

# Scraping

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



**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<sup>1</sup>, Andreas Klingl<sup>2</sup>, Ricardo J. E. Alves<sup>1</sup>, Simon K.-M. R. Rittmann<sup>1</sup>, Michael Melcher<sup>1</sup>, Nikolaus Leisch<sup>1</sup>, Christa Schleper<sup>1</sup>

[+ VIEW AFFILIATIONS](#)

**Correspondence** Christa Schleper [christa.schleper@univie.ac.at](mailto:christa.schleper@univie.ac.at)

*Int J Syst Evol Microbiol*, August 2014 64: 2738-2752, doi: [10.1099/ij.s.0.063172-0](https://doi.org/10.1099/ij.s.0.063172-0)

**Subject:** New Taxa - Archaea

Published Online: 01/08/2014

This is an open access article published by the Society for General Microbiology under the [Creative Commons Attribution License](#)



PDF  
1,006.77 Kb



HTML  
223.50 Kb

**Abstract**

[Fulltext](#)

[Figures \(5\)](#)

[References \(143\)](#)

[Cited By \(11\)](#)

[Supplementary Data \(1\)](#)

[Related Content](#)

A mesophilic, neutrophilic and aerobic, ammonia-oxidizing archaeon, strain EN76<sup>T</sup>, was isolated from garden soil in Vienna (Austria). Cells were irregular cocci with a diameter of 0.6–0.9 µm and possessed archaeella 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<sup>T</sup> 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<sub>2</sub>, 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<sup>T</sup> had a DNA G+C content of 52.7 mol%. Phylogenetic analyses of 16S rRNA genes showed that strain EN76<sup>T</sup> 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

[Preview this:](#)



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* **Title**

**Authors:** Michaela Stieglmeier<sup>1</sup>, Andreas Klingl<sup>2</sup>, Ricardo J. E. Alves<sup>1</sup>, Simon K.-M. R. Rittmann<sup>1</sup>, Michael Melcher<sup>1</sup>, Nikolaus Leisch<sup>1</sup>, Christa Schleper<sup>1</sup>

**+ VIEW AFFILIATIONS**

## Authors

**Correspondence** Christa Schleper [christa.schleper@univie.ac.at](mailto:christa.schleper@univie.ac.at)

*Int J Syst Evol Microbiol*, August 2014 64: 2738-2752, doi: [10.1099/ij.s.0.063172-0](https://doi.org/10.1099/ij.s.0.063172-0)

**Subject:** New Taxa - Archaea

Published Online: 01/08/2014

This is an open access article published by the Society for General Microbiology under the [Creative Commons Attribution License](#)



PDF  
1,006.77 Kb



HTML  
223.50 Kb

## Downloadables

## Followables

**Abstract**

[Fulltext](#)

[Figures \(5\)](#)

[References \(143\)](#)

[Cited By \(11\)](#)

[Supplementary Data \(1\)](#)

[Related Content](#)

A mesophilic, neutrophilic and aerobic, ammonia-oxidizing archaeon, strain EN76<sup>T</sup>, was isolated from garden soil in Vienna (Austria). Cells were irregular cocci with a diameter of 0.6–0.9 µm and possessed archaeella 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<sup>T</sup> 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<sub>2</sub>, 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<sup>T</sup> had a DNA G+C

## Abstract

content of 52.7 mol%. Phylogenetic analyses of 16S rRNA genes showed that strain EN76<sup>T</sup> 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

**Preview this:**



SHUTTLEWORTH  
FUNDED



# 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

61,130  
Views

144  
Shares

Article

Authors

Metrics

Comments

Related Content



## Introduction

Principles

File and Directory  
Organization

The Lab Notebook

Carrying Out a Single  
Experiment

Handling and Preventing  
Errors

Command Lines versus  
Scripts versus Programs

The Value of Version  
Control

Conclusion

Acknowledgments

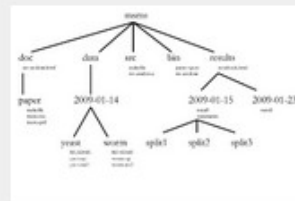
References

Reader Comments (5)

Media Coverage (0)

Figures

## 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

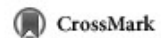
**Copyright:** © 2009 William Stafford Noble. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**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 devoted to learning programming skills and learning how to use existing bioinformatics



Included in the  
Following Collection

*PLOS Computational  
Biology: Education*

## Subject Areas

Computer software

Software engineering

Human learning

Bioinformatics

Biologists

Computational biology

Source code

Scientists

ADVERTISEMENT

## Comments

*Script to implement folder  
structure*

Posted by chendaniely

H





# 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

Article

Authors

Metrics

Comments

Related Content



## Introduction

Principles

File and Directory  
Organization

The Lab Notebook

Carrying Out a Single  
Experiment

Handling and Preventing  
Errors

Command Lines versus  
Scripts versus Programs

The Value of Version  
Control

Conclusion

Acknowledgments

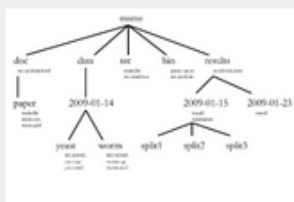
References

Reader Comments (5)

Media Coverage (0)

Figures

## 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

**Copyright:** © 2009 William Stafford Noble. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**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 devoted to learning programming skills and learning how to use existing bioinformatics

61,130  
Views

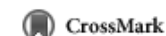
144  
Shares

Downloadable

Download PDF

Print

Share



Included in the  
Following Collection

PLOS Computational  
Biology: Education

Subject Areas

Computer software

Software engineering

Human learning

Bioinformatics

Biologists

Computational biology

Source code

Scientists

ADVERTISEMENT

Comments

Script to implement folder  
structure

Posted by chendaniely

H

# Scrapping

## 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

Article	Authors	Metrics	Comments	Related Content
▼				

Downloadable

Download PDF  
Print Share

### Introduction

Principles  
File and Directory Organization  
The Lab Notebook  
Carrying Out a Single Experiment  
Handling and Preventing Errors  
Command Lines versus Scripts versus Programs  
The Value of Version Control  
Conclusion  
Acknowledgments  
References

Reader Comments (5)  
Media Coverage (0)  
Figures

### 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

**Copyright:** © 2009 William Stafford Noble. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**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 devoted to learning programming skills and learning how to use existing bioinformatics

**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<sup>1</sup>, Andreas Kling<sup>2</sup>, Ricardo J. E. Alves<sup>1</sup>, Simon K.-M. R. Rittmann<sup>1</sup>, Michael Melcher<sup>1</sup>, Nikolaus Leisch<sup>1</sup>, Christa Schleper<sup>1</sup>

VIEW AFFILIATIONS

Authors

Correspondence Christa Schleper [christa.schleper@univie.ac.at](mailto:christa.schleper@univie.ac.at)

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

Subject: New Taxa - Archaea

Published Online: 01/08/2014

This is an open access article published by the Society for General Microbiology under the Creative Commons Attribution License

Followables

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

**Abstract**

A mesophilic, neutrophilic and aerobic, ammonia-oxidizing archaeon, strain EN76<sup>T</sup>, 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<sup>T</sup> 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<sub>2</sub>, 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 2.0 and 1 mM, respectively. The genome of strain EN76<sup>T</sup> had a DNA G+C content of 52.7 mol%. Phylogenetic analyses of 16S rRNA genes showed that strain EN76<sup>T</sup> 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. *Xorarchaeota* and *Algararchaeota*). We propose the name *Nitrososphaera viennensis* gen. nov., sp. nov. to accommodate

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

ADVERTISEMENT

Comments

Script to implement folder structure  
Posted by chendaniely  
some additional

Downloadables

PDF  
1,006.77 Kb  
HTML  
223.50 Kb

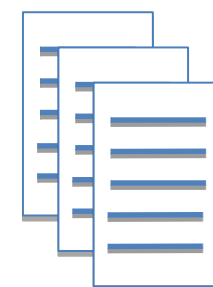
scraper definitions,  
per publisher



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:

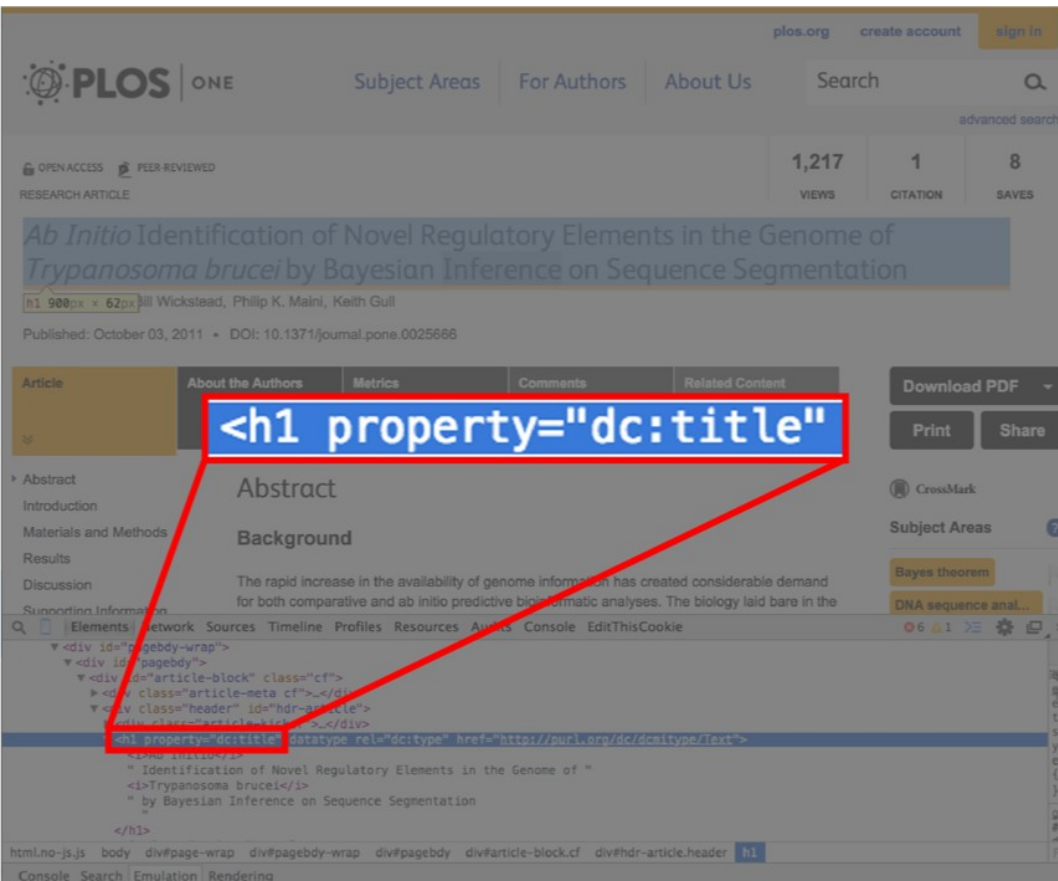
the elements to capture:

element name:

where to find it:

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

# Scraper definitions



The screenshot shows a PLOS ONE article page. A red box highlights the title element in the browser's developer tools. The title is: **<h1 property="dc:title"**. The article title is: *Ab Initio Identification of Novel Regulatory Elements in the Genome of Trypanosoma brucei* by Bayesian Inference on Sequence Segmentation. The authors are: Bill Wickstead, Phillip K. Maini, Keith Gull. The article was published on October 03, 2011, with DOI: 10.1371/journal.pone.0025666.

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



# Output: bibjson

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



# Let's get our hands dirty

We will:

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