Bingo*2*- ANR-07-MDCO-014 (2008-2011)

Knowledge discovery For and By inductive queries in post-genomic applications

CGMC CNRS UMR 5534 – GREYC CNRS UMR 6072
LaHC CNRS UMR 5516 – LIRIS CNRS UMR 5205

Objectifs

- **éviter le déluge de motifs produits :**
  - découverte de la connaissance (partielle) du domaine dans des domaines spécifiques
  - utilisation de la connaissance (partielle) du domaine pour la découverte de connaissances à partir de données

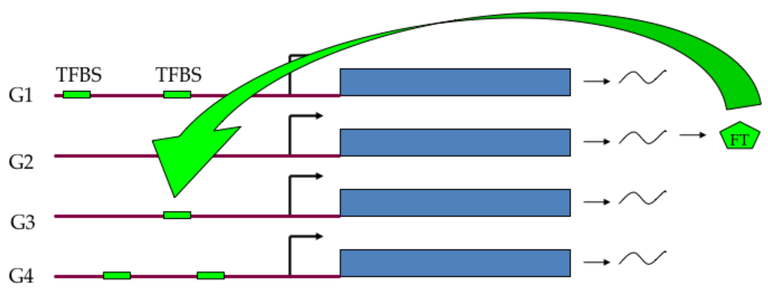
idées forces

- **contribuer à la question de l'origine des contraintes**
- **fournir des résultats sur des applications ciblées**

Exemple de requête

Trouver tous les groupes de synexpression comportant au moins un facteur de transcription parmi les gènes de ce groupe, et comportant, parmi les autres gènes, au moins 50% de gènes comportant dans leur séquence promotrice au moins un site de fixation pour le facteur de transcription co-régulé.

- utilisation de plusieurs sources de données hétérogènes et bruitées
- nécessité de combiner différents prototypes



Nos hypothèses de travail

- le *cadre des bases de données inductives* comme candidat à une théorie de la fouille de données
- cercle vertueux *"théorie/applications"* dans le contexte stimulant de la post-génomique :
  - *trop de motifs* ➠ représentations condensées des collections
  - *données biologiques bruitées* ➠ motifs tolérants au bruit (e.g., motifs locaux " à trous")
  - *sources de données multiples* ➠ concevoir des méthodes de fouille croisant l'information extraite
- *valeur ajoutée :* théorie - méthodes - outils - applications

Consortium

- **CGPHIMC** (Université Claude Bernard, Lyon, équipe BM2A) : expertise sur les données SAGE, applications en biologie, diffusion de SQUAT (Sage Querying and Analysis Tools
- **GREYC** (Université de Caen Basse-Normandie) : découverte de motifs sous contraintes, programmation par contraintes, traitement automatique des langues, algorithmique sur les graphes
- **LaHC** (Université Jean Monnet, Saint-Etienne, équipe "Apprentissage automatique") : inférence grammaticale, apprentissage automatique, programmation logique inductive, fouille de données séquentielles et de textes
- **LIRIS** (INSA de Lyon, équipe Turing) : base de données inductives, fouille de données sous contraintes, fouille de données génomiques et transcriptomiques

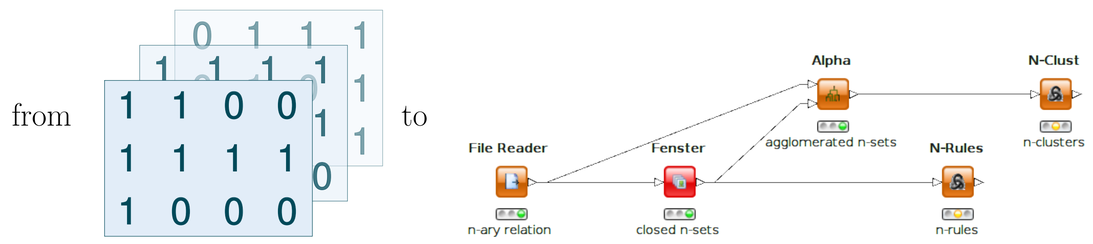
Faits marquants

- publications de grande qualité (e.g., BMC Bioinformatics, In Silico Biology, Data Mining and Knowledge Discovery, Machine Learning, ACM TKDD, SDM, ICDM, CIKM, PAKDD)
- plate-forme SQUAT `http://bsmc.insa-lyon.fr/squat/`
- organisation mini-symposium "data mining"

Résultats

Fouille de relations n-aires

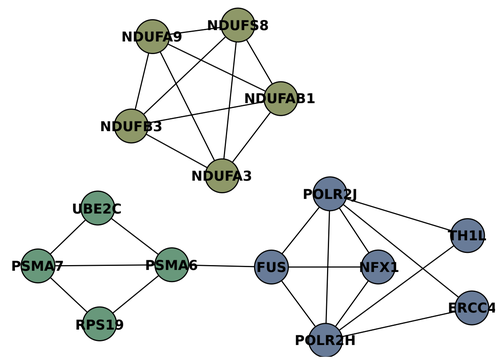
Démarche générale : s'inscrire dans le cadre "Local To Global" de la découverte de motifs **utiles** et de "haut-niveau", généricité des propositions, tolérance aux exceptions



Collections Homogènes de composantes k-cliques Percolées (CoHoP)

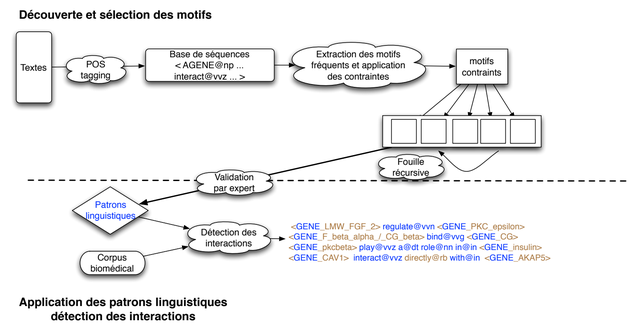
Données : données d'interaction (STRING `http://string-db.org/` + données d'expression

Exemple de résultats : Trois modules de protéines ayant la particularité d'être produits simultanément dans plusieurs situations biologiques



Extraction d'information dans les textes biologiques (interactions géniques et informations biologiques associées)

- *idée directrice :* apports mutuels entre fouille de données séquentielles et traitement automatique des langues
- *méthode :* extraction de motifs séquentiels dans les textes, utilisation de contraintes "linguistiques" et fouille récursive de motifs
- 400 000 textes annotés contenant des interactions : `http://bingotexte.greyc.fr`



Prototypes, plates-formes, applications

- *prototypes :* DATA-PEELER, FITCARE, LSR, SMBIO...
- plates-formes : SQUAT (`http://bsmc.insa-lyon.fr/squat/`), SEDiL (`http://labh-curien.univ-st-etienne.fr/SEDiL/`)

http://bsmc.insa-lyon.fr/squat

Perspectives

données dynamiques (e.g., prise en compte du temps), construction de modèles sous contraintes, comment intégrer les progrès algorithmiques dans un poste de travail pour le biologiste

Contact : Bruno Crémilleux