Effective in-silico sequence-based prediction of FMDV vaccine matching

EuFMD Open Session

Borgo Egnazia, 30-10-2018

Mana Mahapatra
Satya Parida
(Data production)

Yasaman Kalantar-Motamedi,
Sophie Mahendran,
Luca Ferretti, Paolo Ribeca
(Computer)

BBSRC bioscience for the future

The Pirbright Institute receives strategic funding from BBSRC.

www.pirbright.ac.uk

VNT data

Given a serum, focuses on measuring the ratio between the neutralising power against the vaccine (homologous) and the virus being tested (heterologous).

Approach already explored in the past (Reeve et al.) on smaller datasets.

However more and more serological data available, coupled with full-capsid (or full-genome) information.

Approach routinely pursued by WRLFMD for quite some time now.

New attempt based on machine learning approaches.

De-noising

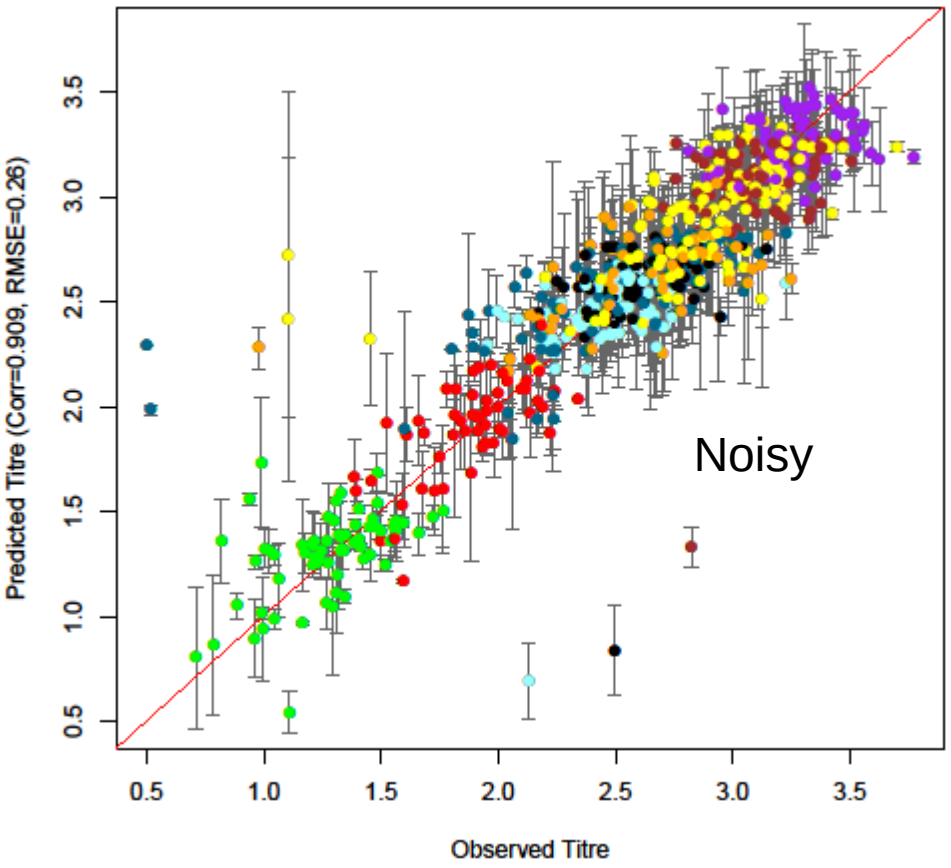
Data intrinsically noisy, with large biological variations.

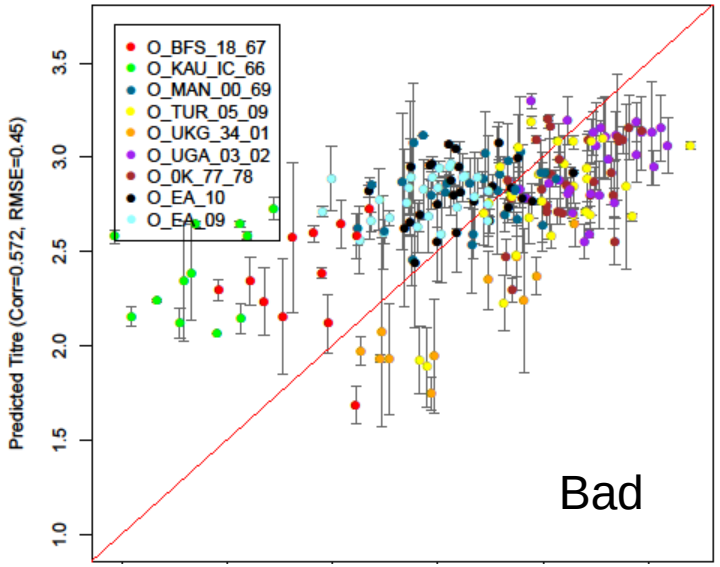
Trying to separate signal from noise
(garbage in, garbage out!).

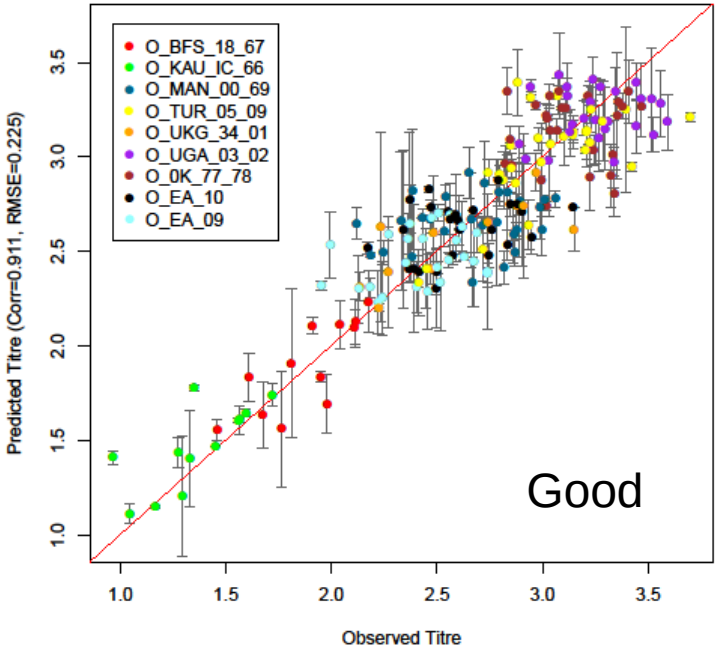
Intuitively:

- If two sequences which are similar generate a different heterologous titre, problem!
  Can be corrected, but needs many sequences
- If a measurement gives a surprising value for the homologous titre, problem!

The goal is being able to filter out bad data.


Predicted Titre (Corr=0.909, RMSE=0.26)
Observed Titre
Noisy


O_BFS_18_67
O_KAU_IC_66
O_MAN_00_69
O_TUR_05_09
O_UKG_34_01
O_UGA_03_02
O_0K_77_78
O_EA_10
O_EA_09
Predicted Titre (Corr=0.572, RMSE=0.45)
Bad


O_BFS_18_67
O_KAU_IC_66
O_MAN_00_69
O_TUR_05_09
O_UKG_34_01
O_UGA_03_02
O_0K_77_78
O_EA_10
O_EA_09
Predicted Titre (Corr=0.911, RMSE=0.225)
Observed Titre
Good

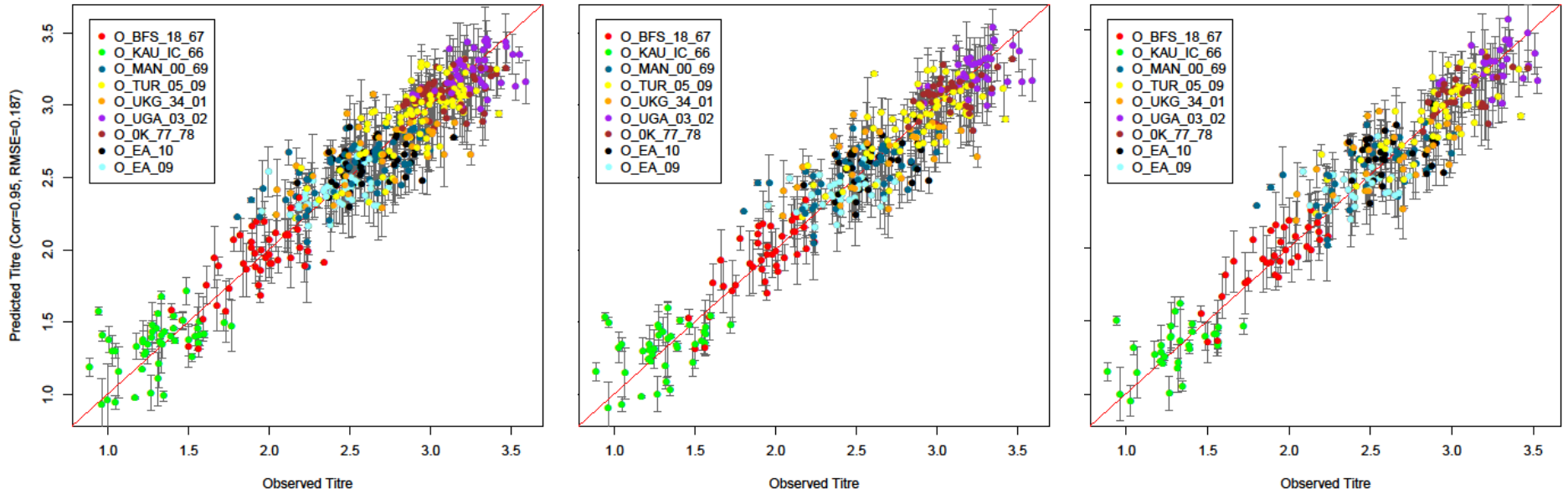
Validation

The model is prone to over-fitting!

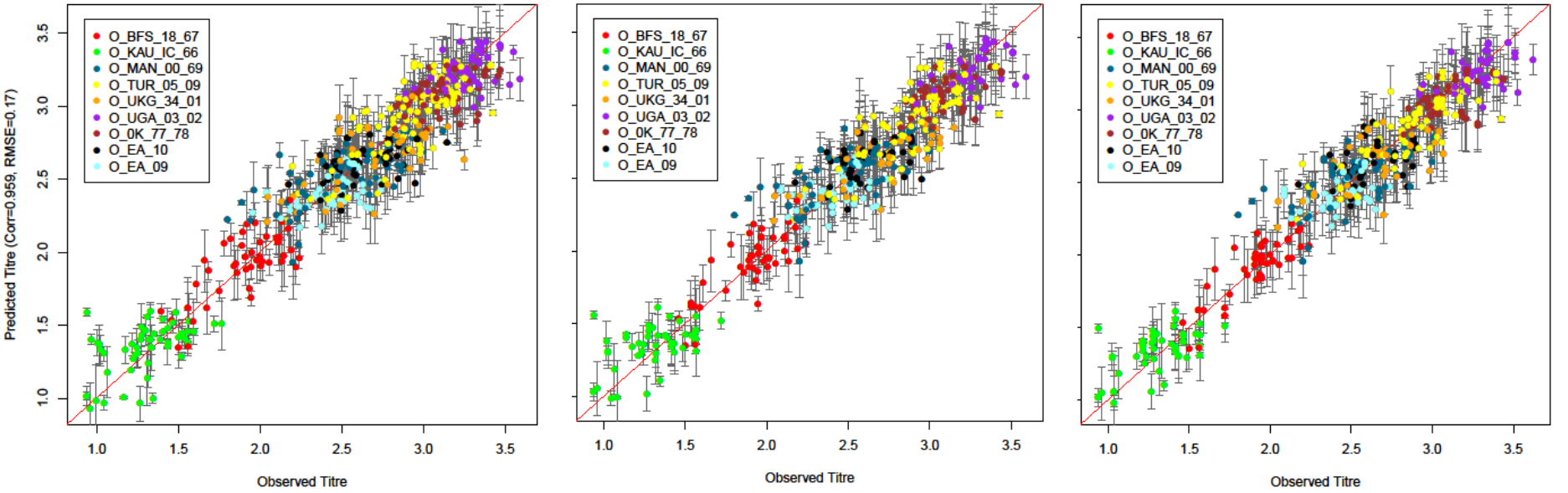
You can get an almost perfect correlation
with no predictive power.

Subtle problems due to the fact that sampling
is not uniform nor random in sequence variant space
(no freedom to choose samples).
Sample size corrections needed when doing validation.

We think we have found a good definition.


O_BFS_18_67
O_KAU_IC_66
O_MAN_00_69
O_TUR_05_09
O_UKG_34_01
O_UGA_03_02
O_0K_77_78
O_EA_10
O_EA_09
Predicted Titre (Corr=0.95, RMSE=0.187)
Observed Titre
1.0
1.5
2.0
2.5
3.0
3.5

Better!

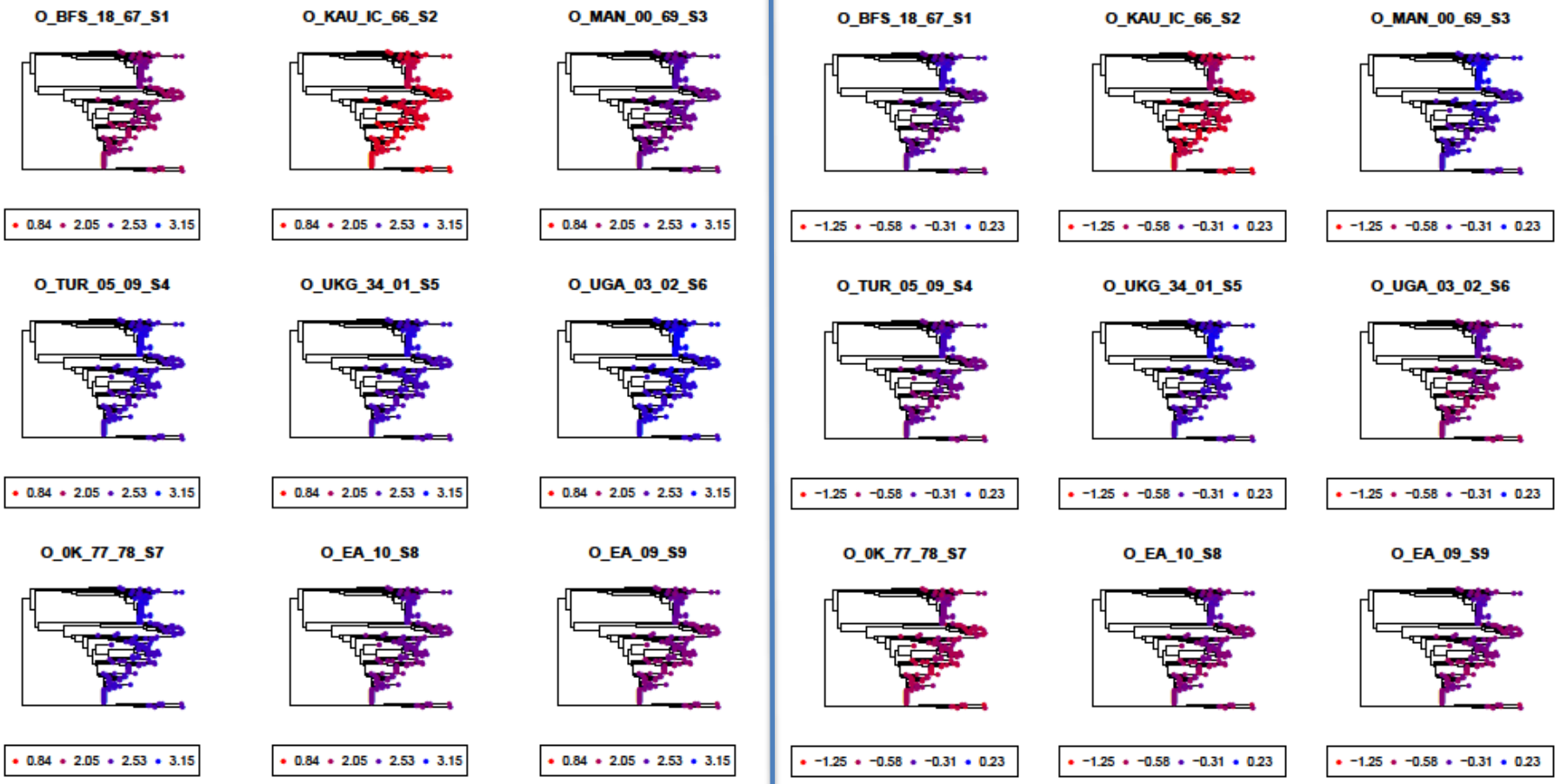

O_BFS_18_67
O_KAU_IC_66
O_MAN_00_69
O_TUR_05_09
O_UKG_34_01
O_UGA_03_02
O_0K_77_78
O_EA_10
O_EA_09
Predicted Titre (Corr=0.959, RMSE=0.17)
Observed Titre
1.0
1.5
2.0
2.5
3.0
3.5

Does it really work?

Predicted titres

All GenBank data

Predicted R1



Outlook

The model works in a similar way for other serotypes (tested A).

Our model allows integration of biochemical features.
We can read out what are the important ones.

We can predict epitopes out of the model
(easy to reproduce most previously published ones—
but that only possible if enough variation in the input).

There is much more data available!
We are integrating it.

Acknowledgements

Yasaman Kalantar-Motamedi

Luca Ferretti

Sophie Mahendran

Mana Mahapatra

Satya Parida

BBSRC bioscience for the future

The Pirbright campus is being redeveloped

BBSRC bioscience for the future

The Pirbright Institute receives strategic funding from BBSRC.

www.pirbright.ac.uk