OFFRE de THÈSE

Sujet de thèse : Learning the grammar of genomic sequences

Description :
Using a linguistic approach for modeling genomic sequences has been advocated for a long time by David Searls [1]. Models may sometimes be designed by experts. In the Symbiose team, we study how to automatically design these models by machine learning and we have proposed a successful approach for learning automata on protein sequences [2,3]. The subject of this thesis is to study how this approach can be extended to learn more expressive grammars like in [4,5,6] to model real genomic sequences with long distance correlations.

Département : D7 - Gestion des données et de la connaissance

Equipe : SYMBIOSE; http://www.irisa.fr/symbiose/positions

Directeur de thèse : Jacques Nicolas

Encadrant(s) : François Coste

Contact : francois.coste@irisa.fr ; tel (33|0) 299 847 491

Début des travaux : Dès que possible

Bibliographie :