

Evaluating Contaminations in Genome Sequences

Pavankumar Videm

Chair for Bioinformatics Freiburg

Prof. Dr. Rolf Backofen

Dr. rer. nat. Dominic Rose

University of Freiburg

- What is genome contamination?
- Possible contaminations are from
 - Vectors
 - Adapters
 - Linkers
- Sequencing errors
- How sequences can be contaminated?

A scenario

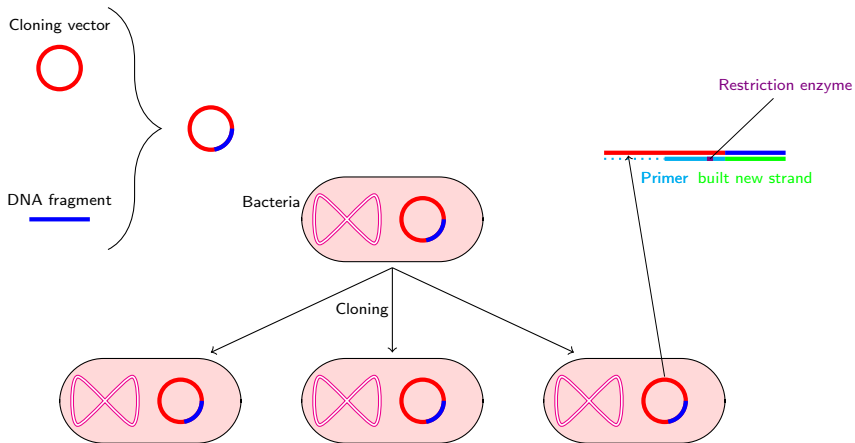


Figure: A scenario showing source of contamination

Consequences of contaminations

- Pollution of public databases
- Meaningless analysis

Our Strategy

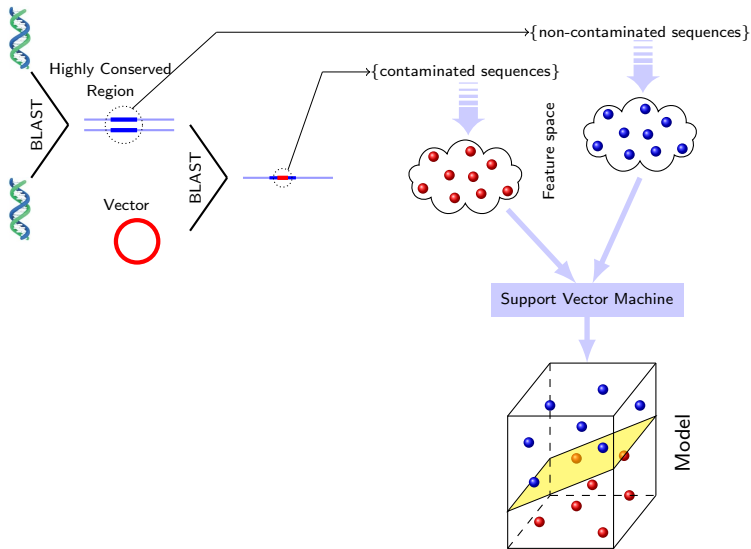


Figure: Model generation - classifies contaminated and non-contaminated sequences

First of all, acquiring data

- Genomes
 - UCSC genome browser¹
 - 58 different metazoans
- Vector sequences
 - UniVec database from NCBI²
 - VectorDB from Stanford³
 - Total of 4696 unique sequences

¹<http://hgdownload.cse.ucsc.edu/downloads.html>

²<ftp://ftp.ncbi.nih.gov/pub/UniVec/>

³<http://genome-www.stanford.edu/vectordb/>

Before choosing genomes

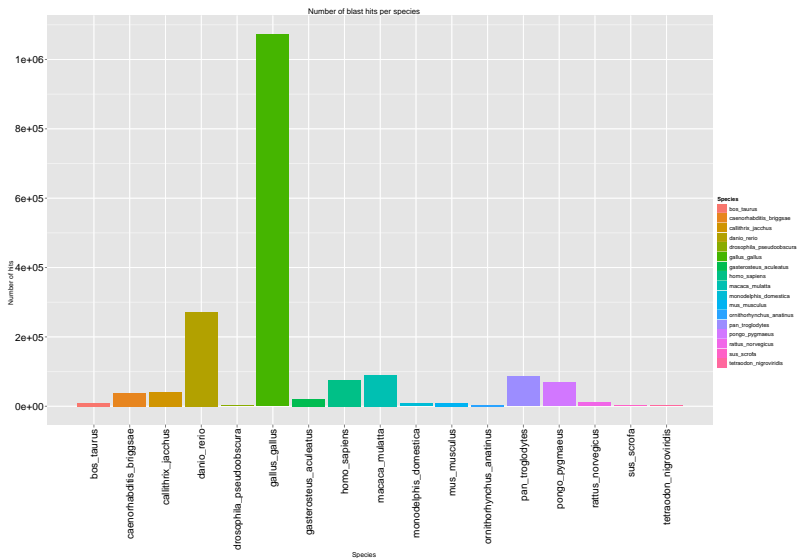


Figure: BLAST hits vectors against genomes

Hunt for Highly Conserved Regions

- Megablast with E-value '1e-5' and identity greater than or equal to 90%.

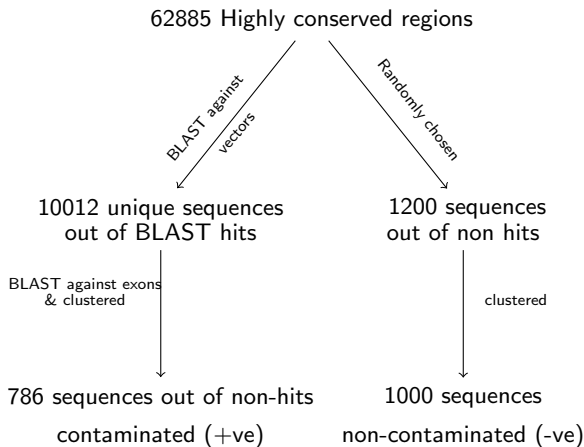
- Callithrix jacchus
- Caenorhabditis briggsae
- Gasterosteus aculeatus
- Macaca mulatta
- Mus musculus
- Pan troglodytes

vs

Danio rerio
and
Gallus gallus

2852087 unique BLAST hits

642885 Highly Conserved Regions



- Contaminated sequences
 - Min length: 33
 - Max length: 528
 - Average length: 91.01
 - Median: 69.0000
 - SD: 55.0659
- Non-contaminated sequences
 - Min length: 33
 - Max length: 683
 - Average length: 105.11
 - Median: 77.00
 - SD: 86.0659

Features and Support Vector Machine

- Features Used:
 - K-mer count
 - K-mer distances
 - K-mer existence as binary class
 - K-mer mismatch scores
 - GC content
 - Stop codons(ATG, CTG, GTG, TTG)
 - Start codons(TAA, TAG, TGA)
- Support Vector Machines:
 - Libsvm 3.0
 - Radial basis kernel
 - Polynomial kernel with degree 3
 - Weka 3.7.3
 - Sequential Minimal Optimization scheme, polynomial kernel of degree 4
 - All other default parameters

Test set	Total Number of Instances	Correctly Classified Instances	Incorrectly Classified Instances
Test set from existing data	357(P:153,N:204)	348(97.48%) TP:146, TN:202	9(2.52%) FP:7, FN:2
Danio rerio ncRNA	4431	4163 (93.96%)	268 (6.04%)
Felis catus ncRNA	738	677 (91.73%)	61 (8.27%)
Gallus gallus ncRNA	1102	1028 (93.29%)	74 (6.71%)
Homo sapiens ncRNA	2647	2580 (97.47%)	67 (2.53%)
Homo sapiens mRNA	1000	978 (97.80%)	22 (2.20%)
Mus musculus ncRNA	752	708 (94.15%)	44 (5.85%)
Rattus norvegicus ncRNA	757	704 (92.99%)	53 (7.01%)
Rfam sequences	786	718 (91.34%)	68 (8.66%)
Human UCRs Watson strand	2081	2008 (96.49%)	73 (3.51%)

Table: Results for different test sets

All ncRNAs from Ensembl ftp – <http://www.ensembl.org/info/data/ftp/index.html>

Human mRNA extracted from UCSC table browser – <http://genome.ucsc.edu/cgi-bin/hgTables>

Human Watson strand UCRs by Bejerano G – <http://users.soe.ucsc.edu/~jill/ultra.html>




- AUC of 0.972 for best model
- Observation: Most of incorrectly classified instances are from predicted sequences or DNA clones.

- Incorrect predictions - evidence for protein coding
- A standalone tool

- Contaminations are sometimes hazardous
- Care should be taken while analysis
- More care should be taken while sequencing genomes

It's an honor to thank

- Chair of Bioinformatiks.
- Dominic Rose for suggesting interesting topic and helping patiently.
- Kousik Kundu, showed an active participation and valuable suggestions in using Weka.

-  [1]Chang, Chih-Chung and Lin, Chih-Jen, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology; Volume 2 Issue 3 Pages 27:1–27:27, 2011
-  [2]Coker JS, Davies, *E Identifying adaptor contamination when mining DNA sequence data*, Biotechniques; Volume 37 Issue 2 Pages 194, 196, 198, Aug 2004
-  [3]Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H., *The WEKA Data Mining Software: An Update*, SIGKDD Explorations; Volume 11 Issue 1, 2009

Thank you for your attention !!