Annotation-independent Search for Synteny Anchors
Karl Kaether
Problem
Orthology Inference

• Mostly based on sequence similarity

• Complicated by (1) sequence divergence, (2) genomic rearrangements:
  
  (1) — genome comparisons which are sensitive enough to detect similarity over large phylogenetic distances become difficult/expensive
  
  (2) — duplications and losses lead to varying number of potential orthologs

• Conserved order of genetic elements (synteny) can help in such situations
Duplications as Major Driver of Evolution

A Current Example

- Protein family of gene responsible for particular fly phenotype underwent frequent duplication/loss events in phylogenetically restricted set of species.

- Other copies do not seem to convey function in D. mel.

- To determine where and possibly how function originated, phylogenetic tracking of functional copy is essential.
Problem
Orthology Inference

• Synteny information can help in such situations

• e.g. maximum parsimony algorithm to build gene tree with minimum amount of gene duplication/loss events

• especially with increasing amount of available genomes this is promising
Another Problem
Finding Good Synteny Anchors

• Usually based on genome annotations and multiple sequence alignments — often unavailable, of bad quality and expensive to compute

• Maybe one can find them irrespective of annotations
Annotation-independent Search for Synteny Anchors

Idea

What are good synteny anchors?

• Conserved across large phylogenetic distances
• Low copy number variation
  —> type of genetic element does not matter
  —> all sequences within a genome which are fairly unique are potential anchors
Annotation-independent Search for Synteny Anchors

Problem Statement

Given: Genome

Wanted: All subsequences which are at least X different from all other subsequences within a genome

X ?? - similarity score like edit distance/hamming distance

Then: If those subsequences are similar above X (+ some tolerance) to a subsequence of another genome, they make a good synteny anchor pair

→ Goal is to create sets of potential anchors per genome and map matching pairs (groups)
Annotation-independent Search for Synteny Anchors

First Attempts

1. Count k-mers

2. Sum up counts within a window

3. Take best x %
Annotation-independent Search for Synteny Anchors

First Attempts

3. Blast against own Genome
   - iteratively with more sensitive word size parameter and considering fragmented hits (chaining of hits)

4. Concatenate consecutive hits

5. Blast and chain again

6. Put into categories according to score value

7. Pairwise blast + chaining of anchor sets of different genomes —> mapping of matching anchors
Moving Parts of Pipeline

• k and allowed error of kmer counting
• Window size, overlap, operation of kmer count aggregation
• Blast parameters - word size, gapped/ungapped
• Categories and respective score values
• Many other open technicalities, e.g. further concatenation / final definition of anchor regions missing as of now
Some Preliminary Results

- 25 Drosophila species
- Pretty good genomes
- Annotations
- High-confidence species tree

$\Rightarrow$ approach can be evaluated well
Some Preliminary Results

How many potential anchors does one get?

- 13-mers, no errors
- 300 nt windows
- overlapping at mid-points
- blast word size of 13 and 9
- categories shown
- sizes between 300 and ~ 6000
How many matches do they have in other genomes?
Recall of Annotated Elements  
*Define Recall

Recall of Annotated Genetic Elements (D.mel - all anchors)

gene
- mRNA
- antisense_RNA

lnc_RNA
- CDS
- pseudogene

exon
- miRNA
- snoRNA
Recall of Annotated Genetic Elements (D.mel - anchors with >= 15 matches)

Recall of Annotated Genetic Elements (D.mel - anchors with >= 20 matches)
Preliminary Results - How to evaluate anchors and their matches

• Reminder: wanted are unique-ish sequences with one-to-one correspondence in other genomes

• evaluation approach:
  - BUSCOs: set of single-copy gene models curated for different taxons
Anchor Evaluation - BUSCOs

Recall of BUSCOs (all anchors)
- willistoni
- pseudoobscurapersimilis
- miranda
- ananassae
- suzuki
- biarmipes
- takahashii
- eugracilis
- erecta
- yakuba
- melanogaster
- sechellia
- simulans
- ficusphila
- busckii
- grimshawi
- arizonae
- mojavensis
- navojoa
- viriliis

Recall of BUSCOs (anchor >= 15 matches incl. multi)
- willistoni
- pseudoobscurapersimilis
- miranda
- ananassae
- suzuki
- biarmipes
- takahashii
- eugracilis
- erecta
- yakuba
- melanogaster
- sechellia
- simulans
- ficusphila
- busckii
- grimshawi
- arizonae
- mojavensis
- navojoa
- viriliis

Recall of BUSCOs (>= 15 matches anchors)
- willistoni
- pseudoobscurapersimilis
- miranda
- ananassae
- suzuki
- biarmipes
- takahashii
- eugracilis
- erecta
- yakuba
- melanogaster
- sechellia
- simulans
- ficusphila
- busckii
- grimshawi
- arizonae
- mojavensis
- navojoa
- viriliis

Caption
Preliminary Results - Duplicated BUSCOs

Recall of Duplicated BUSCOs (all anchors)
miranda

Recall of Duplicated BUSCOs (>= 15 matches anchors)
miranda

Caption

Caption
BUSCOs as measure for quality of matching

Given there is a BUSCO in other genome:

True positives - Match corresponds to other BUSCO
• 88 %

False negatives - No match or wrong match (2 %) despite BUSCO present
• 12 %

Given there is no BUSCO in other genome

False Positives - Match although no BUSCO detected
• 28 %

True negatives - No Match
• 72 %

Recall: 88 %

Precision: 76 %
Sampling of False Positives

Optimal local alignments (match 2, mismatch -1, linear gaps -2):

• Anchor genome x +- 500 nt (with match in genome y) - identified BUSCO genome y:
  - 1345 score points

• Anchor genome x - match region where no BUSCO identified genome y:
  - 1075 score points

• Anchor genome x - random other BUSCO genome y:
  - 638 score points
Annotation-independent Search for Synteny Anchors

More Distant Genomes
BUSCOs as measure for quality of matching

Given there is a BUSCO in other genome:

True positives - Match corresponds to other BUSCO
  • 57 %

False negatives - No match or wrong match (1 - 2 %) despite BUSCO present
  • 43 %

Given there is no BUSCO in other genome

False Positives - Match although no BUSCO detected
  • 22 %

True negatives - No Match
  • 78 %

Recall:  57 %

Precision:  72 %
Sampling of False Positives

Optimal local alignments (match 2, mismatch -1, linear gaps -2):

- Anchor genome x +- 500 nt (with match in genome y) - identified BUSCO genome y:
  - 1004 score points
- Anchor genome x - match region where no BUSCO identified genome y:
  - 903 score points
- Anchor genome x - random other BUSCO genome y:
  - 625 score points
Further Steps

- Systematic Search for good parameters of pipeline
- Qualitative improvements (e.g. local alignments, extension of anchors,…)
- Application to example from Heidelberg group
Additional Slides

• Evaluation with OrthoFinder

• OrthoFinder takes proteomes and clusters proteins into orthogroups
OrthoFinder recall of single-copy proteins with some tolerance

Recall of Single-copy Proteins Identified by Orthofinder

* willistoni pseudoobscuraspersimilis
* miranda
* ananassae

bipectinata
* kikkawai
* rhopaloa
* elegans
* suzukii

biarmipes
takahashii
eugracilis
* erecta
* yakuba

melanogaster
* sechellia
* simulans
* ficusphila
* busckii

grimshawi
* arizonae
* mojavensis
* navojoa
* virilis
OrthoFinder recall of single-copy proteins with some tolerance

Recall of Single-copy Proteins Identified by Orthofinder
willistoni  pseudoobscurapersimilis  miranda  ananassae
bipectinata  kikkawai  rhopaloa  elegans  suzukii
biarmipes  takahashii  eugracilis  erecta  yakuba
melanogaster  sechella  simulans  ficusphila  busckii
grimshawi  arizonae  mojavensis  navojoa  virilis

Caption
Assessment of Anchor Matches

- Relative rate of correct identification for all matches of anchors in single-copy proteins
- Aggregated over all species (no outlier)
- Corresponds to a recall of ~ 67%
Assessment of Anchor Matches

- With tolerance
- Corresponds to a recall of ~ 80%
- Rudimentary re-evaluation showed that a considerable proportion of missing correspondence is due to wrong mapping of proteins to genome