

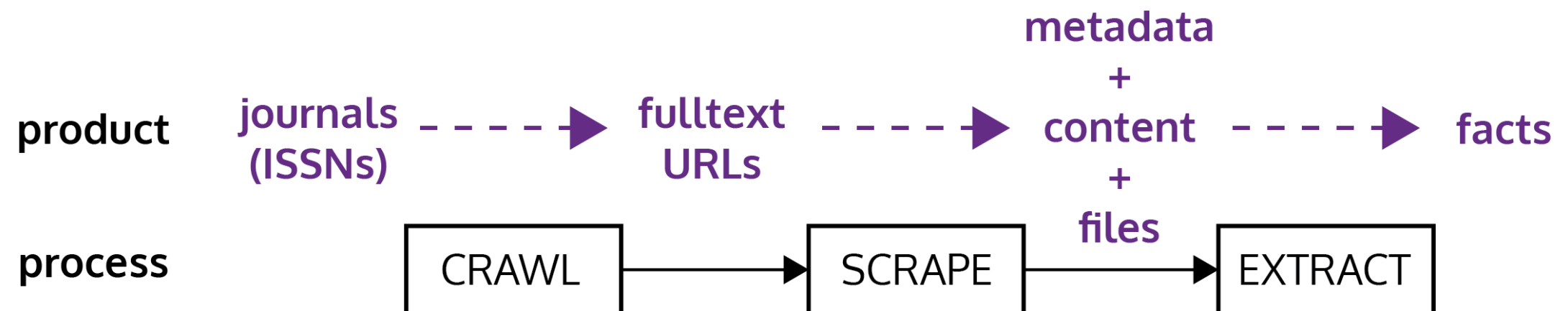
The ContentMine Scraping Stack

Richard Smith-Unna

The scale of the task

- ~ 27,000 peer reviewed journals (Ulrich's)
- > 5,000 publishers
- new papers every day

The pipeline



Introducing scraperJSON

- **observation:** scrapers all have the same plumbing
- **realisation:** each definition should ignore the plumbing, and focus on configuring the software for the target
- **conclusion:** a scraper definition is just a config file (in JSON format)
- **benefits**
 - supports large collections of scrapers
 - no programming required
 - not limited to one piece of software

Basic scraperJSON

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']"  
    }  
  }  
}
```



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

Steven Kelly , Bill Wickstead, Philip K. Maini, Keith Gull

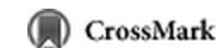
Published: October 03, 2011 • DOI: 10.1371/journal.pone.0025666

Article	About the Authors	Metrics	Comments	Related Content

Download PDF

Print

Share



Subject Areas

Bayes theorem

DNA sequence anal...

Gene expression

Gene regulation

Morphogenic segme...

Nucleotide sequenci...

Sequence analysis

Sequence motif anal...

Abstract

Introduction

Materials and Methods

Results

Discussion

Supporting Information

Author Contributions

References

Reader Comments (0)

Figures

Abstract

Background

The rapid increase in the availability of genome information has created considerable demand for both comparative and ab initio predictive bioinformatic analyses. The biology laid bare in the genomes of many organisms is often novel, presenting new challenges for bioinformatic interrogation. A paradigm for this is the collected genomes of the kinetoplastid parasites, a group which includes *Trypanosoma brucei* the causative agent of human African trypanosomiasis. These genomes, though outwardly simple in organisation and gene content, have historically challenged many theories for gene expression regulation in eukaryotes.

Methodology/Principle Findings

Here we utilise a Bayesian approach to identify local changes in nucleotide composition in the genome of *T. brucei*. We show that there are several elements which are found at the starts and ends of multicopy gene arrays and that there are compositional elements that are common to all intergenic regions. We also show that there is a composition-inversion element that occurs at the position of the trans-splice site.

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

1,217

VIEWS

1

CITATION

8

SAVES

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

900px × 62px Bill Wickstead, Philip K. Maini, Keith Gull

Published: October 03, 2011 • DOI: 10.1371/journal.pone.0025666

Article

About the Authors

Metrics

Comments

Related Content

[Download PDF](#)

Print

Share

► Abstract

Abstract

Introduction

Materials and Methods

Results

Discussion

Supporting Information

Background

The rapid increase in the availability of genome information has created considerable demand for both comparative and ab initio predictive bioinformatic analyses. The biology laid bare in the

 CrossMark

Subject Areas

Bayes theorem

DNA sequence anal...

🔍
📄
Elements
Network
Sources
Timeline
Profiles
Resources
Audits
Console
EditThisCookie


```
▼ <div id="pagebody-wrap">
  ▼ <div id="pagebody">
    ▼ <div id="article-block" class="cf">
      ▶ <div class="article-meta cf">...</div>
      ▼ <div class="header" id="hdr-article">
        ▶ <div class="article-kicker">...</div>
        ▼ <h1 property="dc:title" datatype rel="dc:type" href="http://purl.org/dc/dcmitype/Text">
          <i>Ab Initio</i>
          " Identification of Novel Regulatory Elements in the Genome of "
          <i>Trypanosoma brucei</i>
          " by Bayesian Inference on Sequence Segmentation
          ""
        </h1>
      </div>
    </div>
  </div>
</div>
```

html.no-js.js body div#page-wrap div#pagebdy-wrap div#pagebdy div#article-block.cf div#hdr-article.header **h1**

Console Search Emulation Rendering



1,217

VIEWS

1

CITATION

8

SAVES

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

h1 900px x 62px Bill Wickstead, Philip K. Maini, Keith Gull

Published: October 03, 2011 • DOI: 10.1371/journal.pone.0025666

Article

About the Authors

Metrics

Comments

Related Content

Download PDF

Print

Share

CrossMark

Subject Areas

Bayes theorem

DNA sequence anal...

property="dc:title"

Abstract

Background

The rapid increase in the availability of genome information has created considerable demand for both comparative and ab initio predictive bioinformatic analyses. The biology laid bare in the

Elements Network Sources Timeline Profiles Resources Audits Console EditThisCookie

```
<div id="pagebody-wrap">
  <div id="pagebody">
    <div id="article-block" class="cf">
      <div class="article-meta cf">...</div>
      <div class="header" id="hdr-article">
        <div class="article-kicker">...</div>
        <h1 property="dc:title" datatype rel="dc:type" href="http://purl.org/dc/dcmitype/Text">
          <i>Ab Initio</i>
          " Identification of Novel Regulatory Elements in the Genome of "
          <i>Trypanosoma brucei</i>
          " by Bayesian Inference on Sequence Segmentation
        </h1>
```


html.no-js.js body div#page-wrap div#pagebody-wrap div#pagebody div#article-block.cf div#hdr-article.header h1

Console Search Emulation Rendering

Basic scraperJSON

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']"  
    }  
  }  
}
```

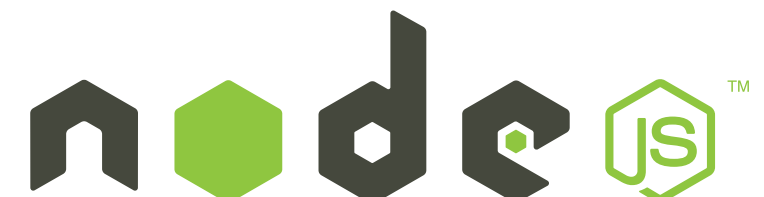


bibJSON output

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

thresher & quickscrape

- reference implementation of scraperJSON
- **thresher** is the scraping library
 - <http://github.com/ContentMine/thresher>
- **quickscrape** is the command-line tool
 - <http://github.com/ContentMine/quickscrape>
- Node.js, **MIT licensed**

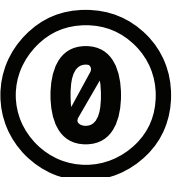


journal-scrapers

<http://github.com/ContentMine/journal-scrapers>

a self-testing collection of scraperJSON scrapers for academic journals

- PLOS
- MDPI
- PeerJ
- Wiley
- ScienceDirect
- Springer
- Taylor & Francis
- NPG, AAAS, RSC, ACS, ...





This repository Search

Explore Gist Blog Help



ContentMine / journal-scrapers

branch: master

journal-scrapers / scrapers / +

remove broken science direct scraper



Blahah authored 4 days ago

..

acs.json

acs must be headless

elife.json

elife scraper

generic_open.json

add renaming to all scrapers

mdpi.json

add fulltext xml to compatible scrapers

peerj.json

add fulltext xml to compatible scrapers

plos.json

add fulltext xml to compatible scrapers

pnas.json

Added PNAS scraper & test

rsc.json

RSC Scrapers