

A Comparative Evaluation of XML Difference Algorithms with Genomic Data

Cornelia Hedeler (Connie) and Norman W. Paton
School of Computer Science, The University of Manchester, UK

Overview

- Motivation
- Problem description
- XML difference algorithms
- Experiments involving controlled modifications
- Evaluation involving real genomic data
- Conclusions

Motivation – Genome sequence data

- Genome sequence data for increasing number of genomes
- Undergoes regular re-annotation (and re-assembly)
- Annotation is done automatically using computational analysis
- Very little information on the changes between releases
- But previous versions are available (Ensembl, EMBL)

Number of species sequenced
(source: NCBI)

Organism group	Number of genomes
Viruses	1952
Eukaryota	1565
Bacteria	956

Motivation – From a biologist's perspective

- Cannot find favourite (group) of gene(s):

-Panic!

What happened to my gene(s)?

– Compare files manually

— Use available websites, tools

```

gene      23584..23925
          /gene=YCL058W-A
mRNA      23584..23925
          /gene="YCL058W-A"
          /note="transcript_id=YCL058W-A"
          /db_xref="SGD:YCL058W-A"
          /db_xref="Uniprot/
          /db_xref="RefSeq_p
          /db_xref="EntrezGe
          /db_xref="EMBL:X59
          /db_xref="AFFY_YG
          /translation="MGKC
          SSETMIPTLLREASTQEP
          KQITMISQFST"
gene      complement(24032..24325)
mRNA      /gene=YCL057C-A
          complement(24032..24325)
          /gene="YCL057C-A"
          /note="transcript_id=YCL057C-A"
          /db_xref="SGD:YCL057C-A"
          /db_xref="Uniprot/SWISSPROT:YCF
          /db_xref="RefSeq_peptide:NP_065
          /db_xref="EntrezGene:850300"
          /db_xref="EMBL:X59720"
          /db_xref="GO:GO:0016020"
          /db_xref="GO:GO:0016021"
          /db_xref="protein_id:CAC42954.1
          /db_xref="AFFY_YG_S98:6759_at"
          /db_xref="AFFY_YG_S98:6916_at"
          /db_xref="AFFY_YG_S98:6919_at"
          /db_xref="AFFY_Yeast_2:1770503
          /db_xref="AFFY_Yeast_2:1771324
          /db_xref="AFFY_Yeast_2:1774225
          /translation="MSEQAQTOQPAKSTPS
          TAMGFGVGVFTSVLFFKRRAFPVWLIGIGFC
gene      24768..26906
          /gene=YCL057W
mRNA      24768..26906
          /gene="YCL057W"
          /note="transcript_id=YCL057W"
CDS       24768..26906
          /gene="YCL057W"

```



gene
misc_RNA

oup) of gene(s):

y gene(s)? Deleted?

ally Moved? Ren Merged?

es, tools.

T: YCHO
P_0098
9"

950.1"

LO74W"
 "AFFY_YG_S98:3488_at"
 "AFFY_YG_S98:3489_f_at"
 "AFFY_YG_S98:3490_f_at"
 "AFFY_YG_S98:6956_s_at"
 eudogene"
 anscript_id=YCL074W
 t(6479..8326)
 073C
 g"YCL074W"
 characterized membrane protein [YCL074W]
 niprot/SWISSPROT;Acc:[P25961]"
 t(6479..8326)
 LO74W"
 "YCL074W"



T:YCHO_YEAST"
P_009863.2"
9"

Motivation – Information on changes (1)

Saccharomyces Genome Database (SGD)

2006-01-11	YCL058W-A YCL058W-A was originally added per Brachat et al., based on conservation with <i>Ashbya gossypii</i>. Kellis et al. also predicted a protein in the same frame, but with a 19 amino acid N-terminal deletion relative to the ORF predicted by Brachat et al. SGD has changed the annotation to that predicted by Kellis et al., based on conservation of that start site in <i>sensu stricto</i> strains of <i>Saccharomyces</i>. Kellis M, et al. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. <i>Nature</i> 423(6937):241-54 Brachat S, et al. (2003) Reinvestigation of the <i>Saccharomyces cerevisiae</i> genome annotation by comparison to the genome of a related fungus: <i>Ashbya gossypii</i>. <i>Genome Biol</i> 4(7):R45
2005-11-14	YCL048W-A Based on genome sequence comparisons among six <i>Saccharomyces</i> species, Cliften et al. 2003 suggested that a new ORF, YCL048W-A, be added to the <i>S. cerevisiae</i> genome annotation. This suggestion was later confirmed by Zhang and Dietrich 2005 using high-throughput identification of transcription start sites. Cliften P, et al. (2003) Finding functional features in <i>Saccharomyces</i> genomes by phylogenetic footprinting. <i>Science</i> 301(5629):71-6 Zhang Z and Dietrich FS (2005) Mapping of transcription start sites in <i>Saccharomyces cerevisiae</i> using 5' SAGE. <i>Nucleic Acids Res</i> 33(9):2838-51
1998-07-24	YCR055C, YCR057C, YCR058C ORFs YCR055C and YCR058C were merged with existing ORF YCR057C. The coding region of YCR057C had been 1320 nt long, but is now 2772 nt in length.
1998-07-24	YCR029C ORF YCR029C has been deleted from the genome annotation due to sequence correction.

<http://www.yeastgenome.org/cache/genomeSnapshot.html#ChrSeqAnnotUpdates>

Motivation – Information on changes (2)

Ensembl

What's New in Ensembl 49

Mus musculus News

▶ Mouse transcripts updated

Several mouse transcripts that were "culled" in Ensembl 47 have been reinstated. Also, a bug in the pipeline that caused ditags to be ignored has been fixed, which has resulted in some shorter UTRs.

▶ Minor updates

A number of minor changes have been made to the core databases.

[Read more...](#)

▶ Variation updates

New variation databases have been released for human, mouse, chicken and zebrafish (dbSNP 128) and rat (dbSNP 126). Transcript variation has been rerun for dog, platypus, chimp, and tetraodon due to gene set changes.

▶ Vega updates

Ensembl Vega has a new Mouse database, and Vega Human no longer shown external Vega annotation.

For the convenience of our users, we have created a web-based 'wizard' which enables you to [convert a GFF file](#) from Mouse assembly m36 to m37.

Fungal Genome Initiative

```
OLD_LOCUS,NEW_LOCUS,UPDATE_TYPE,UPDATE_DETAIL
NCU000001.2,NCU000001.3,UNCHANGED,
NCU000002.2,,DELETE,
NCU000003.2,NCU000003.3,UNCHANGED,
NCU000004.2,NCU000004.3,UNCHANGED,
NCU000005.2,NCU000005.3,CHANGE,UTR_CHANGED
NCU000006.2,NCU000006.3,CHANGE,EVIDENCE_CHANGED
NCU000007.2,NCU000007.3,CHANGE,EVIDENCE_CHANGED
NCU00165.2,NCU11172.3,SPLIT,
NCU00165.2,NCU11173.3,SPLIT,
NCU00166.2,NCU00166.3,CHANGE,EVIDENCE_CHANGED
NCU11163.2,,DELETE,
NCU11165.2,,DELETE,
NCU11166.2,NCU11166.3,UNCHANGED,
NCU11167.2,NCU11167.3,UNCHANGED,
,NCU11174.3,ADD,
,NCU11182.3,ADD,
NCU03818.2,NCU03818.3,UNCHANGED,
NCU03819.2,NCU03819.3,CHANGE,SPLICING_CHANGED
NCU03820.2,NCU11262.3,MERGE,
NCU03821.2,NCU11262.3,MERGE,
```

Neurospora crassa – changes between version 2 and version 3

<http://www.broad.mit.edu/annotation/fgi/>

<http://www.ensembl.org>

Motivation – Tools (1)

Converter – Ensembl

Convert your data

Assembly Converter enables you to convert your data from Mouse assembly m36 to m37. Please note that this facility is not currently compatible with any other species or assembly.

Supported formats: Only [GFF](#) files with chromosomal coordinates are currently supported. Data in this format can be exported from the [August 2007 Ensembl archive](#)

Paste file content

or upload file

or use file URL

[Upload data >](#)

Motivation – Tools (2)

Compare versions - EMBL

Accession Number or Sequence Version: ☐ case sensitive

Snapshot at day-month-year (e.g. 30-11-1998 or 30-NOV-1998)

Differences for X59720 24-AUG-2004 / 29-MAY-2008 [Back to List](#)

Lines unchanged **Lines removed** **Lines inserted**

FT	CDS	complement(11503..12285)
FT		/product="hypothetical protein"
FT		/db_xref="SGD:S0000573"
FT		/db_xref="UniProt/Swiss-Prot:P25593"
FT		/note="ORF YCL068c"
FT		/db_xref="GOA:P25593"
FT		/db_xref="InterPro:IPR008937"
FT		/db_xref="SGD:S000000573"
FT		/db_xref="UniProtKB/Swiss-Prot:P25593"
FT		/product="hypothetical protein"
FT		/protein_id="CAC42951.1"
FT		/translation="MFVLIDNVLAYLLEQDDLFTARFAIQGQIVSRRVKNKIHSNITD
FT		VLLQQFISHTLPYNDNIVPKKILDSMRTAVRQLLEATACVSRECPLVKRSQDIKRARKR
FT		LLSDWYRLGADANMDAVLLVNSAWRFLAVWRPFVNSIQHATQELYQNIHAYLLHGNVN
FT		IQRVTALLQLVMGQDDLLFSMDDVLQEVFRIQLYLNKMLPHNSHKWQKPSPFDSANLLL
FT		NFRDWTDDNALLQELLLSYPTINKNKHKNHSVPRLIQV"

Motivation – Genomic data

Representation

- Flat file formats
 - EMBL, Genbank: information rich, well established
 - GFF (General Feature Format): less information than EMBL or Genbank flat files
- XML
 - **EMBL**
 - INSD, based on Genbank flat file format
 - BIODAS, based on GFF flat file format

Overview

- Motivation
- **Problem description**
- XML difference algorithms
- Experiments involving controlled modifications
- Evaluation involving real genomic data
- Conclusions

Problem description – EMBL

```
gene      335..649
          /gene=YAL069W
          /locus_tag="O13588_YEAST
          /note="Yal069wp (YAL069W)"
mRNA      335..649
          /gene="YAL069W
          /note="transcript_id=YAL069W"
CDS       335..649
          /gene="YAL069W"
          /protein_id="YAL069W"
          /note="transcript_id=YAL069W"
          /db_xref="SGD:YAL069W"
          /db_xref="EMBL:U73805
          /translation="MIVNNTHVLT..."

...

exon      335..64
          /note="exon_id=YAL069W.1«

...

CCACACCACA CCCACACACC CACACACCAC
ACCACACACC ACACCACACC ...
```

Problem description – EMBL XML

```
<feature name="gene"> ← gene 335..649
  <qualifier name="gene">
    YAL069W ← /gene=YAL069W
  </qualifier>
  <qualifier name="locus_tag">
    O13588_YEAST ← /locus_tag="O13588_YEAST"
  </qualifier>
  <qualifier name="note">
    Yal069wp (YAL069W) ← /note="Yal069wp (YAL069W)"
  </qualifier>
  <location type="single" complement="false">
    <locationElement complement="false" type="range">
      <basePosition type="simple">
        335
      </basePosition>
      <basePosition type="simple">
        649
      </basePosition>
    </locationElement>
  </location>
</feature>
```

Problem description – EMBL XML

```
<feature name="mRNA">← mRNA 335..649
  <qualifier name="gene">
    YAL069W ← /gene="YAL069W"
  </qualifier>
  <qualifier name="note">
    transcript_id=YAL069W ← /note="transcript_id=YAL069W"
  </qualifier>
  location
</feature>
<feature name="CDS"> ← CDS 335..649
  <dbreference db="EMBL" primary="U73805" /> ← /db_xref="EMBL:U73805"
  <dbreference db="SGD" primary="YAL069W" /> ← /db_xref="SGD:YAL069W"
  <qualifier name="gene">YAL069W </qualifier> ← /gene="YAL069W"
  <qualifier name="protein_id">YAL069W </qualifier> ← /protein_id="YAL069W"
  <qualifier name="note">
    transcript_id=YAL069W ← /note="transcript_id=YAL069W"
  </qualifier>
  <qualifier name="translation">
    MIVNNTHVLT... ← /translation="MIVNNTHVLT..."
  </qualifier>
  location
</feature>
```

Problem description - Changes

- Update sequence
- Identification of a new gene
- Removal of previously predicted gene
- Merging of two neighbouring genes into one gene
- Splitting of a gene into two neighbouring genes
- ...

```

<feature name="gene">
  <qualifier name="gene">YAL069W </qualifier>
  <qualifier name="locus_tag">
    O13588_YEAST
  </qualifier>
  <qualifier name="note">
    Yal069wp (YAL069W)
  </qualifier>
  <location type="single" complement="false">
    <locationElement complement="false"
      type="range">
        <basePosition type="simple">
          335
        </basePosition>
        <basePosition type="simple">
          649
        </basePosition>
      </locationElement>
    </location>
  </feature>
  <feature name="mRNA">
    <qualifier name="gene">YAL069W</qualifier>
    <qualifier name="note">
      transcript_id=YAL069W
    </qualifier>
    location
  </feature>

```

```

<feature name="CDS">
  <dbreference db="EMBL" primary="U73805"/>
  <dbreference db="SGD" primary="YAL069W"/>
  <qualifier name="gene">YAL069W</qualifier>
  <qualifier name="protein_id">
    YAL069W
  </qualifier>
  <qualifier name="note">
    transcript_id=YAL069W
  </qualifier>
  <qualifier name="translation">
    MIVNNTHVLTL...
  </qualifier>
  location
</feature>
<sequence length="987" type="DNA"
version="0.0">
  gtccagttaaggcctatc...
</sequence>
...
<feature name="exon">
  <qualifier name="note">
    exon_id=YAL069W.1
  </qualifier>
  location
</feature>

```

Overview

- Motivation
- Problem description
- **XML difference algorithms**
- Experiments involving controlled modifications
- Evaluation involving real genomic data
- Conclusions

XML difference algorithms

Summary of key properties

Algorithm	Ordered/ Unordered tree	Changes detected
X-Diff	Unordered	Insert, Delete of leaf nodes; Update of values of text- and attribute nodes
JXyDiff (XyDiff)	Ordered	Insert, Delete, Move of leaf nodes; Update of values of text- or attribute nodes
3-Way Merge and Diff (3DM)	Ordered	Insert, Delete, Move, Copy of leaf nodes; Update of values of text- or attribute nodes

XML difference algorithms

X-Diff

- Pre-processing:
 - Parse, calculate XHash
- Matching:
 - Reduce search space
 - Bottom-up: Match nodes of same type and with matching ancestor names and calculate edit distance
 - Top-down: Create minimum-cost matching using matches and edit distances
 - One-to-one matches
 - Match child nodes of parents that are matched
- Edit script

XML difference algorithms

JXyDiff

- Pre-processing:
 - Bottom-up: Calculate hash values and weight
 - Place sub-trees into priority queue ordered by weight
- Matching:
 - Starting with heaviest sub-tree, match sub-trees
 - If no match found, add sub-trees rooted in child nodes into priority queue
 - Propagate matches further to unmatched nodes with same labels (up - controlled by weight of sub-tree)
- Edit script

XML difference algorithms

3DM

- Matching:
 - For each node in base document
 - Find exact or close matching nodes in updated document (using q-gram string distance measure)
 - For pairs of matched nodes
 - Match sub-trees by depth-first traversal as long as child nodes are matched
- Post-processing:
 - Propagate matches
- Edit script

Overview

- Motivation
- Problem description
- XML difference algorithms
- **Experiments involving controlled modifications**
- Evaluation involving real genomic data
- Conclusions

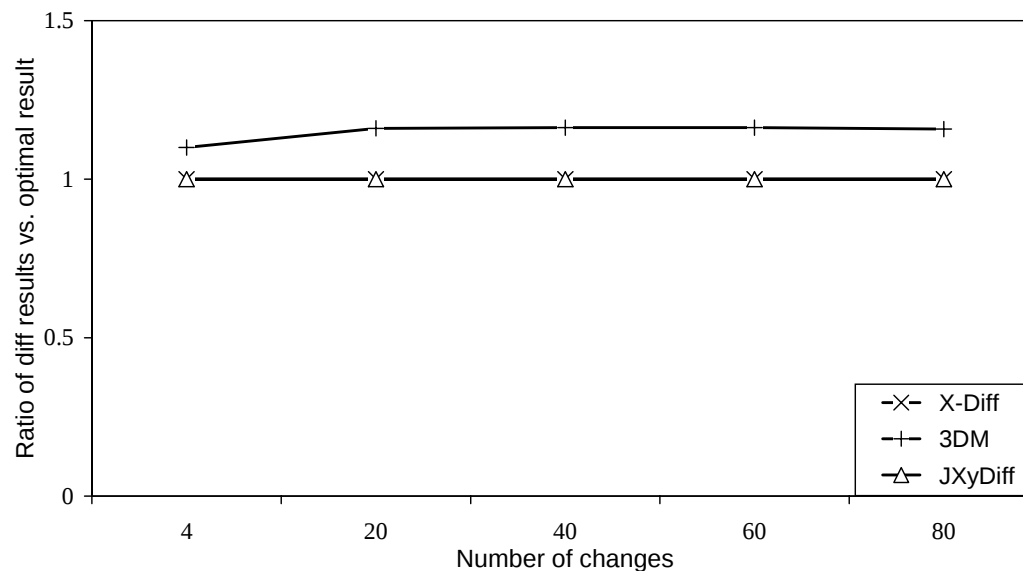
Experiments involving controlled modifications

Experimental setup

- Genomic data of chromosome 10 of yeast with 400 genes
- Introduce n ($n = 4, 20, 40, 60, 80$) changes and pair wise combinations of changes and produce new documents
 - Update sequence
 - Insert gene
 - Delete gene
 - Merge genes
 - Split gene
- Compare base document with each of new documents using X-Diff, JXy-Diff, 3DM
- Process edit scripts

Experiment: Insert gene

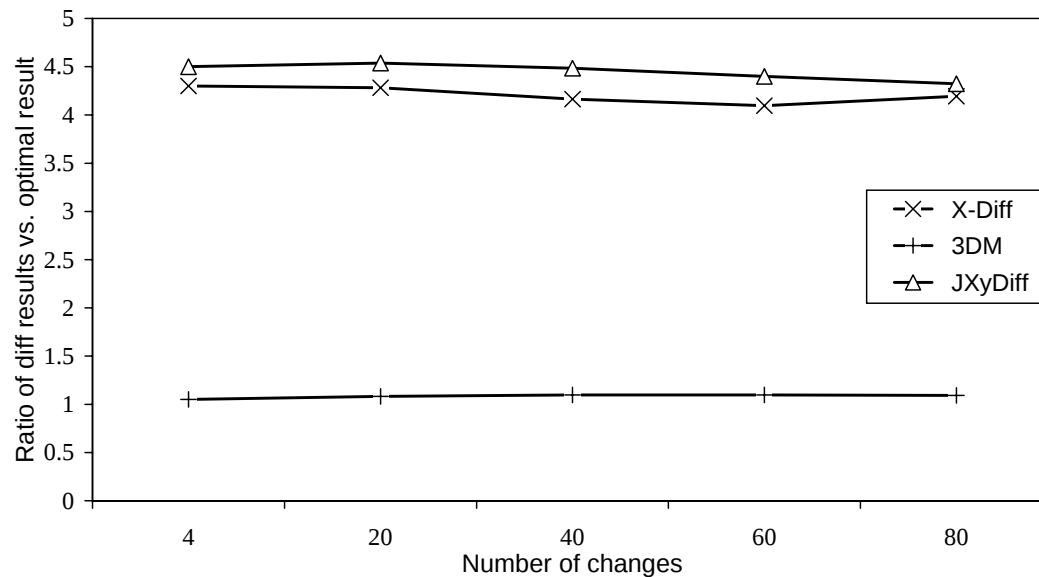
Relative performance



- Correct identification by X-Diff, JXyDiff, 3DM
- Additional copies of inserted sequence elements detected by 3DM

Experiment: Insert gene & Delete gene

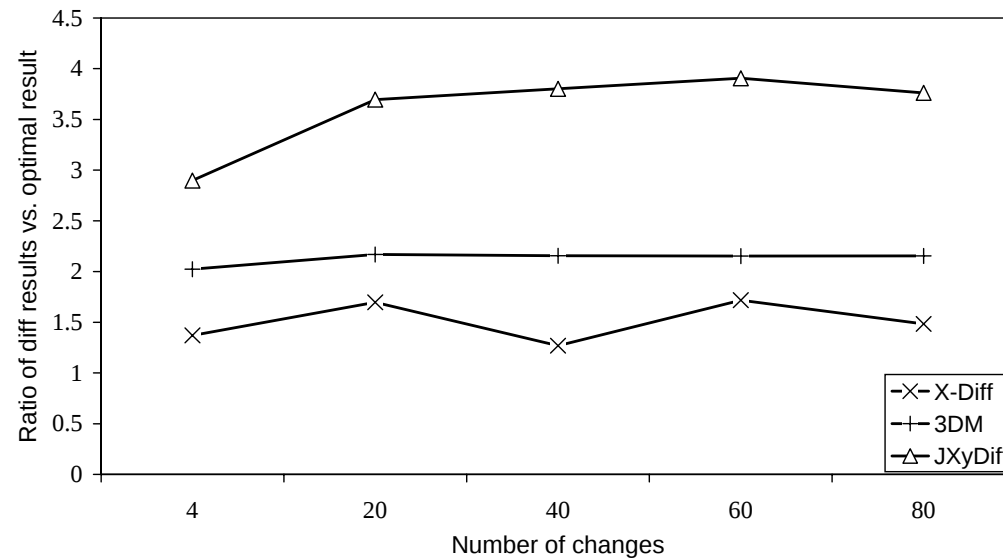
Relative performance



- 3DM:
 - + Inserted and deleted elements
 - Additional copies of sequence
- JXyDiff:
 - + Inserts and deletes of mRNA and gene
 - Incorrect matching of CDS elements
- X-Diff
 - Elements are reported as updates

Experiment: Merge neighbouring genes

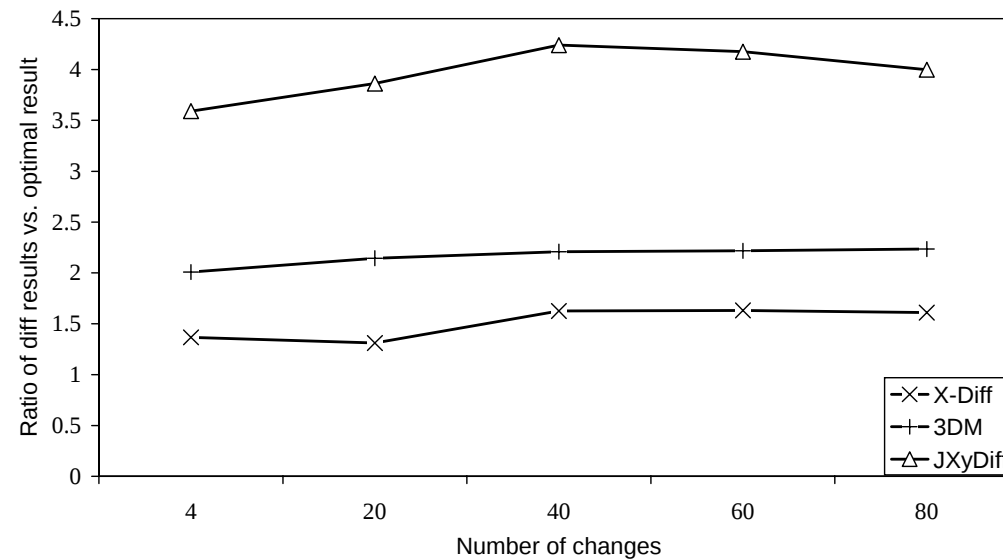
Relative performance



- X-Diff:
 - + Deleted elements
 - Updated elements (mRNAs - CDS)
- JXyDiff:
 - Fraction of deleted elements
 - Very few updated elements (CDS)
- 3DM:
 - + Majority of deleted and updated elements
 - Updates as inserts & deletes

Experiment: Split gene

Relative performance



- X-Diff:
 - + Inserted, majority of updated
 - Updated mRNA matched with gene
- JXyDiff:
 - + Inserted
 - Few updated
- 3DM:
 - + Inserted
 - Updated

Overview

- Motivation
- Problem description
- XML difference algorithms
- Experiments involving controlled modifications
- **Evaluation involving real genomic data**
- Conclusions

Evaluation involving real genomic data

Experimental setup and results

- Ensembl releases 41 and 42 of yeast chromosomes 3 and 5
- X-Diff
 - + Singular change within an element
 - Large number of change within an element
- JXyDiff
 - Matching elements correctly, in particular CDS
- 3DM
 - Detects majority of changes
 - Insert, delete of dbreference, qualifier in CDS
 - Update of sequence
 - Change of location of genes

Chromosome	X-Diff	JXyDiff	3DM
3	1805	13223	1864
5	1130	18475	679

Overview

- Motivation
- Problem description
- XML difference algorithms
- Experiments involving controlled modifications
- Evaluation involving real genomic data
- **Conclusions**

Conclusions

- Algorithms effective for detecting subset of changes
- All require significant post-processing of edit scripts
- Properties of XML representation of genomic data
 - Large number of siblings
 - Related elements not easily identifiable as such
 - Very similar contents of nodes and sub-trees
- Properties of algorithms
 - Treating XML documents as unordered trees
 - Fuzzy matching using q-grams
 - Propagating matches bottom-up

Future work

- Alterations to XML representation of genomic data to address:
 - Large number of siblings
 - Related elements not easily identifiable as such
 - Very similar contents of nodes and sub-trees
- New XML difference algorithm

Thanks!

Questions?