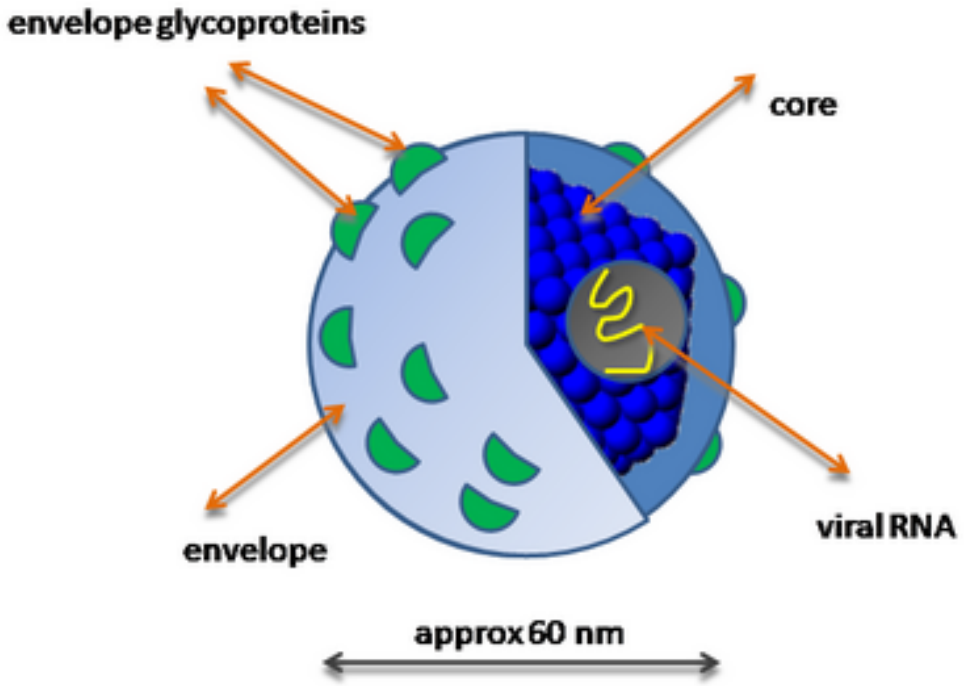
From Genomic Sequence Data to Genotype: A Proposed Machine Learning Approach for Genotyping Hepatitis C Virus

Genaro Hernandez Jr

CMSC 601 Spring 2011

Hepatitis C virus (HCV)

- WHO
  - Estimates 3% of the world population is infected with HCV
  - 170 million people are chronic carriers at risk of developing cirrhosis and/or liver cancer



Structure of Hepatitis C Virus

HCV Genotyping

- HCV is classified into 6 genotypes
  - Several subtypes per genotype
  - Example: 1a denotes genotype 1, subtype a

- Genotyping methods
  - Common diagnostic tools for HCV genotyping
    - Molecular assays: Genome typing assays
      - Examine particular regions of the HCV genome.
      - Example: Amplification of HCV genome followed by digestion with restriction enzymes and restriction length polymorphism analysis

Computational Problem

How do we represent HCV genome sequence data and apply machine learning methods to determine genotype?

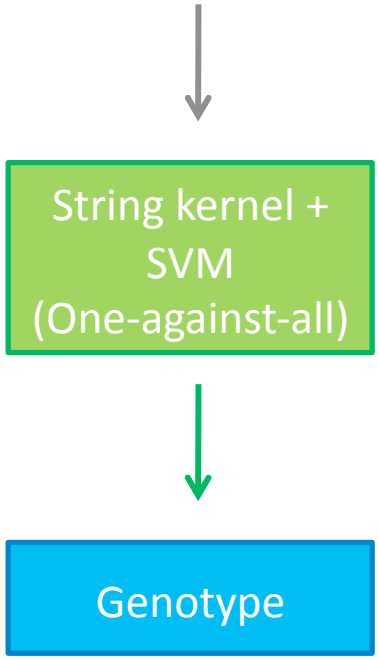
Significance of Genotyping

1. Indicates the severity of infection

2. Informs clinical decision making
   — Different genotypes require particular treatments

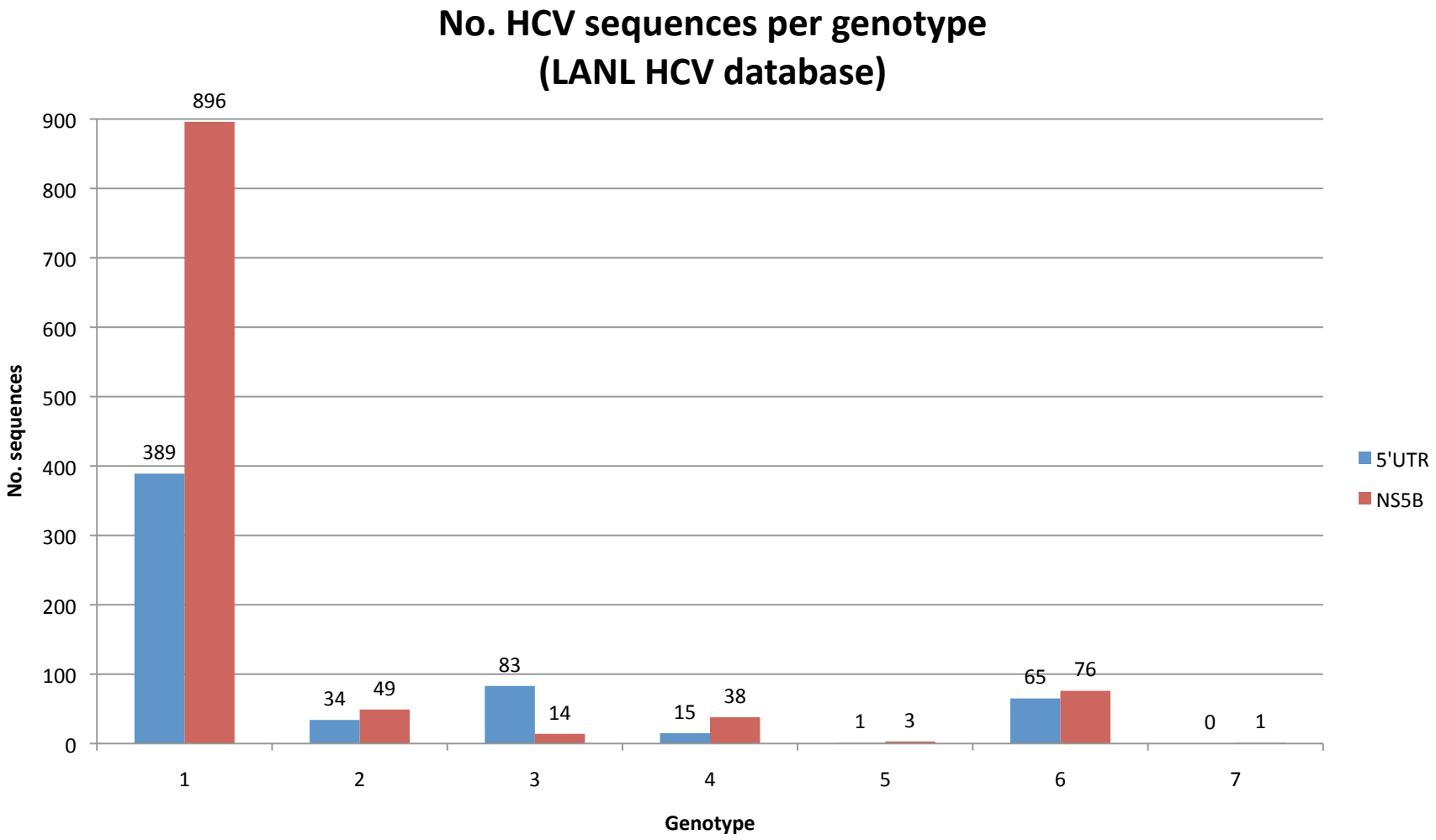
Proposed Approach: Classifier

GCCAGCCCCCGATTGGGGGCGACAC
TCCACCATAGATCACTCCCCTGTGAG
GAACTATTGTCTTCACGCAGAAAGCG
TCTAGCCATGGCGTTAGTATGAGTGTC
GTGCAACCTCCAGGACCCCCCCTCCC
GGGAGAGCCATAGTGGTCTGCGGA



One-against-all (examples)					
C1	C2	C3	C4	C5	C6
+	-	-	-	-	-
C1	C2	C3	C4	C5	C6
-	+	-	-	-	-

Preliminary data: Data collection



Related Work

Research group: Qiu et al. (2009)
Methodology: Position weight matrix, SVMs and random forest
Results: Achieved 99% prediction accuracy.
Strengths: High prediction accuracy
Weaknesses: Several pre-processing steps (e.g. sequence alignment, creation of position weight matrices, concatenation of matrices)

Research group: Hraber et al. (2008)
Methodology: Branching index to combine distance and phylogeny methods
Results: Characterization of 'problematic' sequences
Strengths: Characterization of problematic sequences
Weaknesses: Several pre-processing steps (e.g. creation of phylogenetic trees)

Evaluation of Results

K-fold cross validation

- E.g. 10-fold cross validation
- To estimate the predictive ability or unknown samples

Overall accuracy

- Dataset not used for training
- overall accuracy = (TP + TN)/(TP + TN + FP + FN)
  - TP = true positives
  - TN = true negatives
  - FP = false positives
  - FN = false negatives

Conclusion

Problem

How to represent HCV sequence data, apply machine learning to determine genotype?

Significance

Genotype indicates severity of infection, informs clinical decisions in treating infection.

Proposed methodology

String kernels to represent viral sequence data, SVMs for classification of the virus with one-against-all

Related work

1. Qiu et al. (2009), position weight matrix, SVMs and random forest, achieved 99% prediction accuracy.
2. Hraber et al. (2008), branching index to combine distance and phylogeny methods

Evaluation

K-fold cross validation, overall prediction accuracy.