



LABEL
YOUR
DATA

Case study

Information Services

 NDA

DOCUMENT ANNOTATION FOR COMPLIANCE BUSINESS

Overview

RegTech company from the UK has partnered with a multilingual annotation team at Label Your Data to address the challenge of secure annotation and categorization of over 1 million legal documents.

Dataset size:

1.2 million documents

Team:

30 dedicated annotators

Labeling tool:

LabelStudio

Timeline:

September 2022 — ongoing

Client

A UK-based RegTech company offers regulatory infrastructure and data as a service for organizations and regulatory authorities.

Challenges

The regulatory compliance company struggled with legal document annotation due to complex guidelines requiring legal expertise and multilingual support.

Solutions

Label Your Data's multilingual team with legal expertise annotated and categorized complex legal documents in English, Spanish, and French.

Results

The RegTech company achieved faster turnaround times and streamlined information retrieval by leveraging Label Your Data's secure, multilingual legal document annotation.

Client

A regulatory technology (RegTech) company from the United Kingdom delivers regulatory infrastructure and data as a service. By leveraging AI, they structure complex regulations into a machine-readable format.

Their technology-driven services help both organizations to comply with regulations and regulatory authorities to manage and disseminate them.



Challenges

Inactivation and Regulation of the Aerobic C₄-Dicarboxylate Transport (*dctA*) Gene of *Escherichia coli*

SUZANNE J. DAVIES,¹ PAUL GOLBY,² DAVOOD OMRANI,^{1,†} SUSAN A. BROAD,²
VIKKI L. HARRINGTON,² JOHN R. GUEST,¹ DAVID J. KELLY,¹
AND SIMON C. ANDREWS^{2,*}

Krebs Institute for Biomolecular Research, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN,¹ and School of Animal & Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ,² United Kingdom

Received 16 February 1999/Accepted 2 July 1999

The gene (*dctA*) encoding the aerobic C₄-dicarboxylate transporter (*DctA*) of *Escherichia coli* was previously mapped to the 79-min region of the linkage map. The nucleotide sequence of this region reveals two candidates for the *dctA* gene: *f428* at 79.3 min and the *o157a-o424-o328* (or *orfQMP*) operon at 79.9 min. The *f428* gene encodes a homologue of the *Sinorhizobium meliloti* and *Rhizobium leguminosarum* H⁺/C₄-dicarboxylate symporter, *DctA*, whereas the *orfQMP* operon encodes homologues of the aerobic periplasmic-binding protein-dependent C₄-dicarboxylate transport system (*DetQ*, *DetM*, and *DetP*) of *Rhodobacter capsulatus*. To determine which, if either, of these loci specify the *E. coli* *DctA* system, the chromosomal *f428* and *orfM* genes were inactivated by inserting Sp^r or Ap^r cassettes, respectively. The resulting *f428* mutant was unable to grow aerobically with fumarate or malate as the sole carbon source and grew poorly with succinate. Furthermore, fumarate uptake was abolished in the *f428* mutant and succinate transport was ~10-fold lower than that of the wild type. The growth and fumarate transport deficiencies of the *f428* mutant were complemented by transformation with an *f428*-containing plasmid. No growth defect was found for the *orfM* mutant. In combination, the above findings confirm that *f428* corresponds to the *dctA* gene and indicate that the *orfQMP* products play no role in C₄-dicarboxylate transport. Regulation studies with a *dctA-lacZ* (*f428-lacZ*) transcriptional fusion showed that *dctA* is subject to cyclic AMP receptor protein (CRP)-dependent catabolite repression and ArcA-mediated anaerobic repression and is weakly induced by the DcuS-DcuR system in response to C₄-dicarboxylates and citrate. Interestingly, in a *dctA* mutant, expression of *dctA* is constitutive with respect to C₄-dicarboxylate induction, suggesting that *DctA* regulates its own synthesis. Northern blot analysis revealed a single, monocistronic *dctA* transcript and confirmed that *dctA* is subject to regulation by catabolite repression and CRP. Reverse transcriptase-mediated primer extension indicated a single transcriptional start site centered 81 bp downstream of a strongly predicted CRP-binding site.

Escherichia coli can utilize C₄-dicarboxylates as a carbon and energy source under aerobic and anaerobic conditions (9, 50, 56). Anaerobically, the uptake, exchange, and efflux of C₄-dicarboxylates (fumarate, malate, maleate, and succinate) and L-aspartate are mediated by the three independent dicarboxylate uptake (Dcu) systems, DcuA, DcuB, and DcuC (9, 12, 13, 50, 56). These Dcu systems appear to be active solely under anaerobic conditions (9).

Aerobically, uptake of C₄-dicarboxylates is mediated by a secondary transporter and/or a binding-protein-dependent system, designated Det (20, 24). The Det system has an apparent K_m of 10 to 20 μM for C₄-dicarboxylates and is driven by the electrochemical proton gradient (15), and its activity is induced by succinate and is subject to catabolite repression (20, 27). The corresponding *dctA* mutants cannot utilize the C₄-dicarboxylates malate and fumarate but grow normally on the monocarboxylate lactate (27). Transport across the outer membrane may be mediated by a C₄-dicarboxylate-binding protein (Cbt; K_d for C₄-dicarboxylates of 30 to 50 μM) and a porin (3, 4, 25–30).

Three genetic loci (*dctB* at 16.6 min, *dctC* at 79.3 min, and *dctD* at 16.4 min) are involved in aerobic C₄-dicarboxylate transport (27). The nucleotide sequence of the 76- to 81.5-min region revealed a putative *dctA* gene (*f428*) encoding a protein (*DctA*) most closely resembling the *DctA* proteins of *Sinorhizobium meliloti* and *Rhizobium leguminosarum* (62 to 63% identity) that function as H⁺/C₄-dicarboxylate symporters (51). The *DctA* proteins are members of a family that includes the Na⁺/H⁺ glutamate symporters (*GltB*/*GltT*). A role for the putative *dctA* gene of *E. coli* in the utilization of C₄-dicarboxylates (and the cyclic monocarboxylate orotate) has been suggested by complementation studies with *Salmonella typhimurium* *dctA* or *ouaA* mutants (2, 51). The coding regions corresponding to the *dctB* (predicted to encode an inner membrane protein) and *dctC* (predicted to encode the binding protein) genes have yet to be identified (23).

In addition to the *dctA* (*f428*) gene, *E. coli* contains three apparently cotranscribed genes (*o157a-o424-o328* or *orfQMP*) at 79.9 min that encode products 23 to 32% identical to the DetPQM components of the periplasmic binding protein-dependent C₄-dicarboxylate transport system of *Rhodobacter capsulatus* (11, 46, 51). The *orfQMP* genes are apparently part of a large operon involved in pentose sugar metabolism (11, 42). This suggests that the *orfQMP* products form a pentose sugar transporter, although, given their similarity to the *R. capsulatus* DetPQM components, it is also possible that they transport C₄-dicarboxylates.

* Corresponding author. Mailing address: School of Animal & Microbial Sciences, University of Reading, Whiteknights, P.O. Box 228, Reading RG6 6AJ, United Kingdom. Phone: 118-987-5123, ext. 7045/7886. Fax: 118-931-0180. E-mail: s.c.andrews@reading.ac.uk.

† Present address: Department of Genetics, Uroimeya Medical School, Uroimeya, Iran.

A regulatory compliance business faced a challenge of annotating their legal documents. Their complex guidelines required legal background to understand and categorize the documents' content and structure.

Additionally, strong English skills were crucial for smooth communication during instruction updates.



Solution

To tackle complex legal document annotation, Label Your Data built a multilingual team. A custom test for annotators ensured their strong legal knowledge.

Clear communication with the client was fostered through regular sync-ups in the beginning of the project.

Label Your Data's multilingual team of 30 annotators provided:



Document annotation in English, Spanish, and French



Document categorization using 50 labels



Data extraction



Data validation



Results

The project aimed to streamline human review of legal document annotations. Recognizing the regulatory technology company's need for human oversight throughout the labeling process, Label Your Data provided a highly scalable team of annotators. They efficiently addressed the need for precise and secure document labeling across multiple languages.

1.

1M+ legal documents annotated

2.

Streamlined information retrieval from a massive dataset

3.

Enhanced data accessibility

4.

Faster turnaround times for client inquiries

5.

Efficient regulatory compliance solutions



Run a free pilot for your text analysis project!



Security-compliant

Work with a compliant vendor: PCI DSS Level 1, ISO 27001, GDPR, and CCPA.



Flexible pricing

Choose the pricing option that best suits your project's timeline.



No commitment

Run a free pilot to experience our document annotation services without any obligation.



Tool-agnostic

Get annotated files using your or our tools.