BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 11(4)

BlastXtract2: Improving early exploration of (meta) genomes

Heleen de Weerd[1], Bernd E van der Veen[2] & Marcus J Claesson[3]*

[1]Microbiology and Systems Biology, TNO, Zeist, The Netherlands, [2]Genomics Core Facility, The Netherlands Cancer Institute, Amsterdam, The Netherlands, [3]School of microbiology and alimentary pharmabiotic centre, university college cork, cork, Ireland; Marcus J Claesson – Email: m.claesson@ucc.ie; *Corresponding author

Received March 21, 2015; Revised March 26, 2015; Accepted March 29, 2015; Published April 30, 2015

Abstract:
To manage and intelligently mine the avalanche of genomic sequences intuitive and user-friendly graphical interfaces are required. Here we present BlastXtract2 which exclusively facilitates early exploration of un-annotated genomic and metagenomic sequences. Various formats of translated searches, including the commonly used BlastX, of multiple sequences against multiple protein databases can be uploaded to a relational database server, which can be accessed via a locally installed web-server. There, an intuitive GUI allows straightforward data-mining and enables quick detection of potential frameshifts and poorly sequenced or assembled regions, thereby contributing in making BlastXtract2 a unique and valuable tool for early exploration of (meta) genomic sequences.

Availability: Source code, documentation and an online demo version are available at https://github.com/ ClaessonLab/ BlastXtract2.

Background:
The development of user-friendly data-mining applications have for long been trailing the ever-increasing rate of genomic data generation. Even so, current next-generation sequencing technologies allow both faster and cheaper sequencing of genomes than what was previously possible. Genome projects have progressed from sequencing single chromosomes to genomes of multiple species and whole microbial communities (metagenomics). Translated searches (e.g. BlastX) of such multiple genomic sequences against protein databases are highly valuable for the early exploration of encoded functions. Translated searches normally precede and often complement gene prediction and annotation, which in turn require careful manual curation. Such search results can often also reveal regions of poorer quality of sequencing and/or assembly through detection of frame shifts. However, interpreting BlastX outputs by non-expert users is often limited by large impenetrable text-based flat files. There is therefore a need for intuitive visualisation and flexible mining of encoded functions directly from genomic sequences, which due to confidentiality often need to be analysed in-house. As data management increasingly is becoming a bottleneck for the growing amount of genomic information, text-based flat-files of automatic annotation are not an optimal solution. Relational databases, on the other hand, offer greater speed and a more flexible query-space, in addition to safer and more centralized storage that can be distributed across networks. A number of various tools for analysing, storing and visualising Blast results have been designed, which has been comprehensively reviewed **[1].** Of these, no software application has exclusively exploited translated searches when combined with multi-query analysis aided by an intuitive GUI. To meet these challenges we developed BlastXtract2, which is a substantial improvement over its first version **[2].**

Methodology:
BlastXtract2 was written in Perl with CGI, DBI and BioPerl Graphics modules installed. It runs on a Linux-based server and has been successfully tested on Ubuntu, Debian and Fedora. As it interacts with relational databases BlastXtract2 requires either MySQL or PostgreSQL to be installed in addition to the Apache web-server. The functionality and improvements over its previous version are further described in the Implementation section below.

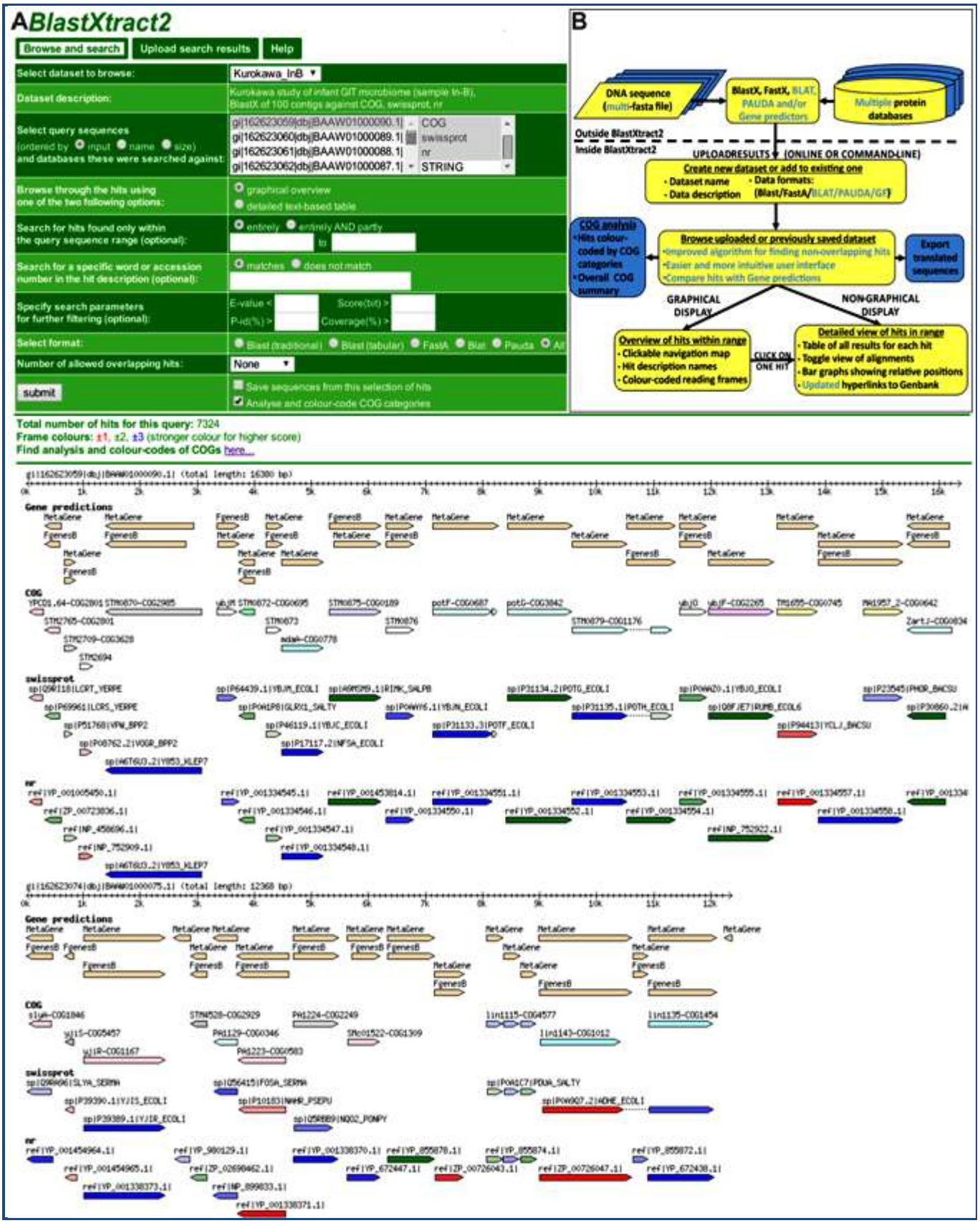

Figure 1: **A)** Screen-dump from BlastXtract2 web-server displaying gene predictions and BlastX hits in two contigs from a human gut metagenome against two databases. Note how the HSP visualisations highlight the incongruent gene predictions and potential frameshifts. COG hits are coloured by functional category; **B)** Flow-diagram of BlastXtract2 functionality where updated features are marked in blue.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 11(4):173-175 (2015)

Implementation

The first version of BlastXtract could only visualize and manage translated search results of a single sequence against a single protein database. **Figure 1b** illustrates the extensive updates of functionality that BlastXtract2 has undergone. Now, translated searches of multiple sequences against multiple sequence databases can be parsed and imported into a relational database as a single dataset, either via a locally installed web-server or from the command-line. The multiple sequences may represent assembled contigs of a draft genome, scaffolds, or a metagenome. Search results against an arbitrary number of databases can subsequently be mined in parallel. In addition to the traditional and tabular outputs of the omnipresent local aligner BlastX **[3],** output data from the slower but more sensitive global aligner FastX **[4],** and the much faster BLAT **[5]** and PAUDA **[6],** can also be parsed and imported into BlastXtraxt2. Moreover, while gene prediction is not the primary aim of BlastXtract2 there is an option to upload GFF files of gene predictions of the (meta)genomic sequences, as useful validation of the same when viewed in parallel with the multiple translated searches. Once uploaded, the search results can be displayed and browsed using either the graphical overview or the more detailed text-based table **(Figure 1a).** By clicking on a hit in the graphical overview, the latter table appears with specific details for that particular hit, while retaining the graphical overview on the same page. Hit features are coloured according to their reading frame, which facilitates immediate detection of potential frameshifts as differentially coloured high-scoring segment pairs (HSPs) belonging to the same hit, which also are connected by dashed lines **(Figure 1a).** Searching datasets in BlastXtract2 can be refined by entering query sequence ranges, words in hit-descriptions, E-value and Percent Identity cut-offs. A new feature also allows extraction of amino acid sequences of any search result, which can be useful for downstream sequence analysis. Functional COG categories are automatically assigned if the (meta)genomic sequences are searched against a slightly modified COG database (see *README.md* for instructions). The hits are subsequently colour coded according to COG categories and summarised in a separate

Utility:

As outlined in the Implementation section above, BlastXtract2 has undergone numerous improvements since its first version. These improvements have significantly increased its usefulness in the early exploration of both assembled sequences of both isolate genomes and metagenomes. This is something which our lab-members and collaborators attest to after having to mine multiple contigs as part of genomic and metagenomic projects.

Conclusion and future development:

In response to the lack of user-friendly graphical tools to augment manual annotation of (meta)genomes, we have developed BlastXtract2. In contrast to existing software applications for managing BLAST results, BlastXtract2 exclusively targets and utilises translated searches from a wider range of commonly used tools of varying speeds and sensitivities. Its intuitive visualisation of hits and HSPs facilitates rapid functional interpretation and detection of frameshifts prior to, and in parallel with, manual gene finding and annotation. That in combination with efficient storage and data-mining features, makes BlastXtract2 a valuable and unique addition to any scientist delving into un-annotated (meta)genomic sequences. Future developments may include parsing of even more formats and a paging function that allows an even higher number of contigs to be simultaneously browsed and searched.

Acknowledgement:

***Funding*:** This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 11/SIRG/B2162.

References:

[1] Neumann, RS *et al. Brief Bioinform* 2014 **15**: 4 [PMID: 23603091]
[2] Claesson MJ & van Sinderen D, *Bioinformatics* 2005 **21**: 3667 [PMID: 16046492]
[3] Altschul SF *et al. Nucleic Acids Res.* 1997 **25**: 3389 [PMID: 9254694]
[4] Pearson WR & Lipman DJ, *Proc Natl Acad Sci U S A.* 1988 **85**: 2444 [PMID: 3162770]
[5] Kent WJ, *Genome Res.* 2002 **12**: 656 [PMID: 11932250]
[6] Huson DH & Xie C, *Bioinformatics* 2014 **30**: 1 [PMID: 23658416]

Edited by P Kangueane

Citation: **Weerd *et al*.** Bioinformation 11(4): 173-175 (2015)