

Machine learning for cancer informatics

Jean-Philippe Vert

Inserm U900 - Mines ParisTech - Institut Curie

Team « Statistical machine learning and modelling of biological systems »



Inserm

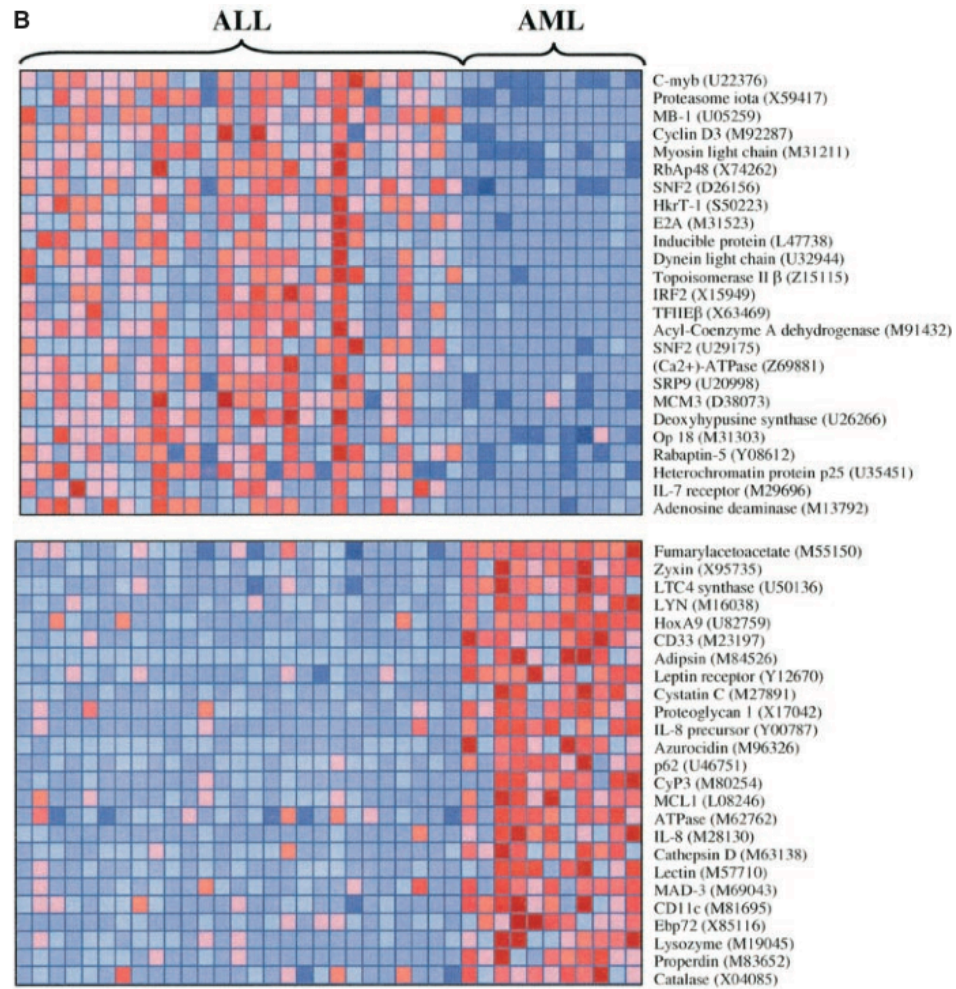
Team's goal

Develop new mathematical/computational models and tools to contribute to:

1. Diagnosis, prognosis and predictive models
2. Identification of important pathways and new drug targets
3. Identification of new drugs

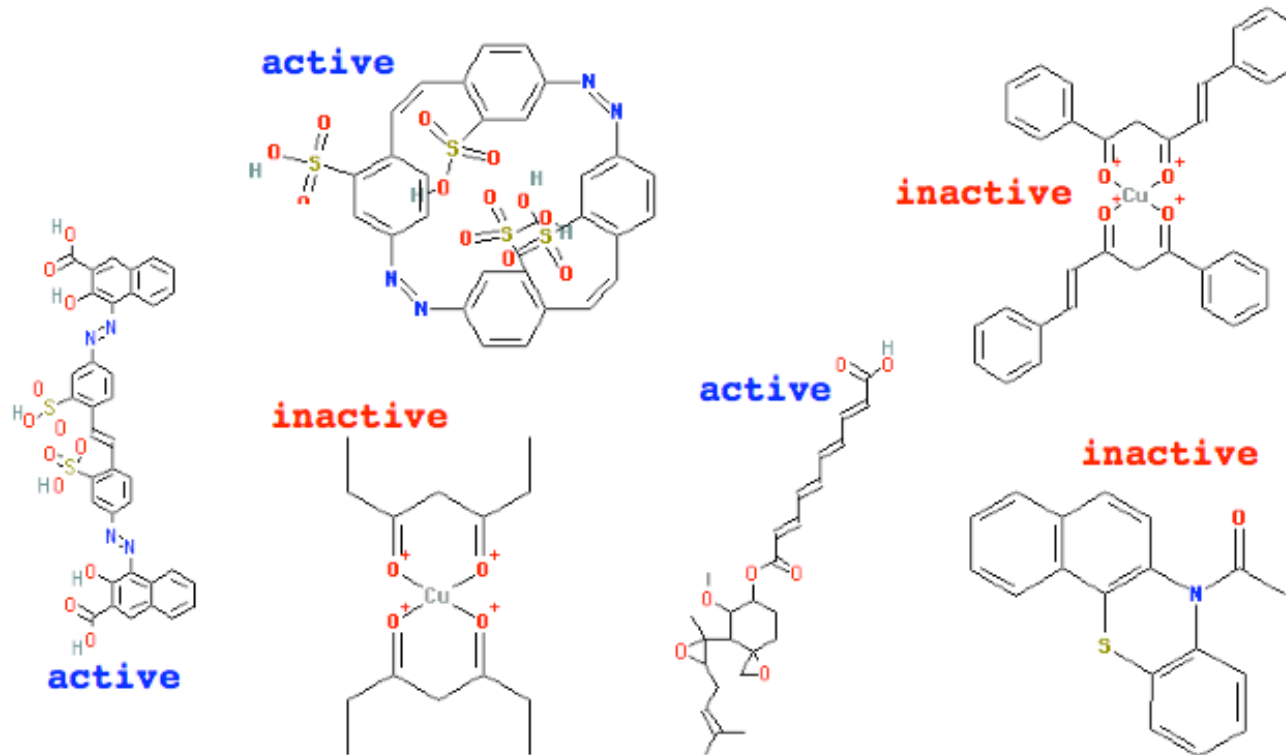


Motivation: Diagnosis / Prognosis from genome / transcriptome



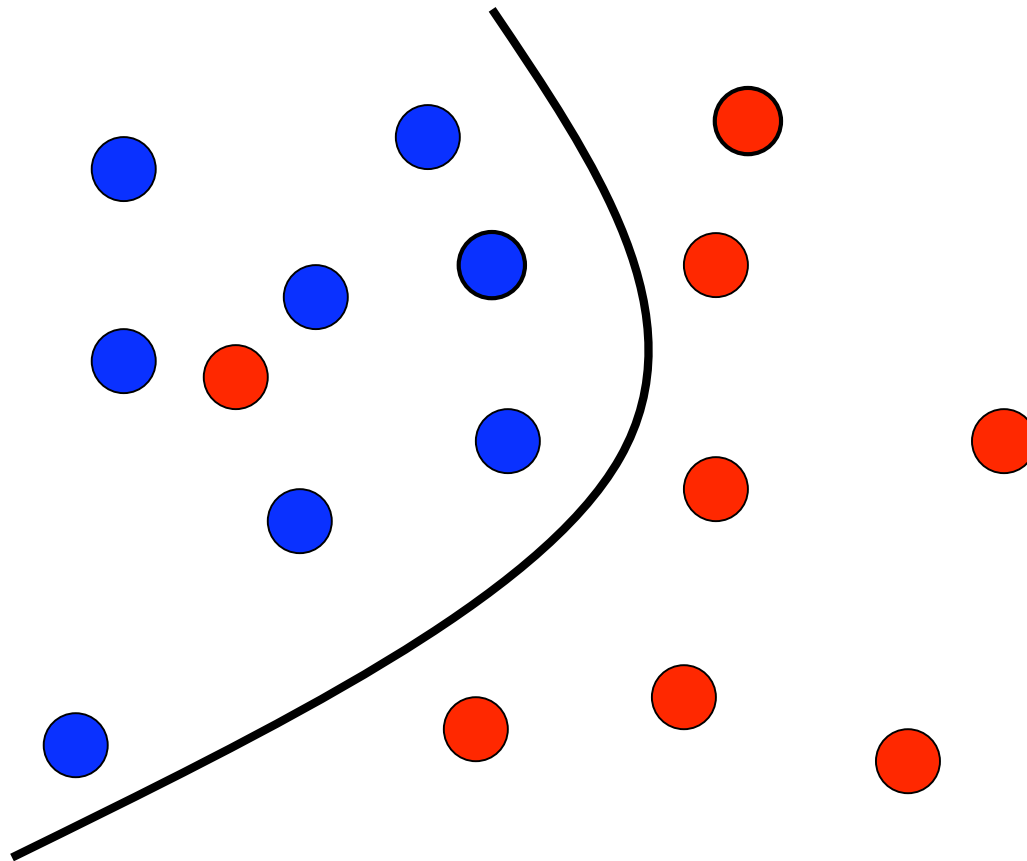
Golub et al., Science, 1999

Motivation: Virtual screening

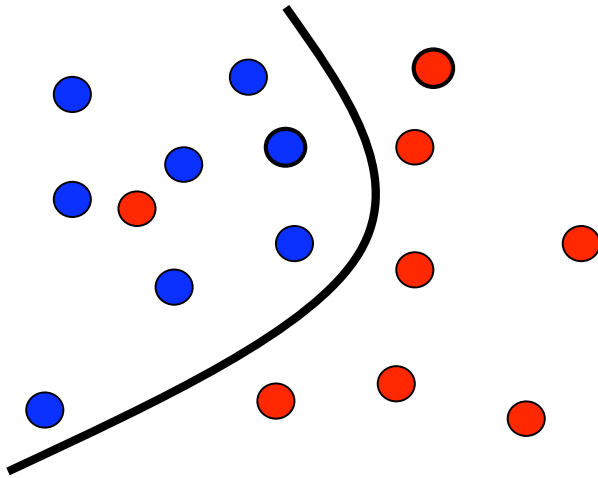


From <http://cactus.nci.nih.gov>

Pattern recognition (aka supervised classification)



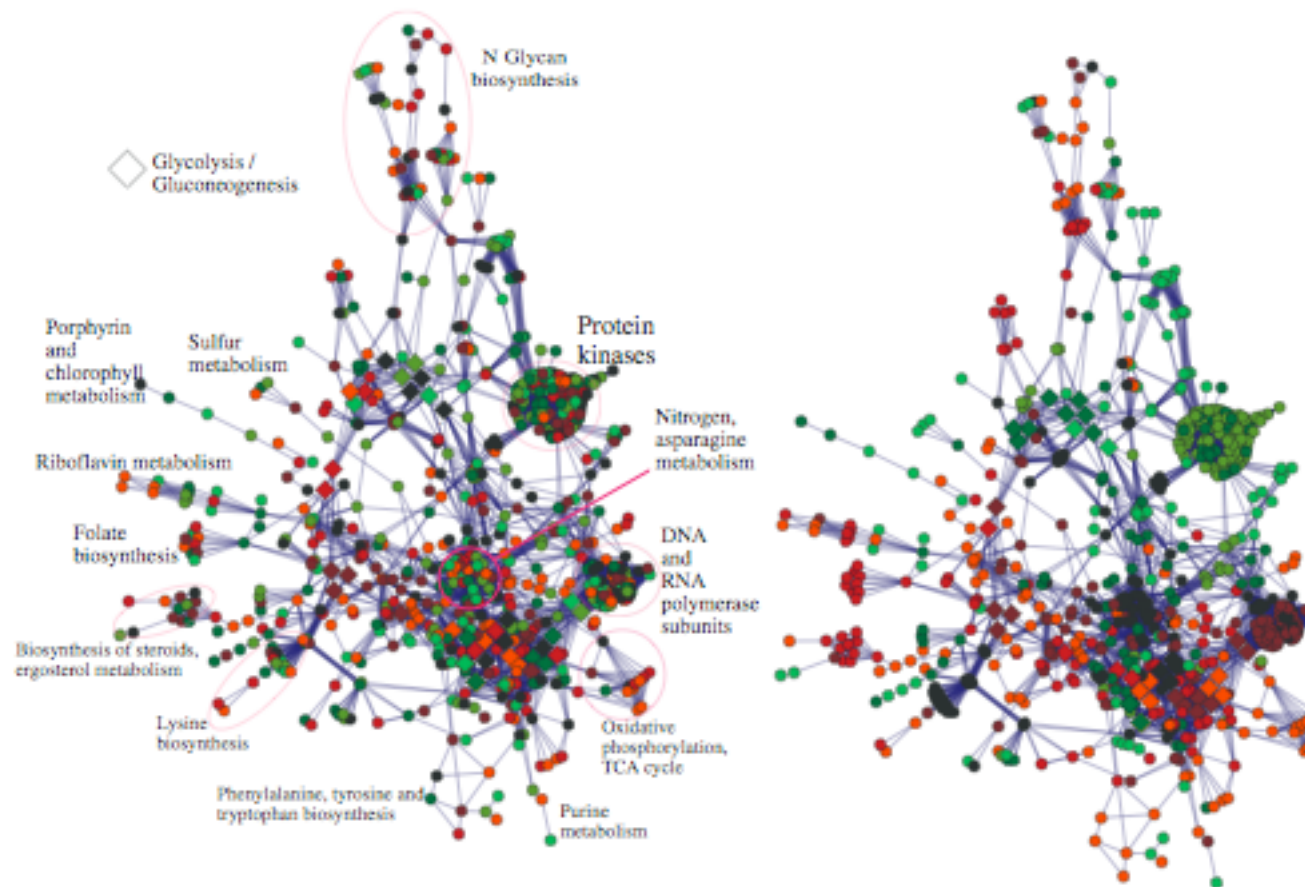
Pattern recognition (aka supervised classification)



Challenges

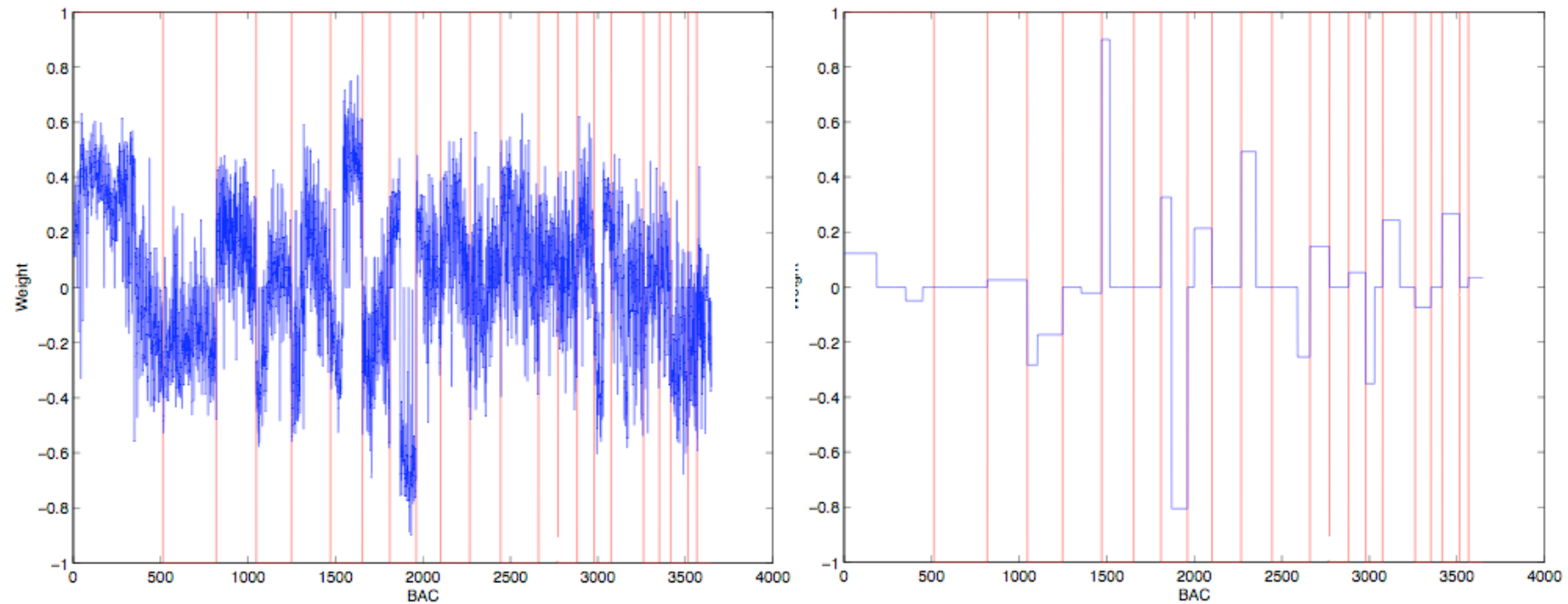
- High dimension
- Few samples
- Structured data
- Inclusion of prior knowledge
- Fast and scalable algorithms

Application: Discriminant signatures from expression data that highlight dysregulated pathways



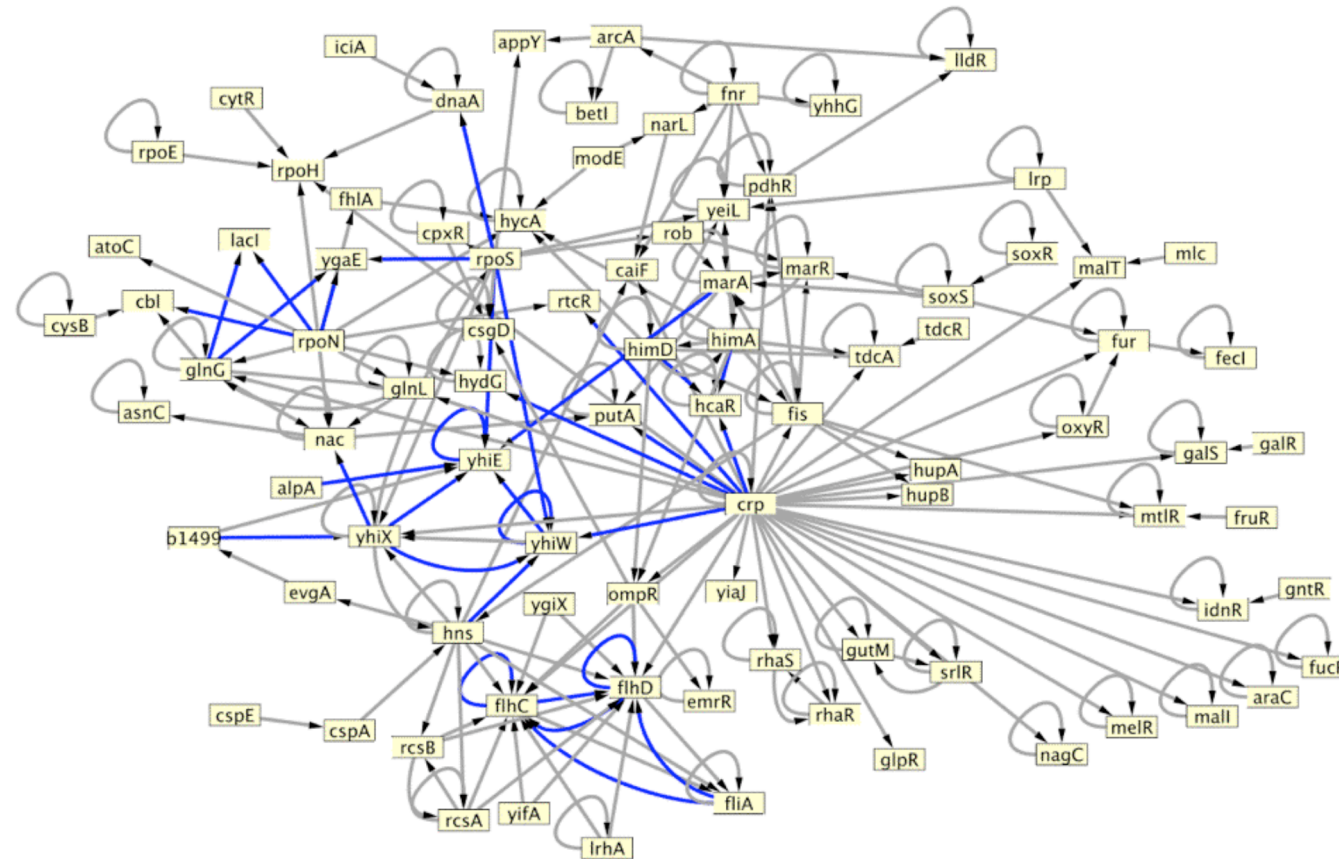
Rapaport et al., Bioinformatics, 2008.

Application: Discriminant CGH signature with automated detection of discriminant regions



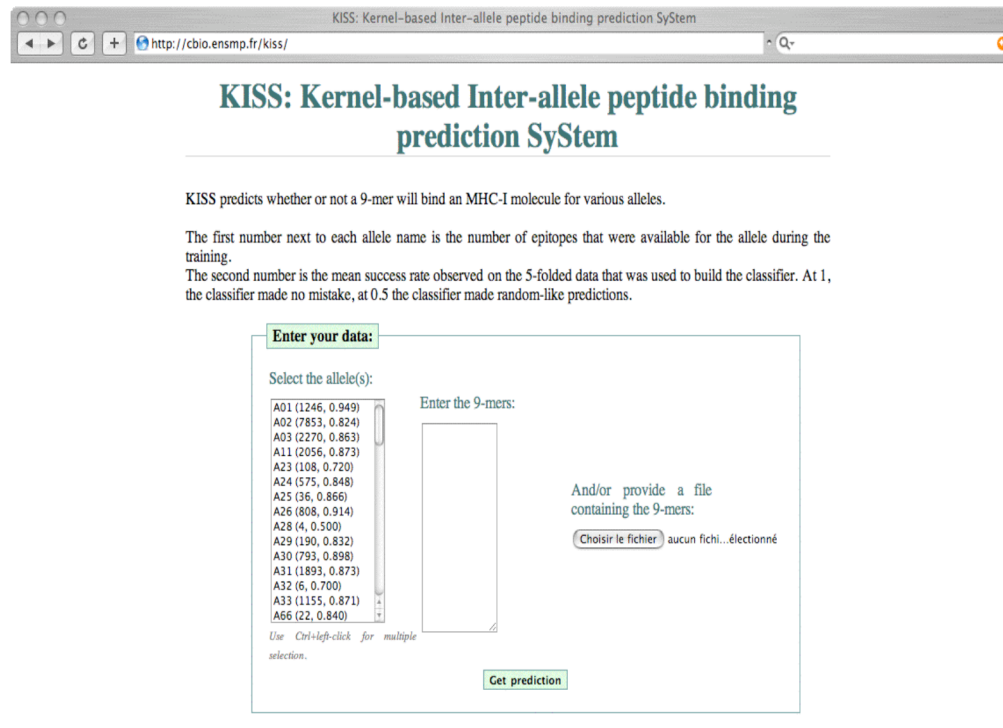
Rapaport et al., Bioinformatics, 2008.

Application: Identification of new regulations from expression data



Mordelet and Vert, *Bioinformatics*, 2008.

Application: Prediction of peptide-MHC I binding for alleles with few known peptides



KISS: Kernel-based Inter-allele peptide binding prediction SyStem

KISS predicts whether or not a 9-mer will bind an MHC-I molecule for various alleles.

The first number next to each allele name is the number of epitopes that were available for the allele during the training.
The second number is the mean success rate observed on the 5-folded data that was used to build the classifier. At 1, the classifier made no mistake, at 0.5 the classifier made random-like predictions.

Enter your data:

Select the allele(s):

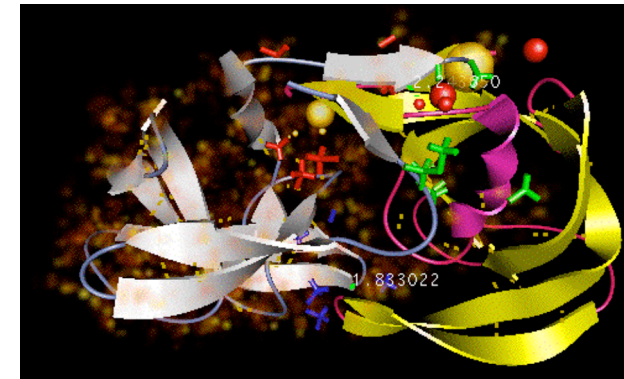
- A01 (1246, 0.949)
- A02 (7853, 0.824)
- A03 (2270, 0.863)
- A11 (2056, 0.873)
- A23 (108, 0.720)
- A24 (575, 0.848)
- A25 (36, 0.866)
- A26 (808, 0.914)
- A28 (4, 0.500)
- A29 (190, 0.832)
- A30 (793, 0.898)
- A31 (1893, 0.873)
- A32 (6, 0.700)
- A33 (1155, 0.871)
- A66 (22, 0.840)

Use **Ctrl**-left-click for multiple selection.

Enter the 9-mers:

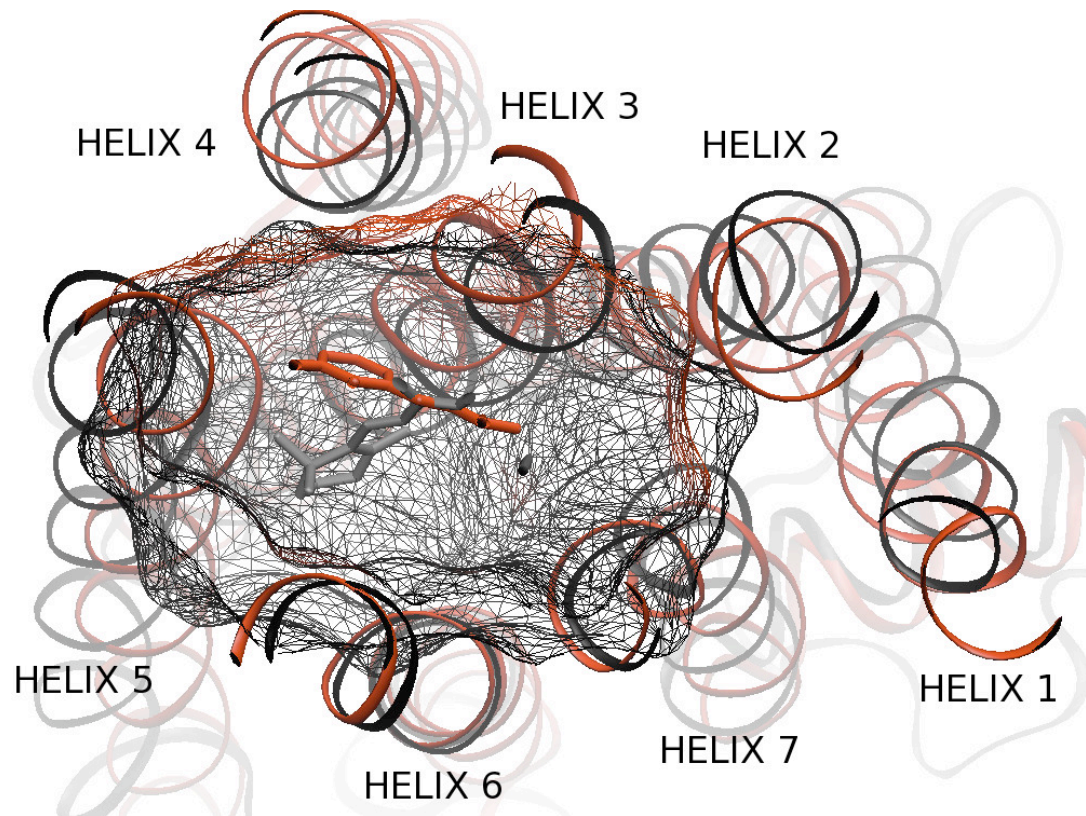
And/or provide a file containing the 9-mers:

aucun fichi...électionné



Jacob and Vert, Bioinformatics, 2008

Application: Chemogenomics and virtual screening of GPCR



Jacob et al., BMC Bioinformatics, 2008.

Conclusion

- Many problems require new methods in statistics / machine learning
- General trend: include prior knowledge in a computational efficient framework
- We seek collaborations!

Jean-Philippe.Vert@curie.fr

Team members:

Kevin Bleakley, Brice Hoffmann, Martial Hue, Laurent Jacob, Christian Lajaunie, Fantine Mordelet, Philippe Rouillier, Isabelle Schmitt, Véronique Stoven, Jean-Philippe Vert, Yoshihiro Yamanishi, Misha Zaslavskiy.

+ many joint work with U900