Identifying Hosts of Families of Viruses: A Machine Learning Approach

Anil Raj, Michael Dewar, Gustavo Palacios, Raul Rabadan, Chris H. Wiggins

(Submitted on 29 May 2011)

Identifying viral pathogens and characterizing their transmission is essential to developing effective public health measures in response to a pandemic. Phylogenetics, though currently the most popular tool used to characterize the likely host of a virus, can be ambiguous when studying species very distant to known species and when there is very little reliable sequence information available in the early stages of the pandemic. Motivated by an existing framework for representing biological sequence information, we learn sparse, tree-structured models, built from decision rules based on subsequences, to predict viral hosts from protein sequence data using popular discriminative machine learning tools. Furthermore, the predictive motifs robustly selected by the learning algorithm are found to show strong host-specificity and occur in highly conserved regions of the viral proteome.

Comments: 11 pages, 7 figures, 1 table
Subjects: **Quantitative Methods (q-bio.QM)**; Applications (stat.AP)
Cite as: **arXiv:1105.5821 [q-bio.QM]**
(or **arXiv:1105.5821v1 [q-bio.QM]** for this version)

Submission history

From: Anil Raj [view email]
[v1] Sun, 29 May 2011 19:36:40 GMT (183kb,A)

Which authors of this paper are endorsers?