Statistical Processing of SELEX Results

Damien Eveillard
LORIA / MAEM
Campus Scientifique, BP 239
F-54506 Vandœuvre-lès-Nancy cedex
Damien.Eveillard@loria.fr

Yann Guermeur
LORIA
Campus Scientifique, BP 239
F-54506 Vandœuvre-lès-Nancy cedex
Yann.Guermeur@loria.fr

The experimental technique of *Systematic Evolution of Ligands by EXponential enrichment* (SELEX) is a method from combinatorial chemistry devoted to the identification of ligand molecules that bind by affinity on given target molecules [7]. It is of widespread use and its theoretical properties have been thoroughly investigated [6]. Although the data produced is very useful, its exploitation remains difficult. We have already highlighted in [3] the limits of the standard approach, which consists in deriving one or several consensus sequences from the whole pool of identified ligands. This approach can exhibit two main drawbacks. On the one hand, consensus sequences can be produced which have no biological significance (do not correspond to a motif). On the other hand, some motifs can be neglected.

To overcome these shortcomings, we propose a statistical pre-processing of the SELEX data resulting in a clustering of the ligands. Distinct sets of consensus are then computed per cluster. The approach is assessed on a real case, for protein SR 9G8. This protein was selected since the corresponding ligands are known not to exhibit any characteristic secondary structure [2]. We thus address the most difficult configuration, in which no biological knowledge source is available *a priori* to guide the classification process. Clustering is performed with methods based on different principles. In addition to standard methods of pattern recognition, such as the k-means algorithm and the hierarchical PCA, two original kernel methods are implemented. The first one is an extension of Kohonen's Self-Organizing Maps (SOM) [4], the second one an extension of Support Vector Clustering (SVC) [1]. The specificity of both extensions rests in the nature of the kernel (the measure of similarity), which takes into account the sequential nature of data. All methods produce almost identical results, namely a classification structure made up of two clusters. The consensus sequences associated with each of these clusters are unaffected by small changes in the sets of ligands used, in contrast with what was observed when one single consensus was computed [3]. Grounding our analysis on this statistical evidence, we put forward a conjecture of biological nature, the existence of multiple motifs. This conjecture is validated *a posteriori* by observations resulting from experiments performed independently by another team [5].

This initial series of experiments has been recently extended to proteins for which the secondary structures of the ligands were experimentally determined in the MAEM laboratory. So far, the classification methods have systematically assigned to different clusters the nucleic segments which had different structures. Our statistical pre-processing thus appears consistent with a structural analysis the usefulness of which is clearly established.

References

[1] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, 2001.

[2] Y. Cavaloc, C. F. Bourgeois, L. Kister, and J. Stevenin. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA*, 5:468–483, 1999.

[3] D. Eveillard and Y. Guermeur. Traitement statistique des résultats de SELEX. In *JOBIM'02*, 2002. (to appear).

[4] T. Kohonen. *Self-Organization and Associative Memory.* Springer-Verlag edition, 1989.

[5] F. Lejeune, Y. Cavaloc, and J. Stevenin. Alternative Splicing of Intron 3 of the Serine/Arginine-rich Protein 9G8 Gene. *The Journal of Biological Chemistry*, 276(11):7850–7858, 2001.

[6] F. Sun, D. Galas, and M.S. Waterman. A mathematical analysis of *in vitro* molecular selection-amplification. *Journal of Molecular Biology*, 258:650–660, 1996.

[7] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249:505–510, 1990.