An online database for the detection of novel archaeal sequences in human ESTs

Neil F. W. Saunders1,∗, Paul M. G. Curmi2,3 and Ricardo Cavicchioli1

∗School of Biotechnology and Biomolecular Sciences and 2School of Physics, The University of New South Wales, Sydney, NSW 2052, Australia and 3Centre for Immunology, St Vincent’s Hospital, Sydney, NSW 2010, Australia

Received on November 11, 2003; revised on January 20, 2004; accepted on January 26, 2004
Advance Access publication April 15, 2004

ABSTRACT
Summary: We have developed a rapid, automated screening system and online database to detect foreign sequences of archaeal origin in human expressed sequence tags. The aim of the screening is to detect transcripts that may be derived from novel, putative archaeal pathogens or symbionts.
Contact: neil.saunders@unsw.edu.au
Supplementary information: http://psychro.bioinformatics.unsw.edu.au/pathogen/about.php

No archaea are known to be pathogens of humans (or any other organism). However, recent analyses indicate that there are few, if any, biological rationales that would preclude archaea from being, or becoming pathogens (Cavicchioli et al., 2003; Eckburg et al., 2003). Despite this, little effort has been directed towards the detection of potentially pathogenic archaea. One way in which novel pathogens may be detected is by examining transcripts from human tissues or cells for foreign sequences. It has been shown recently that a large number of sequences in the human subset of dbEST, the expressed sequence tag (EST) division of GenBank, are in fact of non-human origin and are often derived from known microbial pathogens of humans, such as bacteria, fungi and viruses (Weber et al., 2002; Xu et al., 2003). Therefore, it may be possible to detect putative novel pathogens by screening dbEST.

We have developed a rapid, automated screening system and online database to detect foreign sequences of archaeal origin in human ESTs. Screening is performed in two phases. First, archaeal query sequences are aligned to human EST sequences using BLAT (Kent, 2002). The output from BLAT (as BLAST-like format) is parsed to extract the accession numbers of those ESTs with similarity to an archaeal sequence. This subset of ESTs is then used to search the non-redundant nucleotide database using MEGABLAST (Zhang et al., 2000).

MEGABLAST output is parsed to summarize the data and the accession number of the top hit used to query the NCBI Entrez database to obtain taxonomic data for the hit sequence. Taxonomic information is appended to the MEGABLAST summary and the combined data imported to a MySQL database. A web front-end written in PHP is employed to view and query the final dataset.

The system was tested using a dataset of 5,264,202 EST sequences and 14,977 archaeal sequences, which comprised all human EST and archaeal sequences in GenBank at the time of the test. The majority of sequences identified in the initial BLAT screen were genuine human ESTs that are similar to their archaeal counterparts (e.g. translation elongation factors, ribosomal RNA). The remaining 1501 ESTs had as their top MEGABLAST hit sequences from eukarya, bacteria, viruses/vectors and archaea. These sequences fall into three groups of interest with respect to the query EST sequence. Many of the non-human eukaryal MEGABLAST hits are sequences from organisms closely related to humans (particularly mouse). These sequences are most probably derived from genuine human genes, but the human gene either is not in the BLAST database yet or has a slightly lower ranking MEGABLAST score than the non-human top hit. The majority of the bacterial MEGABLAST hits are sequences from species known to be associated with humans. This reflects both a higher number of shared archaeal/bacterial genes as compared with the archaeal/eukaryal genes and the degree to which human EST sequence data are contaminated by bacterial sequence. The third group of MEGABLAST hits is of the most interest: those for which the top hit is to an archaeal sequence. A total of 30 such ESTs were identified in the human EST dataset, of which 6 aligned to the archaeal sequence across >40 nt. Further BLAST analysis of these six sequences revealed five of them to be most similar to the Pfur DNA polymerase gene, derived from the archaean Pyrococcus furiosus. These sequences are presumably the result of contamination of the Pfur enzyme preparation, used during EST synthesis, with Pfur DNA. The sixth sequence (EST accession number
AW064312) had as its two best BLAST hits (using both blastn and blastx) a gene from the archaeal genus *Methanosarcina* encoding a subunit of a site-specific DNA methyltransferase. Till date, this EST is the best putative archaeal gene candidate from human ESTs.

There are clearly a number of explanations for the presence of an apparently archaeal sequence in human ESTs, besides the presence of archaea *in vivo* in human tissue. These include contamination of the tissue sample, horizontal transfer of an archaeal gene to a non-archaeal human-associated organism, convergent evolution or high conservation of certain sequences and missing human genes in the database leading to a non-human top hit. The identification of an EST as archaeal by BLAST similarity is therefore only the initial step of EST investigation. The database enables researchers to trace the origin and experimental history of a sequence as far as possible via hyperlinks to the NCBI database. Ultimately, it will be of considerable interest to obtain samples of those tissues in which putative archaeal transcripts are detected for further investigation using PCR, and to examine thoroughly tissue samples from undiagnosed diseases for the presence of archaeal DNA.

Our database may best be described as ‘a dataset containing the top MEGABLAST hit to the non-redundant nucleotide database using query sequences from the human EST database that have some similarity to archaeal sequences’. For both the BLAT alignment and the subsequent MEGABLAST search of the non-redundant nucleotide database, default low-stringency search parameters (bit score and *E*-value) are used. This gives rise to a high proportion of ESTs in the final dataset that are genuinely derived from human genes. However, this is preferable to failing to detect foreign transcripts through over-aggressive filtering and reduces the EST dataset to a manageable size (from over $5 \times 10^6$ to a few thousand sequences). Moreover, sequences of immediate interest are tagged as archaeal by querying the Entrez database for taxonomic annotation. The final interpretation is left in the hands of the user through a web front-end to the database. Features of the online analysis system include BLAST searches and hyperlinked accession numbers that allow the user to investigate fully both the EST query and its BLAST hit. The entire process from sequence download to the final database is performed using command-line tools and scripts, and so is an automated discovery system that can be run at set intervals as new sequences are deposited in GenBank.

ACKNOWLEDGEMENTS

We acknowledge valuable discussions with Torsten Thomas, Gregory Tyrelle and Stephen Harrop. This work was funded by the Australian Research Council (ARC) and N.F.W.S. was supported by an ARC Australian Postdoctoral Fellowship.

REFERENCES


