp. (%)	Sn. (%)	Sp. (%)	Sn. (%)
genBlastG	94.10	94.10	98.31	97.85	99.79	99.74
GeneWise	91.07	91.07	97.50	96.92	99.87	99.68
Exonerate[a]	93.73	93.73	98.15	97.77	99.85	99.83
Exonerate[b]	91.03	91.03	97.41	96.26	99.89	99.40

[a]Exhaustive.
[b]Heuristic.

diverged from their common ancestor more than 100 million years ago (Itoh *et al*., 2007). We divided query proteins in eight categories, each of which has a certain global percentage identity (PID) between query (rice) proteins and their orthologous proteins in *A.thaliana*. We then calculated the percentage of the predicted proteins in *A.thaliana* that are identical (thus, perfectly predicted) to the curated orthologs in *A.thaliana*. As shown in Figure 2, for PID of 90–100%, 62.5% of genBlastG predicted *A.thaliana* gene models are base pair to base pair identical to curated orthologs in *A.thaliana*. In contrast, the percentage of correct transcripts predicted by GeneWise is only 43.8%.

This study clearly shows that there is a high correlation between a correct prediction of transcripts and the similarity between query genes and their orthologs.

4 CONCLUSION

BLAST has been extremely successful as a tool for finding local alignments with both high sensitivity and speed. However, the large number of isolated local alignments, in the form of HSPs, cannot be readily and effectively interpreted due to the lack of obvious organization that describes the gene structure from which HSPs are extracted. genBlastG, which builds on the success of genBlastA, presents an approach that constructs the gene models directly from the HSPs returned by BLAST, with the intention of leveraging the wide success of BLAST. Our study shows that genBlastG can find gene models much faster than existing homology-based gene finders including GeneWise and Exonerate, while providing better performance than these programs.

Funding: Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (to K.W. and N.C.); SFU Community Trust funded BCID Project; NSERC Scholarship (to J.S.-C.C.). N.C. is a Michael Smith Foundation for Health Research (MSFHR) Scholar and a Canadian Institutes of Health Research (CIHR) New Investigator.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
Birney,E. *et al*. (2004) GeneWise and Genomewise. *Genome Res.*, **14**, 988–995.
Breathnach,R. and Chambon,P. (1981) Organization and expression of eucaryotic split genes coding for proteins. *Annu. Rev. Biochem.*, **50**, 349–383.
Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C.elegans: a platform for investigating biology. *Science*, **282**, 2012–2018.
Cui,X. *et al*. (2007) Homology search for genes. *Bioinformatics*, **23**, i97-i103.
Itoh,T. *et al*. (2007) Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. *Genome Res.*, **17**, 175–183.
Korf,I. *et al*. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17** (Suppl. 1), S140–S148.
Meyer,I.M. and Durbin,R. (2004) Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res.*, **32**, 776–783.
She,R. *et al*. (2009) GenBlastA: enabling BLAST to identify homologous gene sequences, *Genome Res.*, **19**, 143–149.
Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences, *J. Mol. Biol.*, **147**, 195–197.