Scraping

- **quickscrape** scrapes URLs with the help of
- **scraper definitions** in order to find and retrieve metadata and additional material

SHUTTLEWORTH FUNDED

OA *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum *Thaumarchaeota*

Authors: Michaela Stieglmeier[1], Andreas Klingl[2], Ricardo J. E. Alves[1], Simon K.-M. R. Rittmann[1], Michael Melcher[1], Nikolaus Leisch[1], Christa Schleper[1]

VIEW AFFILIATIONS

Correspondence Christa Schleper christa.schleper@univie.ac.at

PDF 1,006.77 Kb

HTML 223.50 Kb

Int J Syst Evol Microbiol, August 2014 64: 2738-2752, doi: 10.1099/ijs.0.063172-0

Subject: New Taxa - Archaea

Published Online: 01/08/2014

Abstract | Fulltext | Figures (5) | References (143) | Cited By (11) | Supplementary Data (1) | Related Content

Preview this:

A mesophilic, neutrophilic and aerobic, ammonia-oxidizing archaeon, strain EN76^T, was isolated from garden soil in Vienna (Austria). Cells were irregular cocci with a diameter of 0.6–0.9 µm and possessed archaella and archaeal pili as cell appendages. Electron microscopy also indicated clearly discernible areas of high and low electron density, as well as tubule-like structures. Strain EN76^T had an S-layer with p3 symmetry, so far only reported for members of the *Sulfolobales* . Crenarchaeol was the major core lipid. The organism gained energy by oxidizing ammonia to nitrite aerobically, thereby fixing CO_2, but growth depended on the addition of small amounts of organic acids. The optimal growth temperature was 42 °C and the optimal pH was 7.5, with ammonium and pyruvate concentrations of 2.6 and 1 mM, respectively. The genome of strain EN76^T had a DNA G+C content of 52.7 mol%. Phylogenetic analyses of 16S rRNA genes showed that strain EN76^T is affiliated with the recently proposed phylum *Thaumarchaeota* , sharing 85% 16S rRNA gene sequence identity with the closest cultivated relative '*Candidatus* Nitrosopumilus maritimus' SCM1, a marine ammonia-oxidizing archaeon, and a maximum of 81% 16S rRNA gene sequence identity with members of the phyla *Crenarchaeota* and *Euryarchaeota* and any of the other recently proposed phyla (e.g. '*Korarchaeota*' and '*Aigarchaeota*'). We propose the name *Nitrososphaera viennensis* gen. nov., sp. nov. to accommodate

OA *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum *Thaumarchaeota*

Title

Authors: Michaela Stieglmeier[1], Andreas Klingl[2], Ricardo J. E. Alves[1], Simon K.-M. R. Rittmann[1], Michael Melcher[1], Nikolaus Leisch[1], Christa Schleper[1]

VIEW AFFILIATIONS

Authors

Correspondence Christa Schleper christa.schleper@univie.ac.at

PDF 1,006.77 Kb

HTML 223.50 Kb

Downloadables

Int J Syst Evol Microbiol, August 2014 64: 2738-2752, doi: 10.1099/ijs.0.063172-0

Subject: New Taxa - Archaea

Published Online: 01/08/2014

Followables

Abstract | Fulltext | Figures (5) | References (143) | Cited By (11) | Supplementary Data (1) | Related Content

Preview this:

A mesophilic, neutrophilic and aerobic, ammonia-oxidizing archaeon, strain EN76^T, was isolated from garden soil in Vienna (Austria). Cells were irregular cocci with a diameter of 0.6–0.9 µm and possessed archaella and archaeal pili as cell appendages. Electron microscopy also indicated clearly discernible areas of high and low electron density, as well as tubule-like structures. Strain EN76^T had an S-layer with p3 symmetry, so far only reported for members of the *Sulfolobales* . Crenarchaeol was the major core lipid. The organism gained energy by oxidizing ammonia to nitrite aerobically, thereby fixing CO_2, but growth depended on the addition of small amounts of organic acids. The optimal growth temperature was 42 °C and the optimal pH was 7.5, with ammonium and pyruvate concentrations of 2.6 and 1 mM, respectively. The genome of strain EN76^T had a DNA G+C content of 52.7 mol%. Phylogenetic analyses of 16S rRNA genes showed that strain EN76^T is affiliated with the recently proposed phylum *Thaumarchaeota* , sharing 85% 16S rRNA gene sequence identity with the closest cultivated relative '*Candidatus* Nitrosopumilus maritimus' SCM1, a marine ammonia-oxidizing archaeon, and a maximum of 81% 16S rRNA gene sequence identity with members of the phyla *Crenarchaeota* and *Euryarchaeota* and any of the other recently proposed phyla (e.g. '*Korarchaeota*' and '*Aigarchaeota*'). We propose the name *Nitrososphaera viennensis* gen. nov., sp. nov. to accommodate

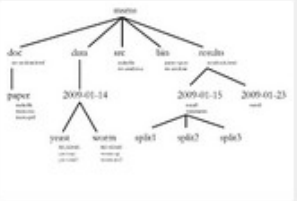
Abstract

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble

Published: July 31, 2009 • DOI: 10.1371/journal.pcbi.1000424 • Featured in PLOS Collections

Figures



Citation: Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424

Editor: Fran Lewitter, Whitehead Institute, United States of America

Published: July 31, 2009

Funding: The author received no specific funding for writing this article.

Competing interests: The author has declared that no competing interests exist.

Introduction

Most bioinformatics coursework focuses on algorithms, with perhaps some components

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble

Author

Title

Published: July 31, 2009 • DOI: 10.1371/journal.pcbi.1000424 • Featured in PLOS Collections

Downloadable

Article | Authors | Metrics | Comments | Related Content

Download PDF
Print
Share

Figures

Citation: Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424

Editor: Fran Lewitter, Whitehead Institute, United States of America

Published: July 31, 2009

Funding: The author received no specific funding for writing this article.

Competing interests: The author has declared that no competing interests exist.

Beginning of fulltext

Introduction

Most bioinformatics coursework focuses on algorithms, with perhaps some components

CrossMark

Included in the Following Collection

PLOS Computational Biology: Education

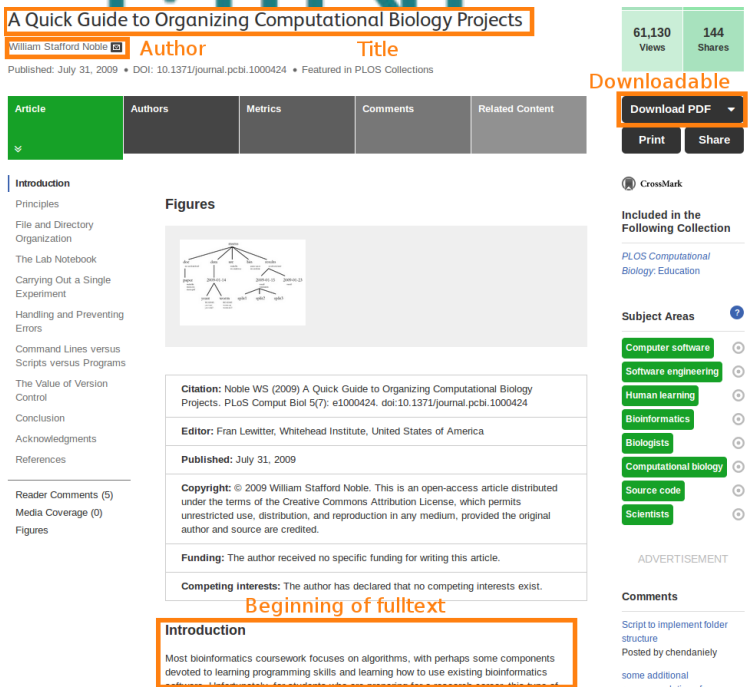
Subject Areas

- Computer software
- Software engineering
- Human learning
- Bioinformatics
- Biologists
- Computational biology
- Source code
- Scientists

Comments

Script to implement folder structure
Posted by chendaniely

Scraping



scraper definitions, per publisher

Paper_A
fulltext.html

- Title
- Authors
- Other Metadata
- Fulltext

possibly XML or PDF

Paper_B
fulltext.html

- Title
- Authors
- Other Metadata
- Fulltext

possibly XML or PDF

SHUTTLEWORTH FUNDED

BASIC SCRAPER JSON

name of the scraper:

the URL(s) it applies to:

the elements to capture:

element name:

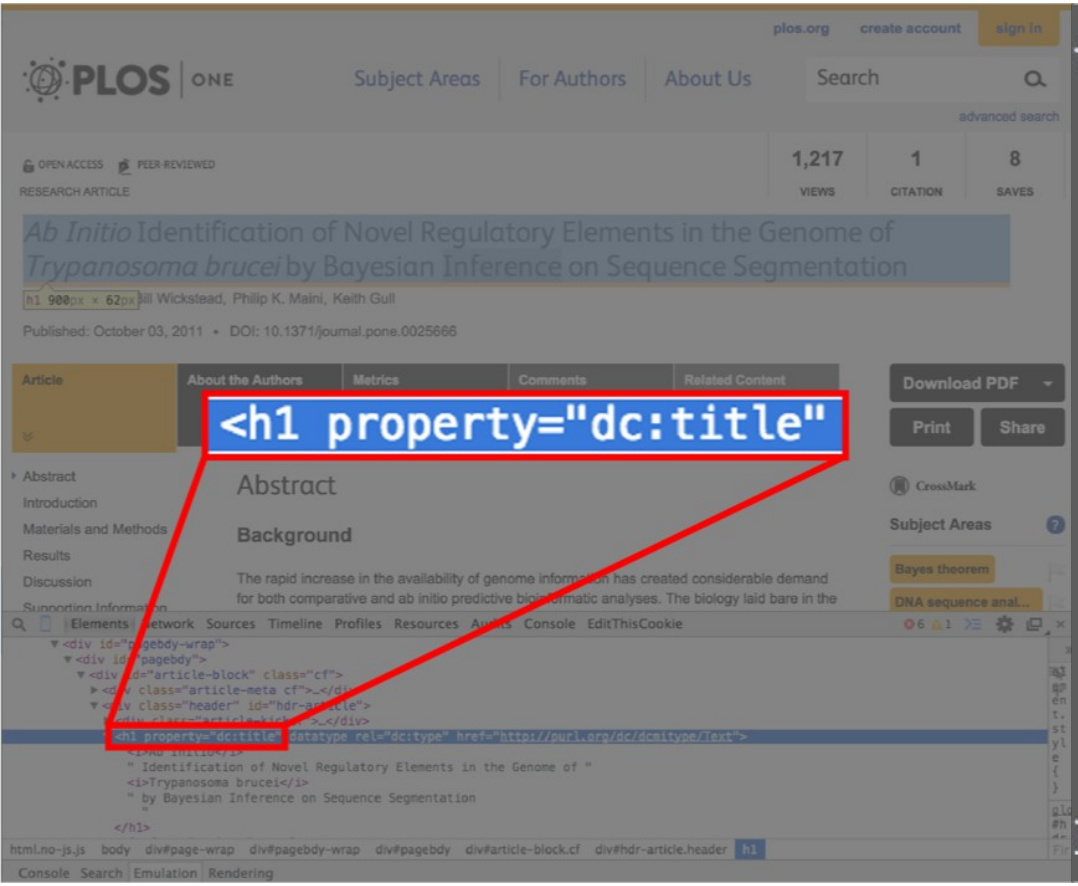
where to find it:

```
{
  "name": "PLOS",
  "url": "plos\\w*.org",
  "elements": {
    "title": {
      "selector": "//h1[@property='dc:title']",
    }
  }
}
```

http://github.com/ContentMine/scraperJSON

SHUTTLEWORTH FUNDED

Scraper definitions




```
{
  "name": "PLoS",
  "url": "plos\\w*.org",
  "elements": {
    "title": {
      "selector": "//h1[@property='dc:title']",
    }
  }
}
```

SHUTTLEWORTH FUNDED

Output: bibjson


```
{

  "title": "Ab Initio Identification of Novel
Regulatory Elements in the Genome of Trypanosoma
brucei by Bayesian Inference on Sequence
Segmentation"

}
```


SHUTTLEWORTH FUNDED

Let's get our hands dirty

We will:

- Run quickscrape
- Compare source-html with scraper output
- Get some intuition about scraping