

Algorithms for the Analysis of Expressed Sequence Tags

Thesis defense for the degree of *dr.scient.*

Ketil Malde

March 14, 2005

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

Assembly

What now?

An Integrated Tool

The Lost Genes

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

Assembly

What now?

An Integrated Tool

The Lost Genes

What is a gene?

- ▶ Proteins perform functions
- ▶ DNA stores their blueprints

What is a gene?

- ▶ Proteins perform functions
- ▶ DNA stores their blueprints

Simplified definition:

Gene (n): A region of the DNA that encodes for one particular protein.

What is a gene?

The central dogma:

- ▶ Proteins perform functions
- ▶ DNA stores their blueprints
- ▶ mRNA provides a copy for their construction

It's all sequences

- ▶ DNA is sequences of A C G and T
- ▶ RNA is sequences of A C G and U
- ▶ Proteins are sequences over 20 amino acids

It's all sequences

- ▶ DNA is sequences of A C G and T
- ▶ RNA is sequences of A C G and U
- ▶ Proteins are sequences over 20 amino acids

→ Practical to work with, both for:

- ▶ the biomechanical machinery of the cell
- ▶ the electronic machinery of the computer

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

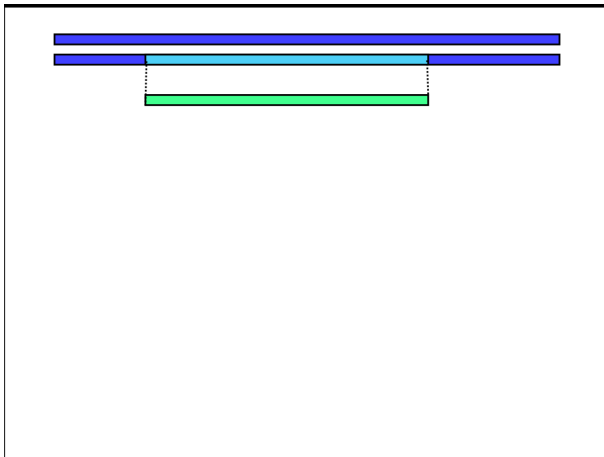
Assembly

What now?

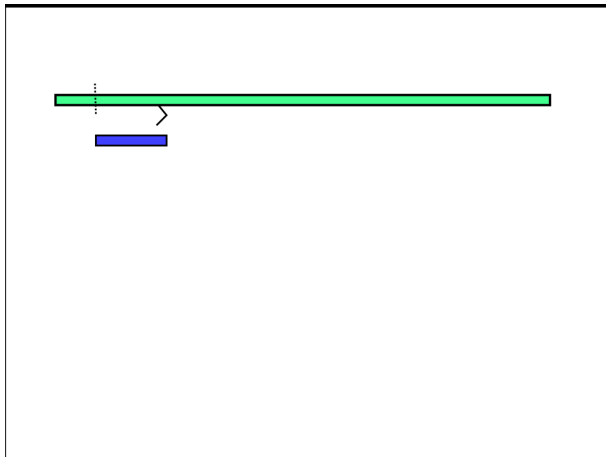
An Integrated Tool

The Lost Genes

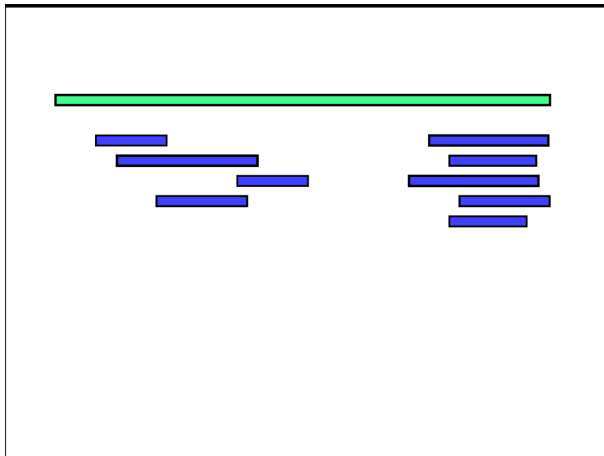
What are ESTs?



What are ESTs?



What are ESTs?



Why are ESTs important?

ESTs are:

- ▶ Cheap to manufacture on a large scale

And provide information about:

- ▶ Gene structure
- ▶ Gene regulation and modification

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

Assembly

What now?

An Integrated Tool

The Lost Genes

Clustering

Task: *Group ESTs that originate from the same gene*

Clustering

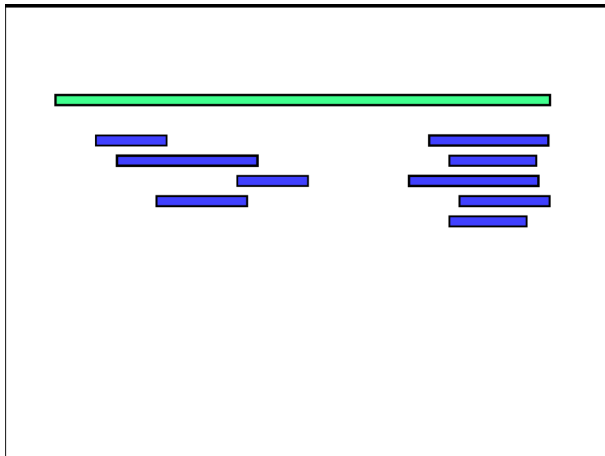
Task: *Group ESTs that originate from the same gene*
— in practice, that are sufficiently similar

Clustering

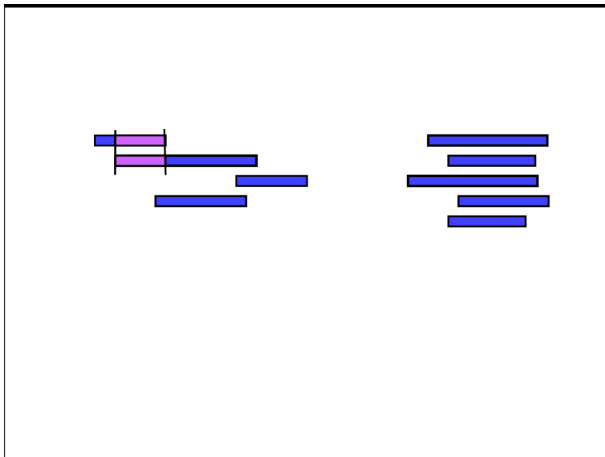
Task: *Group ESTs that originate from the same gene*
— in practice, that are sufficiently similar

Straightforward approach:
check all pairs of sequences

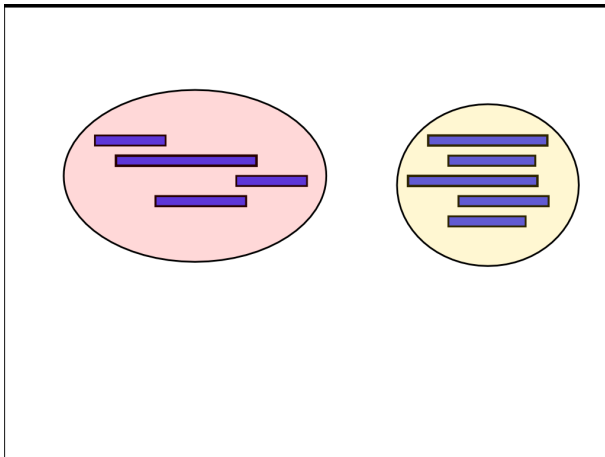
EST Clustering



EST Clustering



EST Clustering



Clustering

Task: *Group ESTs that originate from the same gene*
— in practice, that are sufficiently similar

Straightforward approach:
check all pairs of sequences

For n sequences, there are $(n^2/2)$ pairs.

The Suffix Array

The Suffix Array data structure
is a sorted array of all *suffixes*:

- ▶ Efficient and compact data structure
- ▶ Indexes the data set
- ▶ Makes it easy (and fast) to identify all matches

Idea: use a suffix array to find the matches

The Suffix Array

The Suffix Array data structure
is a sorted array of all *suffixes*:

- ▶ Efficient and compact data structure
- ▶ Indexes the data set
- ▶ Makes it easy (and fast) to identify all matches

Idea: use a suffix array to find the matches

Tool: *xsact* (Paper II)

The Suffix Array

The Suffix Array data structure
is a sorted array of all *suffixes*:

- ▶ Efficient and compact data structure
- ▶ Indexes the data set
- ▶ Makes it easy (and fast) to identify all matches

Idea: use a suffix array to find the matches

Tool: *xsact* (Paper II)

PaCE, quasar

Genome-based clustering

Previously, we only used the ESTs themselves.

But: we now know the genome for many organisms.

Genome-based clustering

Previously, we only used the ESTs themselves.

But: we now know the genome for many organisms.

Idea:

Find the position of the EST in the genome
then cluster based on position

Genome-based clustering

Previously, we only used the ESTs themselves.

But: we now know the genome for many organisms.

Idea:

Find the position of the EST in the genome
then cluster based on position

Useful to compare the *ESTs-only* tools
(Paper IV)

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

Assembly

What now?

An Integrated Tool

The Lost Genes

The masking process

Task: *Remove noise and troublesome parts of the ESTs*

The masking process

Task: *Remove noise and troublesome parts of the ESTs*

- ▶ Quality clipping
- ▶ Vector masking
- ▶ Repeat masking

The masking process

Task: *Remove noise and troublesome parts of the ESTs*

- ▶ Quality clipping
- ▶ Vector masking
- ▶ Repeat masking

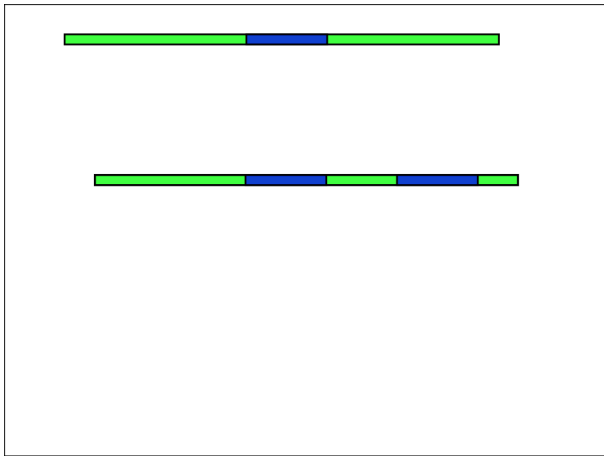
The masking process

Task: *Remove noise and troublesome parts of the ESTs*

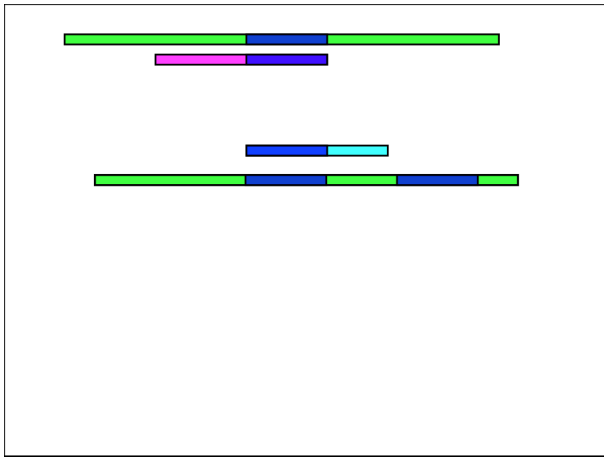
- ▶ Quality clipping
- ▶ Vector masking
- ▶ Repeat masking

(our) definition: a repeat is a region of a gene that match one or more other genes.

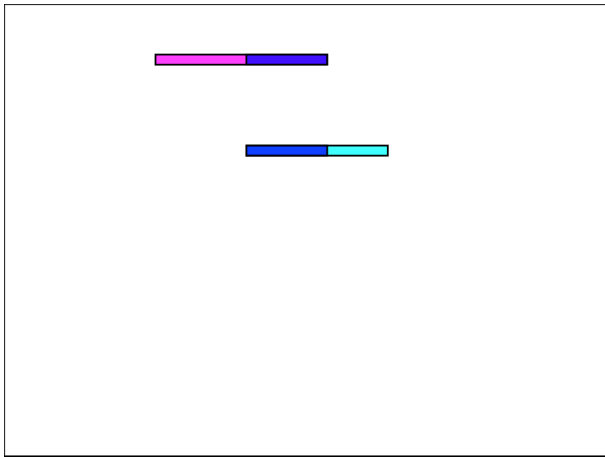
Repeats



Repeats

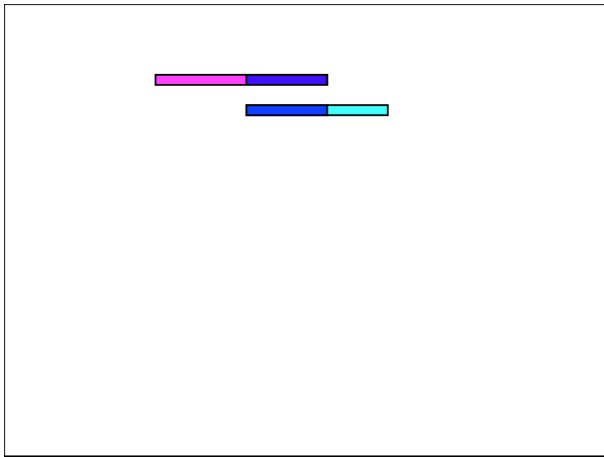


Repeats

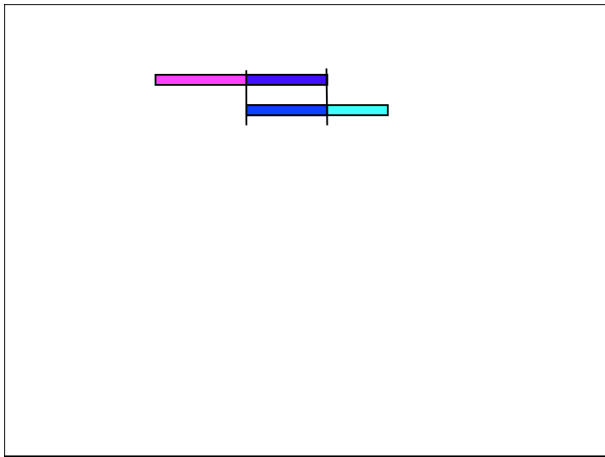




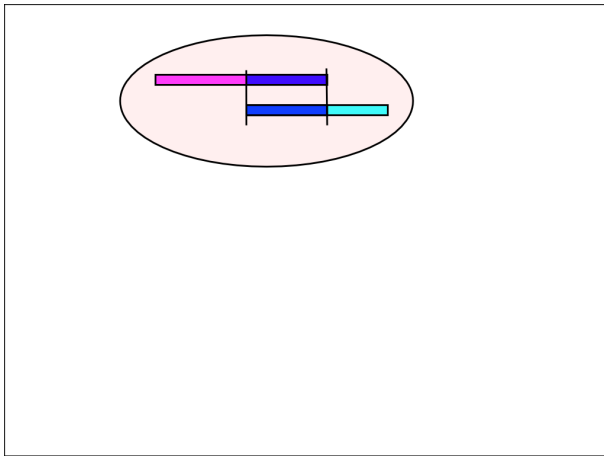
Repeats



Repeats



Repeats



Masking with a library

The Standard method: Library based repeat masking

Library construction:

- ▶ Search for repeats in the genome
- ▶ Collect all repeated sequences

Masking with a library

The Standard method: Library based repeat masking

Library construction:

- ▶ Search for repeats in the genome
- ▶ Collect all repeated sequences

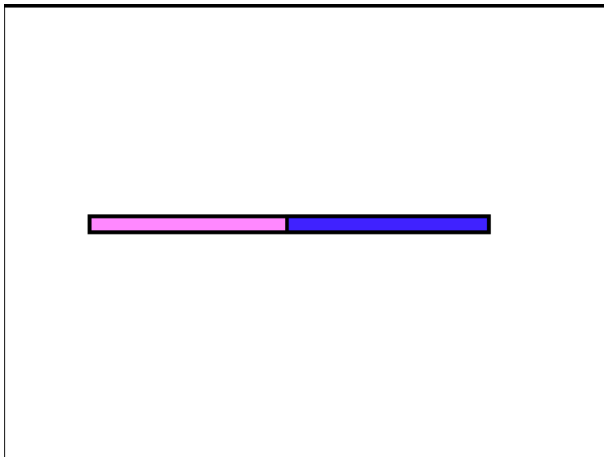
Masking against the library:

- ▶ Search for these sequences in the EST data
- ▶ And cross them out..

Disadvantages

- ▶ requires a good library
(which requires a well-known organism)
- ▶ slow (with large library)
- ▶ many repeats are outside the genes
- ▶ repeats depend on comparison method

Library-less masking

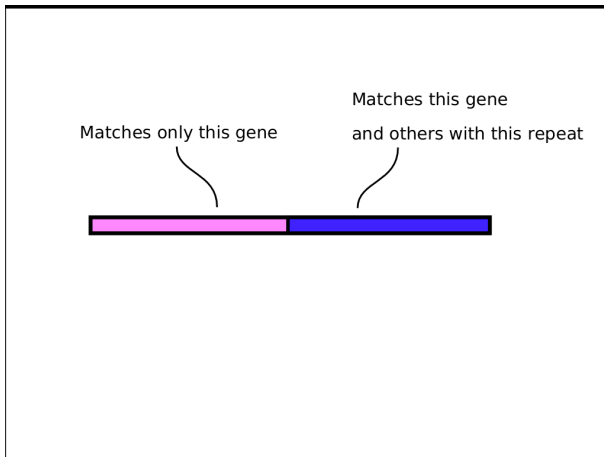


Library-less masking

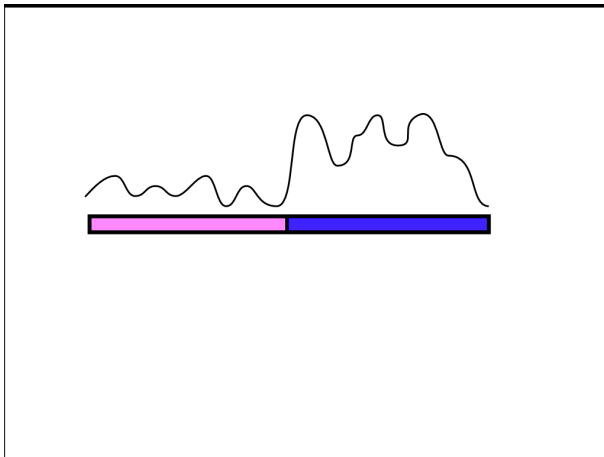
Matches only this gene



Library-less masking



Library-less masking



Library-less masking

Tool: *Repeatbeater* (Paper I)

- ▶ Works well in practice
- ▶ Detects different repeats
- ▶ More exploration required

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

Assembly

What now?

An Integrated Tool

The Lost Genes

Assembly

Task: Align the ESTs in a cluster to reconstruct the original mRNA sequence

Assembly

Task: *Align the ESTs in a cluster to reconstruct the original mRNA sequence*

Commonly used tools: Phrap, CAP3

Assembly

Task: *Align the ESTs in a cluster to reconstruct the original mRNA sequence*

Commonly used tools: Phrap, CAP3

Typical algorithm:

1. Identify overlaps
2. Find the optimal set of overlaps

⇒ Hamiltonian graph traversal

Also: not always a single correct assembly.

Word-based assembly

Alternative approach:

- ▶ break the data into words
- ▶ traverse the words in a consistent manner

Word-based assembly

Alternative approach:

- ▶ break the data into words
- ▶ traverse the words in a consistent manner

Tool: *xtract* (Paper V)

- ▶ Fast (proportional to output sequences)
- ▶ Better results
- ▶ but difficult to tune right

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

Assembly

What now?

An Integrated Tool

The Lost Genes

An Integrated Tool

Write a software tool that incorporates:

- ▶ Masking
- ▶ Clustering
- ▶ Assembly

Based on the same efficient indexing data structure.

An Integrated Tool

Write a software tool that incorporates:

- ▶ Masking
- ▶ Clustering
- ▶ Assembly

Based on the same efficient indexing data structure.

Apply to “new” organisms (salmon, cod,...)

Background

Genes and stuff

Expressed Sequence Tags

Contributions

Clustering

Masking

Assembly

What now?

An Integrated Tool

The Lost Genes

Identifying Genes in the Genome

20,000 protein coding genes vs.
180,000 EST clusters
→ junk? → non-coding genes?

Identifying Genes in the Genome

20,000 protein coding genes vs.
180,000 EST clusters

→ junk? → non-coding genes?

When matching ESTs against the genome,
15-25% do not match anywhere.

→ junk? → genes in heterochromatin? → other mechanisms?