Predicting isoform transcripts: What does the comparison of known transcripts in human, mouse and dog tell us?
Nicolas Guillaudeux, Catherine Belleannée, Samuel Blanquart, Jean-Stéphane Varré

To cite this version:
Nicolas Guillaudeux, Catherine Belleannée, Samuel Blanquart, Jean-Stéphane Varré. Predicting isoform transcripts: What does the comparison of known transcripts in human, mouse and dog tell us?. JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques, Jul 2019, Nantes, France. 8, pp.1, 2019, 10.7490/f1000research.1117311.1 . hal-02267357

HAL Id: hal-02267357
https://inria.hal.science/hal-02267357
Submitted on 19 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Modelling gene structures using comparative genomics to predict isoforms [1]

Our structure of a target gene (fig.1a)
- based on functional sites of known transcripts:
  - a letter: (part of) a coding exon
  - using functional sites of known transcripts in an orthologous source gene (fig.1b):
    - to reveal new functional sites or coding exons on the target gene (fig.1c)
    - to predict new functional sites of the target gene, using predicted functional sites (fig.1d)
- Between two functional sites each node is a functional site of a gene
- We build two conservation graphs for each gene:
  - conservation of functional sites: between two functional sites
  - conservation of transcripts: between two transcripts
  - a shared in the 3 species (fig.3a), in 2 species (fig.3b), specific to a species (fig.3c)
  - conservation only graphs without “ambiguous” components (fig.3d) are shown.

Conservation of functional sites:
- each node is a functional site of a gene
- each edge represents an orthology relationship between two functional sites
- Genes with all functional sites shared (only triplets)
- At least one triplet
- At least one duplicate
- At least one singleton
- Nothing aligned (artefact)

Conservation of transcripts:
- each node is a transcript of a gene
- each edge represents an orthology relationship between two transcripts
- Genes with all transcripts shared (only triplets)
- At least one triplet
- At least one duplicate
- At least one singleton

Analyzing conservation over functional sites and transcripts

We build two conservation graphs for each gene:
- conserved functional site graph (fig.4) & conserved transcript graph (fig.5)
- a graph component shows orthology relationships between species
- classification of graph components (3-species case):
  - shared in the 3 species (fig.3a), in 2 species (fig.3b), specific to a species (fig.3c)
- only graphs without “ambiguous” components (fig.3d) are considered for analysis

Application: Predicting transcripts in human, mouse and dog

Data used:
- 2,167 orthologous genes & 18,109 known transcripts:
  - human and mouse: CCDS
  - dog: ENSEMBL

Pairwise comparisons [1] and merging results
- 6,861 new predicted transcripts (fig.2):
  - 1,540 at least one site shared in mouse and dog
  - 2,112 at least one site shared in human and dog
  - 3,209 at least one site specific in human
  - 2,051 at least one site specific in dog

Phylogenetic interpretation of gene structures

Most analyzed genes encountered structure divergence
- At least one site shared in human and mouse
- At least one site shared in human and dog
- At least one functional site specific in human
- At least one functional site specific in mouse

Some genes have the same gene structure but different isoforms
- 255 genes have all functional sites conserved over the 3 species but 120 express different transcripts.
  - Example:
    - Frameshifts
    - Alternative transcription
    - False negative

Figure 1. CREM gene: pairwise comparison between human (a) and mouse (b) reveals new functional sites or exons (red) on human gene (c). This lead to predict new expressible transcripts (d) in human.

Figure 2. Estimated transcriptomes: known and predicted transcripts obtained using pairwise comparisons.

Figure 3. Considered gene components: (a) triplet, (b) duplicate, (c) singleton. “Ambiguous” graph components (d) are not considered.

Figure 4. Functional site graph of CREM gene (left). Analysis of 1,663 considered functional site graphs. 255 graphs reveal genes having common gene structures (right).

Figure 5. Transcript graph of CREM gene (ambiguous) (left). Analysis of 986 considered transcript graphs. 135 graphs reveal genes having common potential transcriptomes (right).

Figure 6. Distribution of duplicates (left) and singletons (right) components in functional site graphs.