Available online at www.sciencedirect.com
SCIENCE DIRECT®
Experimental Parasitology 110 (2005) 178–183
Experimental Parasitology
www.elsevier.com/locate/yexpr

Entamoeba histolytica: Construction and applications of subgenomic databases

Margit Hofer, Michael Duchêne *

Department of Specific Prophylaxis and Tropical Medicine, Center for Physiology and Pathophysiology, Medical University of Vienna, Kinderspitalgasse 15, A-1095 Vienna, Austria

Received 1 February 2005; received in revised form 15 March 2005; accepted 15 March 2005
Available online 22 April 2005

Abstract

Knowledge about the influence of environmental stress such as the action of chemotherapeutic agents on gene expression in *Entamoeba histolytica* is limited. We plan to use oligonucleotide microarray hybridization to approach these questions. As the basis for our array, sequence data from the genome project carried out by the Institute for Genomic Research (TIGR) and the Sanger Institute were used to annotate parts of the parasite genome. Three subgenomic databases containing enzymes, cytoskeleton genes, and stress genes were compiled with the help of the ExPASy proteomics website and the BLAST servers at the two genome project sites. The known sequences from reference species, mostly human and *Escherichia coli*, were searched against TIGR and Sanger *E. histolytica* sequence contigs and the homologs were copied into a Microsoft Access database. In a similar way, two additional databases of cytoskeletal genes and stress genes were generated. Metabolic pathways could be assembled from our enzyme database, but sometimes they were incomplete as is the case for the sterol biosynthesis pathway. The raw databases contained a significant number of duplicate entries which were merged to obtain curated non-redundant databases. This procedure revealed that some *E. histolytica* genes may have several putative functions. Representative examples such as the case of the δ-aminolevulinate synthase/serine palmitoyltransferase are discussed.

Index Descriptors and Abbreviations: EC, enzyme commission; LGT, lateral gene transfer; ORF, open reading frame

Keywords: *Entamoeba histolytica*; Genome project; Database; Enzyme; Cytoskeleton; Stress response

1. Introduction

Entamoeba histolytica causes up to 100,000 deaths per year and represents a major health problem especially in developing countries (WHO, 1997). The *E. histolytica* genome project (Loftus et al., 2005) carried out by the Wellcome Trust Sanger Institute in Hinxton, Cambridge (UK) and the Institute for Genomic Research (TIGR) in Rockville, Maryland (USA), creates many new opportunities for the scientific community. Hopes are high to gain new insights into the molecular mechanisms of amoebiasis and find new targets for chemotherapy.

In our laboratory we plan to monitor the effects of metronidazole and alkylphosphocholines on the gene expression of *E. histolytica* with the help of oligonucleotide microarrays. Microaerophilic *E. histolytica* reduces the nitro group of metronidazole to cytotoxic nitro radicals via its reduced ferredoxin (Wassmann et al., 1999). Little is known about the mode of action of alkylphosphocholines such as miltefosine, but it is believed to interact with the lipid metabolism of the cells (Croft et al., 2003). Hybridization of drug-treated and control cDNAs to a focussed microarray with all

* Corresponding author. Fax: +43 1 4277 64899.
E-mail address: michael.duchene@meduniwien.ac.at (M. Duchêne).

0014-4894/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.exppara.2005.03.018

known enzymes of the redox system of the amoebae and their lipid metabolism enzymes could be an appropriate way to gain insight into the mechanism of action of the two chemotherapeutic agents. When we started our work, the genome sequences had not yet been annotated, therefore we had to compile our own database with genome information. A software program for such a database should be stable, create highly convertible data sets, be flexible, and easy to use and maintain. We decided to use the Microsoft Access database software. Here we describe the compilation of three different subgenomic databases containing enzymes, cytoskeleton genes, and stress genes found in the *E. histolytica* genome.

2. Materials and methods

The Microsoft Access 2000 database software was chosen as program for compiling our data. First we selected enzymes with known protein sequences from the web version of the Roche Applied Science (previously Boehringer–Mannheim) Biochemical Pathways (http://us.expasy.org/cgi-bin/show_thumbnails.pl). We followed the link to the ExPASy website (http://www. expasy.org/) with the Swiss-Prot and Trembl (http://www.expasy.org/sprot/) as well as the Enzyme (http://www.expasy.org/enzyme/) databases, where reference protein sequences for these enzymes were retrieved. These were used for a tblastn search with default settings against the TIGR (http://tigrblast.tigr.org/er-blast/index.cgi?project=eha1) and Sanger (http://www.sanger.ac.uk/cgi-bin/blast/submitblast/e_histolytica) DNA sequences of the *E. histolytica* genome project. Tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands) using the BLAST algorithm (Altschul et al., 1990). The tblastn output top hits were considered a significant match if their *P* values were in the range of 10^{-5}or better (lower than 9.9E−05), and a complete database entry was generated. We also recorded the cases where no significant match was found. The exact position of the homologous sequence in the contig was established with the help of the Expasy translate tool (http://www.expasy.org/tools/dna.html) as well as the UWGCG package (Program Manual for the Wisconsin Package Version 8.1-Unix, see References). The theoretical isoelectric point (p*I*) and molecular weight of the found homolog were calculated with the help of the ExPASy p*I*/Mw tool (http://www.expasy.org/tools/pi_tool.html). In total, a set of up to 30 parameters were recorded for each database entry. Biochemical pathways including all found *E. histolytica* enzyme homologs were established with the help of the Kyoto Encyclopedia of Genes and Genomes (KEGG) website (http://www.genome.jp/kegg/tool/search_pathway.html) and displayed in comparison to the Standard Data Set.

3. Results and discussion

3.1. Establishment and use of the annotation databases

The purpose of our database work was to get a collection of annotated *E. histolytica* genes which should be the basis of our focussed oligonucleotide microarray. Although we planned to specifically target enzymes of the lipid metabolism and those known to be involved in reactions to metronidazole chemotherapy, we also wanted to get an overall picture of the metabolism of *E. histolytica*. Therefore we searched for all enzymes in the Roche/Boehringer "Biochemical Pathways" chart. In addition, we tried to find homologs of a list of cytoskeleton genes and stress genes. For better organisation we compiled three independent databases of enzymes, cytoskeleton genes, and stress genes.

A number of data sets were extracted for each gene into our Microsoft Access database. Each homologous *E. histolytica* sequence found with a *P* value in the range of 10^{-5} or better was used to generate a complete database entry. The *P* value and the percentage of protein sequence identity were entered. For the reference sequence the accession number, the enzyme commission number, and the species were recorded, as well as how many hits had been found in total. Next, we entered the *E. histolytica* protein coding DNA sequence, the deduced open reading frame (ORF), the position of the ORF within the contig, if it was a full-length ORF, and the theoretical p*I*/Mw. References were given where *E. histolytica* genes had been published before, and sufficient space for comments was included, for example, we recorded if there were conflicts with the published sequence. Recently we also introduced a field for the date of last revision to keep track of changes.

For classification of the entries we introduced status numbers ranging from 1 to 6. Number 1 means a published *E. histolytica* gene sequence, 2 is a sequence with high degree of end-to end similarity, 3 a sequence with intermediate degree of similarity, 4 a sequence with weak similarity, 5 means that there was a sequence for searching but no homology found, and 6 indicates that the enzyme activity is known but there is no protein reference sequence in the databases for searching.

The intermediate result of our searches after going through the Biochemical Pathways chart was that out of 1536 enzyme genes searched for, 962 genes with at least some sequence similarity (status 1–4) were detected, searching for 295 cytoskeleton genes led to 281 hits (status 1–4), and 109 hits were obtained by searching for 120 putative stress-related genes.

Inspection of the databases revealed that the same *E. histolytica* genes sometimes had been detected in different searches. When performing a tblastn search with human alcohol dehydrogenase α-chain (EC 1.1.1.1; Accession No. P07327) for instance, we obtained three weakly homologous *E. histolytica* sequences which were entered