OASIS User's Manual

David Robinson

dgrtwo@princeton.edu

OASIS, or Optimized Annotation System for Insertion Sequences, is a system for automatic annotation of Insertion Sequence (IS) elements in prokaryotic genomes, based on NCBI Genbank files.

OASIS is a free open source package developed on a Mac platform and implemented in Python, available at `http://github.com/dgrtwo/OASIS`.

1 Installation

1.1 Requirements

- Python 2.4-2.7 (not 3.0)
  - Python can be obtained from `http://www.python.org/`.
- BioPython 1.46 or above
- NCBI BLAST, specifically the `blastall` and `formatdb` executables

1.2 Setup

- If necessary, install the prerequisites of Python and BioPython, and download NCBI BLAST
- In the file `src/OASIS/data/data.cfg`, set the lines:

  ```
  BLAST_EXE=/PATH/TO/BLAST/EXECUTABLE/HERE
  FORMAT_EXE=/PATH/TO/FORMATDB/EXECUTABLE/HERE
  ```

  to the full path to the blastall and formatdb executables on your computer, respectively.
- Install with the commands:

  ```
  python setup.py build
  sudo python setup.py install
  ```

2 Input

OASIS takes as input a microbial genome in Genbank format. The genome must already have annotated transposases, as do most NCBI publicly available prokaryotic genomes, for OASIS to find the insertions sequences. Sequenced microbial genomes in Genbank format are available at

```
ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria
```

OASIS can alternatively be given a folder of Genbank files, in which case it will output to a folder of annotations.

3 Output

OASIS outputs two files for each genome- a fasta file and a gene feature file. The gff file contains one line for each insertion sequence. An example line looks like:

```
AE009952.1 OASIS IS 776493 777934 . + . set_id "1"; family "IS3";
group "IS150"; IRL "TGCACTGAACCCCA"; IRR "TGCACTGAACCCCA";
```

Along with the usual gff data (chromosome, source, feature, start, end, score, direction, and strand), OASIS provides attributes that describe the IS:

- **set_id**- ISs within a genome are divided into sets, each containing nearly identical copies.
- **family**- The identified family of the insertion sequence, based on the *ISfinder* database of IS elements (see `http://www-is.biotoul.fr/`).
- **group**- The group of the insertion sequence (subclass of *ISfinder* family)
- **IRL**- The left (defined as upstream of the coding sequence) inverted repeat sequence, if one is found
- **IRR**- The right (defined as downstream of the coding sequence) inverted repeat sequence, if one is found

The fasta file contains the nucleotide sequence of each IS and one amino acid sequence for each ORF in each. For example:

```
>AE009952.1_776493_777934
TGCACTGAACCCCAGATCTTGGATCTTTTTATCCATGATCTGGAGGTTCGGTTCAATGAA
ACACCCTTTTTCAACCCGCCTAGCGGCGGTTCAACATTACCTTTCAGGAAAGGCCACTCT
GCGGGAAACCGCACGTCAATTCAGTGTTGGCAAATCCCCTCTTACGCGTTGGATCCGAGC
```

```
TTTTCGCCGTCAAGGTGAGGCTGGACTGGAGCACCATCTTTCCAGAACTTATACTCCAGA
...
>AE009952.1_776493_777934_ORF_1
MIWRFGSMKHPFSTRLAAVQHYLSGKATLRETARQFSVGKSPLTRWIRAFRRQGEAGLEH
HLSRTYTPEFRLCVVRYMMANRCSAADASAHFNIPNETIIQNWMKRYREGGKEALNPSKT
GPTMPKDKYEHDSKPFSEMTHAELEKELEYLRAENAYLKKRKALREEKALREQQKKPE
>AE009952.1_776493_777934_ORF_2
MARSTYYYHASKPDGVIDDYADAVKAIGALSRRHAQRYGYRRMTVALRKEGFTLNHKTVR
KLMNQHGLLSLIRRKKYRSYRADGGRASDNLLARNFTSEISGLKWCTDVTEFRVGAQKLY
LSVIQDLFNNEIISWHMSERAALILTCKTLEKALKVKGRKEGLMLHSDQGWHYRTPMWRS
MLVEAGIRQSMSRKGNCLDNAVMENFFSHLKAEMYHRKKYDSATVLKRDIVEYIHYYNTE
RISLKTGGMSPAEYRTQVEKQ
```

4 Algorithm

OASIS finds multiple-copy IS elements in each genome by identifying conserved regions surrounding transposase genes. First, groups of already-annotated transposase genes that could compose multiple copies of an IS were identified by length and sequence similarity. Those that fit a high similarity threshold were assumed to be in the same group.

The edges of each element were then found by finding a region of conservation around each group of transposases. The windows upstream and downstream of a group of transposases were compared. We are thus given a set of sequences and asked to find the edge at which conservation ends, allowing for mismatches within IS elements, in a more computationally efficient matter than a multiple alignment. The edge of conservation was found via the following maximum likelihood approach.

We assumed that upstream and downstream regions each consist of a contiguous conserved region (within the IS) and an unconserved region (outside the IS). For computational efficiency, it was assumed that there were no gap mutations in the conserved region. Consider the set of upstream or downstream n-windows, in a group that contains m transposases. Let k be the length of the conserved region- the length which the IS element extends past the transposase on this side.

The first k characters of the sequences each come from multinomial distributions that are identical across the m sequences and are biased heavily towards a single nucleotide (the true sequence of the IS element). Let c_j be the consensus nucleotide for the jth character of the true sequence, and let $x_{i,j}$ represent the jth character of the ith sequence.

$$\begin{array}{cccc}
x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\
x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\
 & \vdots & \vdots & \\
x_{m,1} & x_{m,2} & \cdots & x_{m,n}
\end{array}$$

The distribution of $x_{i,j}$ is assumed to be follows:

$$x_{i,j} \sim \begin{cases} \text{Multinom}(\mathbf{p_{c_j}}) & : i \leq k \\ \text{Multinom}(\mathbf{p_b}) & : i > k \end{cases}$$

Where $\mathbf{p_{c_j}}$ represents the distribution of a conserved nucleotide with consensus value c_j, and $\mathbf{p_b}$ represents a background distribution vector (the distribution of nucleotides in unconserved regions). $\mathbf{p_c}$ can be any of 4 vectors that are biased towards a consensus nucleotide, with some probability of error. Let c, ε be adjustable parameters specifying the probability of a consensus match or an error in a conserved region.

The value of k is then found that maximizes the likelihood of the set of sequences. The likelihood is found by a dynamic programming algorithm as follows. For each nucleotide $N \in \{A, G, C, T\}$, let N_i be the count of nucleotides N in $x_{1\ldots m,i}$. Also let M_i be the count of the most common nucleotide (the MLE estimate of the consensus nucleotide in a conserved region). The log-likelihood function is then

$$l(k) = \sum_{i=1}^{k} \log(\frac{m!}{A_i!C_i!T_i!G_i!} c^{M_i} \varepsilon^{m-M_i}) + \sum_{i=k+1}^{n} \log(\frac{m!}{A_i!C_i!T_i!G_i!} (\frac{1}{4})^m)$$

Our estimate of k, $\hat{k}$, is thus

$$\hat{k} = \underset{k}{\text{argmax}}\ l(k)$$

OASIS computes this through a dynamic programming algorithm.

The putative edge of the IS element is then marked as $\hat{k}$ nucleotides from the edge of the transposase. This process is repeated for the windows upstream and downstream of the repeated transposase, for each group of transposases in the genome.

Once the edge of conservation has been found, the edges were then checked for inverted repeats using a Smith-Waterman alignment between the regions surrounding either edge of an IS element. If inverted repeats are found that disagree with the putative edges, the edges are adjusted to a limited extent.

Single-copy IS elements were found separately by finding transposases that were not placed in multiple-copy groups, and checking for inverted repeats in the surrounding regions. A Smith-Waterman alignment of the upstream and reverse-complement of the downstream regions was used to recognize significant inverted repeats and thus possible edges, a method first developed by IScan.

Once groups of IS elements were identified, a sample IS element from the set is selected, and blastn (NCBI) was used against the genome sequence to identify missing and partial copies of the IS element.

The family and group of each IS element is identified using blastp on the identified IS element ORFs against the ISFinder database, and classifying based on the family and group of the best match.

5 Tutorial

What follows is an example use of OASIS for annotation of a prokaryotic genome.

5.1 Input

The example genome is the `AE009952` chromosome from Yersinia pestis KIM, a Gram-negative bacterium belong to the Enterobacteriaceae family. The chromosome is available to be downloaded from

```
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/
Yersinia_pestis_KIM_10_uid288/AE009952.gbk
```

5.2 Usage

To annotate the chromosome, execute the command

```
OASIS -g AE009952.gbk -o Yersinia_Pestis_Annotations
```

5.3 Output

Once the annotation is complete, there are two output files specifying the annotated IS elements in the genome. The gff file, `Yersinia_Pestis_Annotations.gff`, will have the format specified in the Output section above. The fasta file, `Yersinia_Pestis_Annotations.fasta`, will have the nucleotide sequence and any protein sequences of one example from each group.