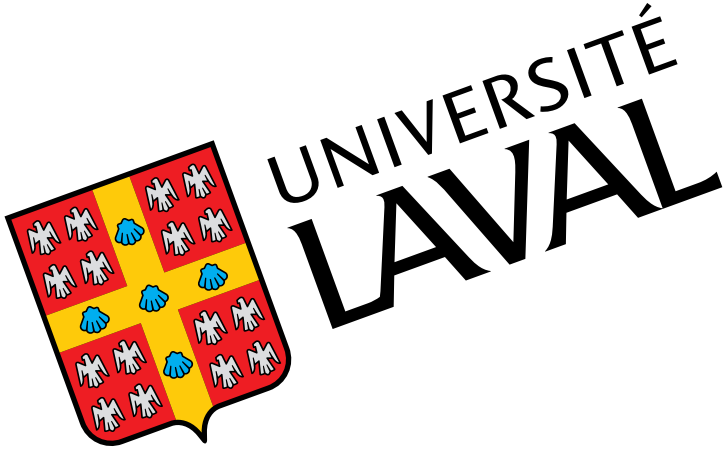
HIV-1 coreceptor usage prediction without multiple alignments



Sébastien Boisvert, M.Sc. student, Université Laval
www.graal.ift.ulaval.ca
Directors: Jacques Corbeil and Mario Marchand

HIV

- HIV (human immunodeficiency virus) is the causative agent of the deadly disease known as AIDS (acquired immunodeficiency syndrome)
- HIV integrates its genome in the host genome.
- genome size: 10 kb
- molecule type: RNA
- 9 genes
- HIV-1 (spread world-wide) and HIV-2

HIV infection

- HIV uses a CD4 receptor and a chemokine receptor to infect cells
- chemokine receptors are CCR5 and CXCR4
- CXCR4-using viruses are associated with faster depletion of T cells CD4+
- HIV usually infects with CCR5 and switches to CXCR4 with disease progression
- The V3 loop inside the gp120 protein of the retroviral envelope is a strong determinant of the coreceptor usage

Fighting HIV

- Many drugs are available, each having a specific molecular target (integrase, envelope, reverse transcriptase, coreceptor, etc.)
- Coreceptor inhibitors (CCR5- or CXCR4-specific)
- If one knows if a virus uses CCR5 and/or CXCR4, then a coreceptor inhibitor can be selected accordingly

Determination of the coreceptor usage

- Phenotypic assays and genotypic assays
- Phenotypic assays rely on recombinant DNA
- Genotypic assays rely on DNA sequencing (only the env gene of HIV is relevant here) and machine learning
- We investigated how the machine learning component can be enhanced.

A mathematical view of the problem

- $\mathcal{X}$: V3 loop protein sequences
- $\mathcal{Y} = \{-1, +1\}$ is a binary output space (ex.: CXCR4: yes or no)
- training set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, with $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\ \forall i$
- Each example (x_i, y_i) is distributed identically and independently with an unknown, but constant distribution $P_{\mathcal{X},\mathcal{Y}}$
- Learn from the patterns in the training set

Machine learning

- An algorithm A learns a classification function $h : \mathcal{X} \to \mathcal{Y}$
- only the observations in the training set $\mathcal{S}$ can be utilized
- h is a classifier
- h must be accurate on examples that are not in the training set

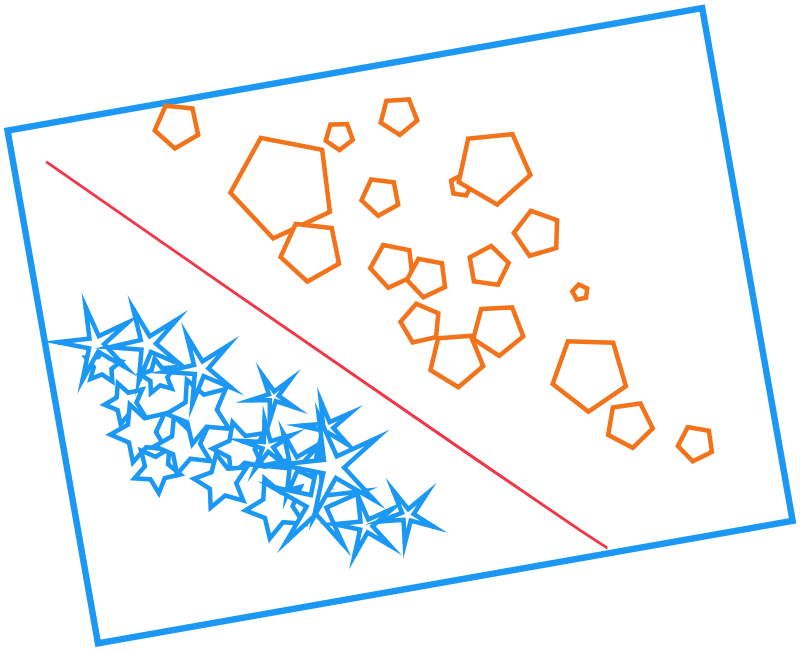
A kernel is a measure of similarity

- mapping function $\phi : \mathcal{X} \rightarrow \mathcal{R}^n$
- a kernel is a dot product in a feature space: $k(x, x') = \phi(x) \cdot \phi(x')$
- the kernel measures similarity: $k : \ \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{R}$ (biologically, we look for common motifs)

Linear classifiers

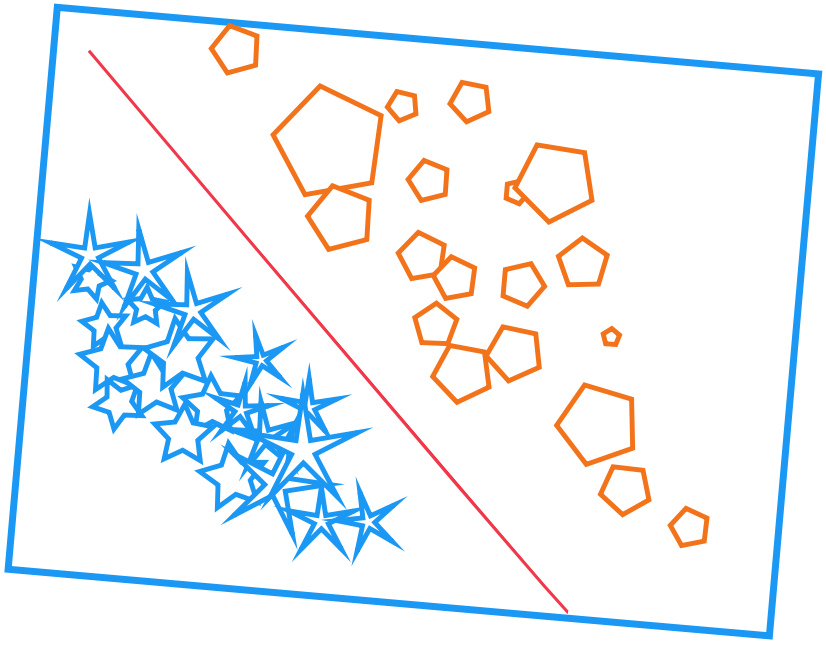
- We are interested in classifiers that can be written as $w \cdot \phi(x)$ because the predicted class is simply the sign of the dot product
- The support vector machine is a linear classifier

Support vector machines



- binary classifier $h: \mathcal{X} \to \{-1, +1\}$
- primal representation: (w, b), w is the normal vector and b is the bias
- separation surface: $\{\phi(x) : w \cdot \phi(x) + b = 0\}$
- $h(x) = \text{sgn}(w \cdot \phi(x) + b)$

Duality



- dual representation: (α, b), α is the lagragian and b is the bias
- the vector w can be computed from α: $w = \sum_{i=1}^{m} \alpha_i y_i \phi(x_i)$
- $h(x) = \operatorname{sgn}(w \cdot \phi(x) + b) = \operatorname{sgn}(\sum_{i=1}^{m} \alpha_i y_i k(x, x_i) + b)$
- ϕ is not needed at all
- only $k(x, x')$ appears in the dual representation

The charge rule

The simpliest method for coreceptor usage prediction. (Fouchier et al. 1992)

1. Build a multiple alignment with all sequences
2. Check the (basic) charge of positions 11 and 25 only

Drawbacks

- Some sequences need to be discarded to have a good alignment
- Using only 2 positions reduces the information the data

Other methods

- SVM (support vector machines) with linear kernel
- Random forests
- Neural networks

Issues

Multiple alignments are needed in all cases because those methods need the same amount of attributes for each example. (many sequences have to be discarded to yield a good multiple alignment and therefore we do not use the maximun amount of information.)

Our solution

- SVM with string kernels instead of linear kernels
- We describe a new string kernel: the distant segments kernel

Pros

1. no multiple alignment needed at all.
2. string kernels are natural similarity measures.
3. V3 sequences don't need to be aligned.
4. can be applied to a great number of biologically similar questions

Summary

1. We define a new kernel for HIV-1 coreceptor usage prediction
2. We compare it to existing kernels (data not shown) and we show that multiple alignments are not necessary

The distant segments kernel

Let the following set be the occurances of subsequences of exactly δ symbols beginning with sequence α and ending with α':

$$\mathcal{S}^{\delta}_{\alpha,\alpha'}(s) \stackrel{\text{def}}{=} \{(\mu,\alpha,\nu,\alpha',\mu') : \; s=\mu\alpha\nu\alpha'\mu' \;\; \wedge \;\; 1\leq|\alpha| \;\; \wedge \;\; 1\leq|\alpha'| \;\; \wedge \;\; 0\leq|\nu| \;\; \wedge \;\; \delta=|s|-|\mu|-|\mu'|\}$$

Then, let the mapping function be the size of such sets for many (δ,α,α') :

$$\phi^{\delta_m,\theta_m}_{DS}(s) \stackrel{\text{def}}{=} \left(\left|\mathcal{S}^{\delta}_{\alpha,\alpha'}(s)\right|\right)_{\{(\delta,\alpha,\alpha'):\, 1\leq|\alpha|\leq\theta_m \wedge 1\leq|\alpha'|\leq\theta_m \;\wedge\; |\alpha|+|\alpha'|\leq\delta\leq\delta_m\}}$$

The kernel is the inner product of sequences in feature space.

$$k^{\delta_m,\theta_m}_{DS}(s,t) \stackrel{\text{def}}{=} \langle\phi^{\delta_m,\theta_m}_{DS}(s),\phi^{\delta_m,\theta_m}_{DS}(t)\rangle$$

Comparison for CXCR4

- charge rule (Pillai et al. 2003) : 87.45%
- SVM with linear kernel (Pillai et al. 2003) : 90.86%
- SVM with structural descriptors (Sander et al. 2007): 91.56%
- SVM with distant segments kernel: 94.80%
- **Our method is the only one without multiple alignments!**
- we used a test set to validate our classifier whereas other methods rely on the cross-validation method (which is biaised)

Perspectives

- Sequencing technologies are improving (Roche/454, Illumina/Solexa, ABI SOLiD)
- Machine learning is an emerging science (multiple kernel learning, theoritical risk bounds)
- The next generation of bioinformatic programs for the prediction of HIV-1 coreceptor usage promises improvements for treatment selection in clinical settings.
- Submitted to the journal *Retrovirology*

Acknowledgements

- Mario Marchand, François Laviolette, Jacques Corbeil
- Canadian Institutes of Health Research
- Natural Sciences and Engineering Research Council of Canada
- Canada Research Chair in Medical Genomics
- Los Alamos National Laboratory HIV Databases

Links

- Web server: genome.ulaval.ca/hiv-dskernel
- Our machine learning research group: www.graal.ift.ulaval.ca
- Jacques Corbeil's group: genome.ulaval.ca/corbeillab
- Machine learning course: cours.ift.ulaval.ca/65764
- Kernel methods: www.kernel-methods.net
- Support vector machines: www.support-vector.net