Robert Cedergren Bioinformatics Colloquium 2008

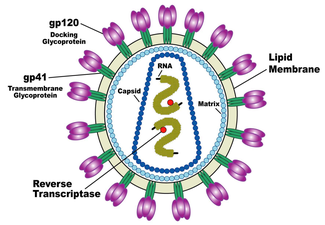
The distant segments kernel and the support vector machine: an alignment-free method for HIV type 1 coreceptor usage prediction

Sébastien Boisvert, Mario Marchand, François Laviolette and Jacques Corbeil
Université Laval

Background

The determination of the coreceptor usage of HIV isolates is critical in clinical settings [1]. Bioinformatics programs geared to tackle this problem brought new perspectives to the field [2].

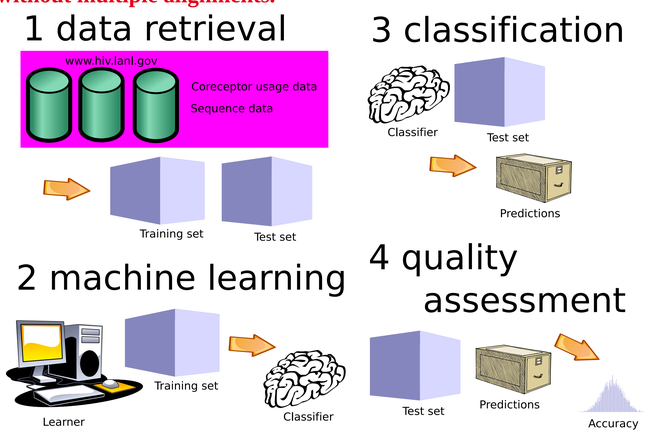
Figure 1: Structural representation of a HIV virion (image from wikipedia).



- Learning algorithms are trained with a multiple alignment of V3 loop sequences (in gp120, see Figure 1) and associated coreceptor usage data.
- Sequences are discarded to have good multiple alignments.

Methods

Figure 2: A bioinformatics workflow for computational prediction of coreceptor usages without multiple alignments.



An example is a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ where x is a V3 sequence and y is the class. The training and test sets contain examples.

Kernels

A kernel is a similarity measure.

$$k : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$$

A mapping function is associated to a kernel.

$$\phi : \mathcal{X} \to \mathcal{R}^n$$

A kernel is a dot product in feature space.

$$k(x, x') = \phi(x) \cdot \phi(x')$$

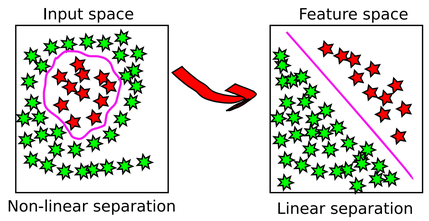
Support vector machines

The support vector machine is a linear separator, i.e.:

$$h(x) = \text{sgn}(w \cdot \phi(x) + b) = \text{sgn}(\sum_{i=1}^{m} \alpha_i y_i k(x, x_i) + b).$$

x is the example to classify, $\{(x_i, y_i)\}_{i=1,\ldots,m}$ is the training set and $\alpha = (\alpha_1, \ldots, \alpha_m)$ is the dual representation.

Figure 3: A linear separation in feature space is a non-linear separation in input space.



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' in protein sequence x:

$$\mathcal{S}^{\delta}_{\alpha,\alpha'}(x) = \{(\mu, \alpha, \nu, \alpha', \mu') : \; x = \mu\alpha\nu\alpha'\mu' \;\wedge\; 1 \le |\alpha| \;\wedge\; 1 \le |\alpha'| \;\wedge\; 0 \le |\nu| \;\wedge\; \delta = |x| - |\mu| - |\mu'|\}.$$

Then, let the mapping function be the size of such sets for many $(\delta, \alpha, \alpha')$:

$$\phi^{\delta_m,\theta_m}_{DS}(x) = \left(\left|\mathcal{S}^{\delta}_{\alpha,\alpha'}(x)\right|\right)_{\{(\delta,\alpha,\alpha') :\, 1 \le |\alpha| \le \theta_m \wedge 1 \le |\alpha'| \le \theta_m \wedge |\alpha| + |\alpha'| \le \delta \le \delta_m\}}$$

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

$$\mathcal{K}^{\delta_m,\theta_m}_{DS}(x, x') = \phi^{\delta_m,\theta_m}_{DS}(x) \cdot \phi^{\delta_m,\theta_m}_{DS}(x')$$

Algorithm 1 solves this computational task with combinatorics.

Algorithm 1: The algorithm for computing the distant segments kernel. s and t are protein sequences. δ_m and θ_m are integer parameters.

DISTANT-SEGMENTS-KERNEL$(s, t, \delta_m, \theta_m)$

$c \leftarrow 0$

FOR any two j_s, j_t such that $s[j_s{+}1] = t[j_t{+}1]$ DO

$n \leftarrow \min(|s| - j_s,\ |t| - j_t,\ \delta_m)$

$k \leftarrow -1$; $i \leftarrow 1$

WHILE $i \le n$ DO

$k \leftarrow k + 1$; $i_{2k} \leftarrow i$

DO $i \leftarrow i{+}1$ WHILE $(i \le n$ AND $s[j_s{+}i] = t[j_t{+}i])$

$i_{2k+1} \leftarrow i$; $l_k \leftarrow i_{2k+1} - i_{2k} + 1$

DO $i \leftarrow i{+}1$ WHILE $(i \le n$ AND $s[j_s{+}i] \neq t[j_t{+}i])$

$c \leftarrow c + \binom{l_0}{3} - 2\binom{l_0-\theta_m}{3} + \binom{l_0-2\theta_m}{3}$

$c \leftarrow c + \min(\theta_m, i_1 - i_0) \cdot \sum_{r=1}^{k} \left(\binom{l_r}{2} - \binom{l_r-\theta_m}{2}\right)$

RETURN c

Results

The SVM equipped with the distant segments kernel performs better than the published algorithms.

Table 1: State-of-the-art classification results. Training set: 1425 sequences. Testing set: 1425 sequences.

Coreceptor usage	Accuracy	Specificity	Sensitivity
CCR5	0.9635	0.8355	0.9875
CXCR4	0.9480	0.9756	0.8768
CCR5 and CXCR4	0.9515	0.9920	0.6589

Table 2: Available methods. The results column contains the metric and what the classifier is predicting.

Reference	Training set	Testing set	Multiple alignments	Results
Pillai et al. 2003	271	-	yes	Accuracy (CXCR4): 0.8745
Resch et al. 2001	181	-	yes	Specificity (X4): 0.9000
Pillai et al. 2003	271	-	yes	Accuracy (CXCR4): 0.9086
Jensen et al. 2003	213	175	yes	Specificity (CXCR4): 0.9600
Jensen et al. 2006	279	-	yes	Specificity (CXCR4): 0.9400
Sander et al. 2007	432	-	yes	Accuracy (CXCR4): 0.9156
Xu et al. 2007	651	-	yes	Accuracy (R5): 0.9510
Lamers et al. 2008	149	-	yes	Accuracy (R5X4): 0.7550
Our method	1425	1425	**no**	Accuracy (CXCR4): **0.9480**

Outlook

We introduced a new string kernel that has broad applicability to various similar problems. We showed that multiple alignments are not necessary and that using more sequences provide opportunities to the learning algorithm to find better solutions. Our approach promises improvements for the application of bioinformatics software to the coreceptor usage prediction in clinical settings.

References

[1] T. Lengauer, O. Sander, S. Sierra, A. Thielen, and R. Kaiser. Bioinformatics prediction of HIV coreceptor usage. *Nat. Biotechnol.*, 25:1407–1410, Dec 2007.

[2] S. Pillai, B. Good, D. Richman, and J. Corbeil. A new perspective on V3 phenotype prediction. *AIDS Res. Hum. Retroviruses*, 19:145–149, Feb 2003.

NSERC CRSNG
UNIVERSITÉ LAVAL
CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada