Integrated gene prediction for eukaryotic and prokaryotic genome using EuGène

Erika Sallet, Jérôme Gouzy, Brice Roux, Delphine Capela, Laurent Sauviac, Claude Bruand, Pascal Gamas and Thomas Schiex

Journée PLUME Biologie
Terminology:
eukaryotic gene structure

Diagram showing the structure of a eukaryotic gene, including:
- DNA
- Pre-mRNA
- mRNA
- Protein

Key steps:
1. Transcription
2. Processing/SPLICING
3. Translation
Structural annotation of prokaryotic genomes

- Sequencing technology gives access to prokaryote transcriptome

- Need to reconsider prokaryotic genome annotation taking into account such data

- Already used in eukaryote gene prediction
RNA-seq

- Strong gene density in prokaryotic genomes
  - RNAseq reads not so informative?

Apollo view (Lewis SE et al Genome Biology 2002)
Oriented RNA-seq

→ Informative signal to detect transcription
EuGène

► Integrative modular eukaryotic gene finder

► Integrates different kinds of evidence such as:
  - Statistical properties of region
  - Known protein similarity,
  - EST (Expressed Sequence Tag)

► Extensible and generic software

► Artistic license

► Forge  https://mulcyber.toulouse.inra.fr/projects/eugene
EuGène

► Contributors

- Created by Thomas Schiex (INRA, Applied Mathematics and Informatics) in 1999
- 8 different developers
- Current EuGène team = 3 members

► Users

- Intended for bioinformaticians
- We are the main user of EuGène

► Genome annotation of *M. truncatula, S. lycopersicum, A. thaliana*, fungi (*Botrytis cinerea, ...*), *M. incognita, ...*
Eukaryotic gene

▶ = A sequence of “regions” separated by “sites”
EuGène eukaryotic automaton
How does EuGène work?

- Transforms a nucleic sequence into a graph
- Edits/weights it according to biological & statistical evidence
How does EuGène work?

- Transforms a nucleic sequence into a graph
- Edits/weights it according to biological & statistical evidence
How does EuGène work?

- Transforms a nucleic sequence into a graph
- Edits/weights it according to biological & statistical evidence

CCTACGTGCA
CATGACGTAGG
CCTACGTGCA
CATGACGTAGG
CCTACGTGCA
CATGACGTAGG
CCTACGTGCA
CATGACGTAGG

May 2012 – PLUME – Erika.Sallet@toulouse.inra.fr
How does EuGène work?

- Transforms a nucleic sequence into a graph
- Edits/weights it according to biological & statistical evidence
How does EuGène work?

- Transforms a nucleic sequence into a graph
- Edits/weights it according to biological & statistical evidence
How does EuGène work?

- Transforms a nucleic sequence into a graph
- Edits/weights it according to biological & statistical evidence
How does EuGène work?

- Transforms a nucleic sequence into a graph
- Edits/weights it according to biological & statistical evidence
- A gene structure = a path in the graph
- The prediction = an optimal path in the graph
From eukaryote to prokaryote
EuGène prokaryote automaton

- Prokaryotic genomes:
  - Protein genes on both strands (possibly overlapping)
  - Transcribed regions (operon)
  - RNA genes (possibly antisense)
Integrating evidence

► Translation starts
  ▪ Ribosome hybridation energetic model \(^{(1)}\)

► Transcription
  ▪ Based on the expression level of RNAseq data

► Transcription start
  ▪ Simple signal processing of RNAseq data

► Segmentation of transcription profiles G. Rigaill, S. Robin. An exact Algorithm for the Segmentation of NGS Profiles Using Compression

Application to *Sinorhizobium meliloti* Sm2011

**Sinorhizobium meliloti**
- Nitrogen fixing bacteria
- Symbiosis with legumes (ex: *M. truncatula*)
- 6.69 Mb (62% GC)

**EuGène « recipe » to annotate the genome**
- 19 RNAseq oriented libraries (48Gb)
  - Paired-end and single-end Solexa reads
  - 6 triplicates
  - 3 biological conditions
- Results of BlastX vs SwissProt and Sm1021
- RNAmer and tRNAscanSE results
- IMM (Interpolated Markov Model)
Application to *Sinorhizobium meliloti* Sm2011

Prediction strand +

UTRs → Operon → CDS

Lib 1 long inserts

Lib 2 long inserts

Lib1 short inserts

Lib2 short inserts

Prediction strand -

ncRNA
Overall

- Results of the EuGène prediction
  - 6,367 protein coding genes
    - 135 new (compared to Sm1021)
    - Covering 87.35% of the genome
  - 1,969 ncRNA genes (54 tRNA, 9 rRNA)
    - Covering 2.98% of the genome
    - Published ncRNA (Valverde 2008): 76/140 found
    - Antisense ncRNA candidate
  - 88.8% of the genome is annotated (86.5% previously)
Many thanks

► ANR Symbimics projects participants:
  - Pascal Gamas (coordinator)
  - Brice Roux, Claude Bruand, Delphine Capela, Laurent Sauviac, Nathalie Rhodde

► “EuGène team” members
  - Thomas Schiex and Jérôme Gouzy
