

Modélisation grande échelle de réseaux biologiques : vérification par contraintes booléennes de la cohérence des données

Philippe Veber

Projet SYMBIOSE, IRISA

17 décembre 2007

Présentation générale

Problématique

Interprétation et analyse des données haut-débit en biologie moléculaire

Spécificité

On étudie des systèmes physiques qui sont par essence quantitatifs à l'aide

- ▶ de mesures bruitées
- ▶ de connaissances partielles et qualitatives

Présentation générale

Approche

formuler des questions à un niveau suffisamment abstrait (graphe d'influence/variations) en s'appuyant sur une **règle de consistance** naturelle, inspirée par nos collaborateurs biologistes

Contributions

- ▶ déterminer un cadre de validité pour la règle de consistance
- ▶ identifier des questions biologiques d'intérêt, les exprimer comme requêtes sur un modèle formel, déterminer leur classe de complexité
- ▶ développer des algorithmes répondant aux requêtes en un temps "raisonnable"

Différentes sources de données

On distingue deux catégories :

Comportements observés

- ▶ puces à ADN,
- ▶ puces à oligonucléotides
- ▶ gels bidimensionnels
- ▶ chromatographie/spectrométrie

Mécanismes moléculaires

- ▶ bases de données d'interactions
 - ▶ RegulonDB, EcoCyc pour la bactérie *E. coli*
 - ▶ IntAct, BIND, TRANSPATH pour l'homme
- ▶ données ChIP-chip
- ▶ données de double-hybride

Comment exploiter conjointement ces informations, en tirer des conclusions pertinentes ?

L'approche de la biologie systémique

Biologie systémique

construction d'un modèle physique du système, souvent à partir de données expérimentales, et dans le but de prédire son comportement

Modèles physiques

- ▶ représentant l'évolution du système dans le temps
- ▶ mécanismes moléculaires (réactions chimiques, régulations)
- ▶ données sur les (variations d') états

Reconstruction de modèle : méthodes probabilistes

= Retrouver un modèle à partir des données/mesures

- ▶ La principale difficulté porte sur **le bruit** présent dans les mesures.
- ▶ Cette observation a motivé le développement d'approches **probabilistes**.
- ▶ Plus précisément, de modèles paramétriques
 - ▶ paramètres pour la structure du système
 - ▶ paramètre pour modéliser le bruit dans les données
- ▶ estimation par un critère d'optimalité
 - ▶ maximum de vraisemblance
 - ▶ critère BIC (Bayesian Information Criterion)
 - ▶ modèle de régression (moindres carrés)
- ▶ ces approches proposent des prédictions, sous la forme de paramètres optimaux (= les plus probables)

Reconstruction de modèle : méthodes probabilistes

Quelques exemples :

- ▶ Réseaux bayésiens (Friedman 2000, Segal 2003)
- ▶ Modèles probabilistes des cascades de régulation (Ideker 2002)
- ▶ Modèles différentiels mixtes génétiques/métaboliques (Palsson 2004)
- ▶ Modèles différentiels linéaires (Bansal 2007)

Limites :

- ▶ problèmes d'optimisation particulièrement difficiles
 - ▶ problèmes d'optima locaux
- ▶ robustesse des prédictions
 - ▶ relation bruit / modèles sous-optimaux
- ▶ modèles quantitatifs, qui requièrent une masse de données rarement disponible ($\#$ variables $\gg \gg$ $\#$ échantillons)

Notion de consistance – Méthodes discrètes

La discrétisation des données/du modèle est une alternative pour :

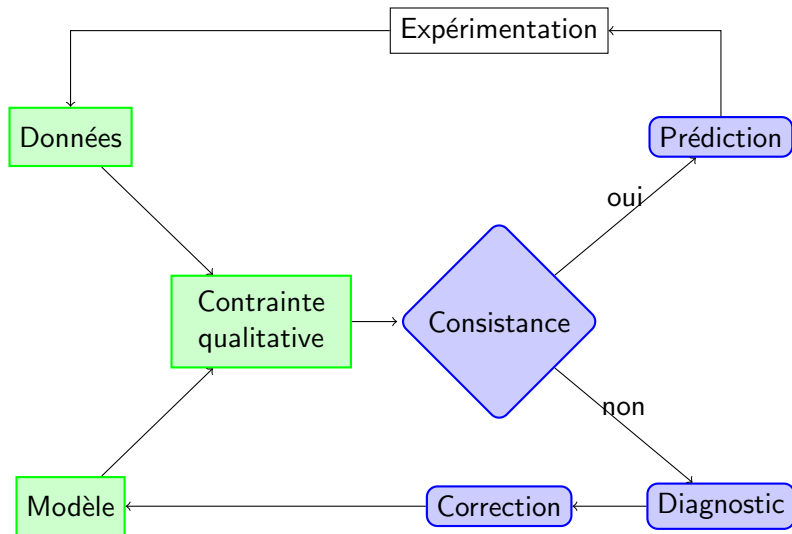
- ▶ traiter des données bruitées
- ▶ raisonner malgré l'incertitude sur les paramètres

Deux approches “orthogonales” :

- ▶ modèles différentiels linéaires par morceaux (Snoussi 1993)
 - ▶ notion d'abstraction discrète, formalisme adapté à une étude fine de “petits” systèmes
- ▶ modèle global pour la régulation transcriptionnelle chez *E. coli*, règles de consistance *ad hoc* (Gutierrez-Rios 2003)

Dans les 2 cas, introduction d'une notion de **consistance** entre des données et un modèle.

Méthodologie



Plan

Problématique

Équations qualitatives et graphe d'interaction

Algorithmes et validation

Un exemple de données d'expression

ex : bactéries *E. coli* sous différents pH (Maurer *et al*
PMID :15601715)

Gène	Niveau d'expression (pH 5)					Niveau d'expression (pH 7)				
	1	2	3	4	5	1	2	3	4	5
agaA	128.7	347.1	344.1	346.6	381.4	620.5	558.5	420.0	393.7	419.2
agaB	12.7	18.8	14.5	35.6	16.3	6.5	17.6	4.6	12.6	16.0
agaC	23.5	66.5	78.1	70.8	71.3	85.6	78.4	53.9	63.5	53.5
agal	51.7	65.1	125.3	116.4	104.0	248.4	104.9	167.7	198.2	175.4
agp	657.3	1019.4	1142.2	1254.3	1060.8	1711.4	1048.8	1551.6	1040.5	1289.6
alaX	6481.9	8344.6	8435.7	7064.6	4838.6	51.9	5855.5	5646.2	6318.4	6245.0
aldA	1588.1	1689.7	1489.3	1494.4	1227.9	1126.5	631.1	553.6	526.6	918.1

Un exemple de données d'expression

ex : bactéries *E. coli* sous différents pH (Maurer *et al*
PMID :15601715)

Gène	pH = 5		pH = 7	
	μ	σ	μ	σ
agaA	309.58	91.48	482.38	90.13
agaB	19.58	8.26	11.46	5.12
agaC	62.04	19.62	66.98	12.98
agal	92.50	28.96	178.92	46.51
agp	1026.80	201.40	1328.38	267.98
alaX	7033.08	1327.25	4823.40	2398.56
aldA	1497.88	153.56	751.18	233.50

Un exemple de données d'expression

ex : bactéries *E. coli* sous différents pH (Maurer *et al*
PMID :15601715)

Gène	pH = 5		pH = 7	
	μ	σ	μ	σ
agaA	309.58	91.48	482.38	90.13
agaB	19.58	8.26	11.46	5.12
agaC	62.04	19.62	66.98	12.98
agal	92.50	28.96	178.92	46.51
agp	1026.80	201.40	1328.38	267.98
alaX	7033.08	1327.25	4823.40	2398.56
aldA	1497.88	153.56	751.18	233.50

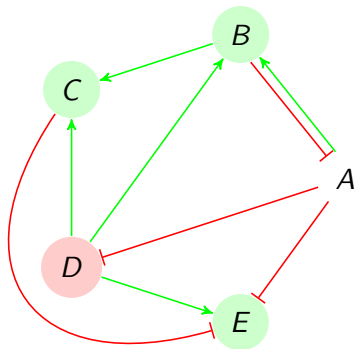
→

Gène	signe de la variation pH = 5 → pH = 7
agaA	+
agaB	0/?
agaC	0/?
agal	+
agp	+
alaX	-
aldA	-

Interprétation en signes de variation

Graphe d'interaction

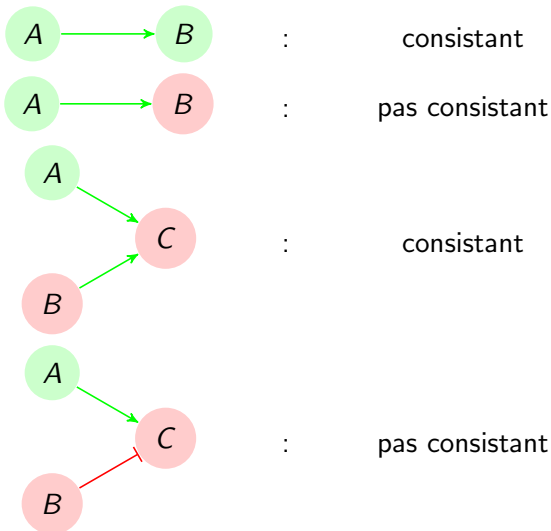
Une représentation courante en génétique et en modélisation



- ▶ sommets = gènes, protéines
- ▶ arcs = régulations
 - $j \rightarrow i$ si la **vitesse** de production de i dépend du **niveau** de j
- ▶ signe = activation/inhibition
- ▶ mesure = étiquetage des sommets
 - ▶  augmentation
 - ▶  diminution

Règle de consistance

la variation de l'expression d'un gène doit être expliquée par la variation d'au moins un de ses prédécesseurs



Énoncé formel : algèbre des signes

L'algèbre des signes est l'ensemble $\mathbb{S} = \{+, -, 0, ?\}$

- muni des opérations $+$ et $\times$:

+	-	0	+	?
-	-	-	?	?
0	-	0	+	?
+	?	+	+	?
?	?	?	?	?

$\times$	-	0	+	?
-	+	0	-	?
0	0	0	0	0
+	-	0	+	?
?	?	0	?	?

- et d'une relation de **compatibilité** $\approx$

$\approx$	-	0	+	?
-	<i>T</i>	<i>F</i>	<i>F</i>	<i>T</i>
0	<i>F</i>	<i>T</i>	<i>F</i>	<i>T</i>
+	<i>F</i>	<i>F</i>	<i>T</i>	<i>T</i>
?	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>

- qui n'est **pas transitive**

$$+ \approx ? \text{ et } ? \approx - \text{ mais } + \not\approx -$$

Règle de consistance : équations qualitatives

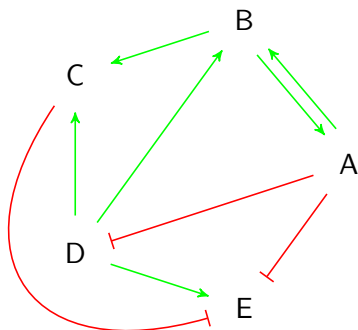
Plus généralement,

$$X_i \approx \sum_{j \in \text{pred}(i)} S_{ji} X_j$$

avec les variables

X_i : variation de l'espèce i durant l'expérience

S_{ji} : effet de la régulation de j sur i (activation/inhibition)



$$A, B, C, D, E \in \{+, -, 0\}$$

$$A \approx B \quad (1)$$

$$B \approx D + A \quad (2)$$

$$C \approx B + D \quad (3)$$

$$D \approx -A \quad (4)$$

$$E \approx D - A - C \quad (5)$$

Justification de la règle de consistance

Objectif : déterminer le cadre de validité de la règle de consistance

La règle de consistance est vérifiée sous les hypothèses :

- ▶ dynamique différentielle
- ▶ état initial et état final **stationnaires** (déplacement d'équilibre)
- ▶ signe des régulations **constants**
 - ▶ *i.e.* indépendants de l'état du système
 - ▶ vrai pour une classe raisonnable de réseaux biochimiques
 - ▶ faux pour certains types de régulation génétique
- ▶ dégradation suffisante

La preuve passe par la démonstration d'une version globale du théorème des fonctions implicites.

[Siegel 2006, Radulescu 2006]

Remarques importantes

- ▶ Cette règle est une **abstraction valide** de toutes les dynamiques admettant un graphe d'interaction donné
 - ▶ Elle sur-approxime les comportements observables du système
 - ▶ C'est donc une condition nécessaire que modèles et données doivent vérifier
- ▶ Les équations qualitatives sont modulaires
 - ▶ hypothèses de validité "locales"

Contrainte qualitative

Rappel :

X_i : variation de l'espèce i durant l'expérience

S_{ji} : effet de la régulation de j sur i (activation/inhibition)

Consistance au sommet i

$$C_i(X, S) = \text{vrai ssi } X_i \approx \sum_{j \in \text{pred}(i)} S_{ji} X_j$$

Contrainte qualitative globale

$$C(X, S) = \bigwedge_{i \in G} C_i(X, S)$$

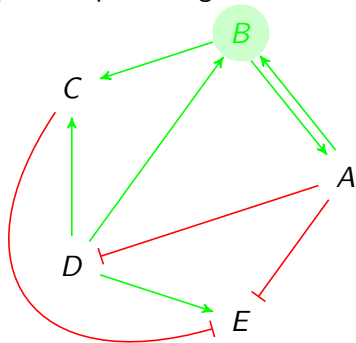
Vérifier la consistance

consistance $\stackrel{\text{def}}{=}$ existence d'une solution à la contrainte qualitative

- ▶ construire la contrainte qualitative à partir des **variations** et des **régulations**
- ▶ en chercher une solution

Prédiction : invariants de l'ensemble de solutions

Supposons que B augmente



Solutions de
 $C(X, S) \wedge (X_B = +)$:

A	C	D	E
A	C	D	E
A	C	D	E

- ▶ A et D ont toujours la même valeur,
- ▶ ce sont des **invariants** de l'ensemble des solutions

Invariants de l'ensemble des solutions

Definition

variable qui prend la même valeur dans toutes les solutions de la contrainte qualitative

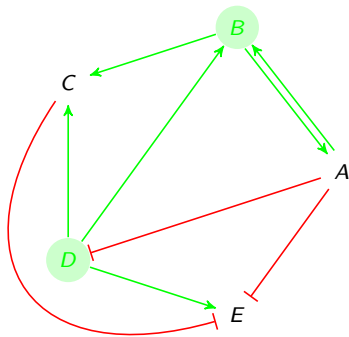
- = conséquence de la contrainte
- = prédiction du modèle

Diagnostic : exemple

- Supposons que B et D ont augmenté
- $C(X, S) \wedge (X_B = +) \wedge (X_D = +)$ n'a pas de solution
- debugger : trouver des sous-ensembles non satisfiables de la conjonction

$$C_A \wedge C_B \wedge C_C \wedge C_D \wedge C_E \wedge (X_B = +) \wedge (X_D = +)$$

- $C_A \wedge C_D \wedge (X_B = +) \wedge (X_D = +)$ n'a pas de solution

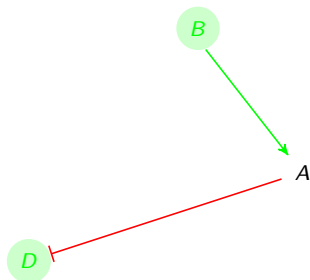


Diagnostic : exemple

- ▶ Supposons que B et D ont augmenté
- ▶ $C(X, S) \wedge (X_B = +) \wedge (X_D = +)$ n'a pas de solution
- ▶ debugger : trouver des sous-ensembles non satisfiables de la conjonction

$$C_A \wedge C_B \wedge C_C \wedge C_D \wedge C_E \wedge (X_B = +) \wedge (X_D = +)$$

- ▶ $C_A \wedge C_D \wedge (X_B = +) \wedge (X_D = +)$ n'a pas de solution



Diagnostic

- ▶ à chaque sommet est associée une contrainte
- ▶ si la conjonction n'a pas de solution
- ▶ chercher des sous-système qui n'ont pas de solution non plus
- ▶ chaque contrainte correspond à un sous-graphe
- ▶ (sommet + prédécesseurs)
- ▶ trouver un “petit” sous-système avec aucune solution
- ▶ minimal au sens de l'inclusion
- ▶ ces sous-systèmes sont appelés défauts

Plan

Problématique

Équations qualitatives et graphe d'interaction

Algorithmes et validation

Approches usuelles pour la résolution

Systèmes d'équations qualitatives utilisés :

- ▶ en intelligence artificielle (Travé-Massuyes 2003) et en physique qualitative (Kuipers 1994)
- ▶ formalisation de raisonnements, aide à la compréhension qualitative de systèmes

La résolution d'un système qualitatif est un problème NP-complet

Résolution étudiée dans le cas linéaire (Dormoy 1988) :

- ▶ en analogie avec les techniques d'algèbre linéaire, règles de résolution des systèmes qualitatifs
- ▶ peu efficace en pratique (système de quelques dizaines de variables)

→ contraintes sur domaine fini

Approches usuelles pour la résolution

Systèmes d'équations qualitatives utilisés :

- ▶ en intelligence artificielle (Travé-Massuyes 2003) et en physique qualitative (Kuipers 1994)
- ▶ formalisation de raisonnements, aide à la compréhension qualitative de systèmes

La résolution d'un système qualitatif est un problème NP-complet

Résolution étudiée dans le cas linéaire (Dormoy 1988) :

- ▶ en analogie avec les techniques d'algèbre linéaire, règles de résolution des systèmes qualitatifs
- ▶ peu efficace en pratique (système de quelques dizaines de variables)

→ contraintes sur domaine fini

Approches usuelles pour la résolution

Systèmes d'équations qualitatives utilisés :

- ▶ en intelligence artificielle (Travé-Massuyes 2003) et en physique qualitative (Kuipers 1994)
- ▶ formalisation de raisonnements, aide à la compréhension qualitative de systèmes

La résolution d'un système qualitatif est un problème NP-complet
--

Résolution étudiée dans le cas linéaire (Dormoy 1988) :

- ▶ en analogie avec les techniques d'algèbre linéaire, règles de résolution des systèmes qualitatifs
- ▶ peu efficace en pratique (système de quelques dizaines de variables)

→ contraintes sur domaine fini

Approches usuelles pour la résolution

Systèmes d'équations qualitatives utilisés :

- ▶ en intelligence artificielle (Travé-Massuyes 2003) et en physique qualitative (Kuipers 1994)
- ▶ formalisation de raisonnements, aide à la compréhension qualitative de systèmes

La résolution d'un système qualitatif est un problème NP-complet
--

Résolution étudiée dans le cas linéaire (Dormoy 1988) :

- ▶ en analogie avec les techniques d'algèbre linéaire, règles de résolution des systèmes qualitatifs
- ▶ peu efficace en pratique (système de quelques dizaines de variables)

Problème algébrique ? → contraintes sur domaine fini

Approches usuelles pour la résolution

Systèmes d'équations qualitatives utilisés :

- ▶ en intelligence artificielle (Travé-Massuyes 2003) et en physique qualitative (Kuipers 1994)
- ▶ formalisation de raisonnements, aide à la compréhension qualitative de systèmes

La résolution d'un système qualitatif est un problème NP-complet
--

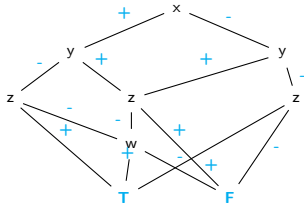
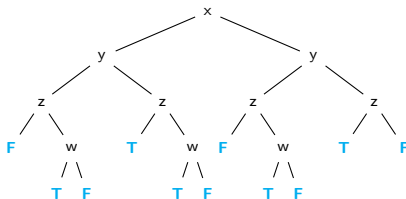
Résolution étudiée dans le cas linéaire (Dormoy 1988) :

- ▶ en analogie avec les techniques d'algèbre linéaire, règles de résolution des systèmes qualitatifs
- ▶ peu efficace en pratique (système de quelques dizaines de variables)

~~Problème algébrique~~ ? → contraintes sur domaine fini

Diagrammes de décision

- ▶ structure de données pour les ensembles de vecteurs booléens [Bryan 1986]
- ▶ utilisés en vérification de circuit, en *model-checking*
- ▶ chaque chemin de la racine à une feuille représente un vecteur
- ▶ chaque sous arbre n'est représenté qu'une fois en mémoire



Construction de l'ensemble des solutions d'une contrainte

Principe

- ▶ construire l'ensemble de toutes les solutions
- ▶ diagramme de décision = **représentation compacte**
- ▶ on construit récursivement le diagramme associé à chaque expression dans le système qualitatif
- ▶ par composition de diagrammes

$X_B - X_C$	$X_D \approx X_A$	$X_A \approx X_B - X_C$	$X_A \approx X_B - X_C$ $X_B \approx X_A - X_C$ $X_D \approx X_A$

Utilisation du diagramme

Une fois le diagramme construit, un **parcours récursif** permet de calculer des propriétés non-triviales de l'ensemble de solutions :

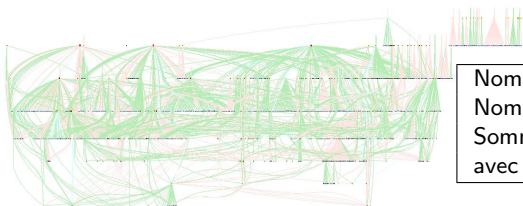
- ▶ existence d'une solution
- ▶ invariants (prédictions)
- ▶ nombre/proportion de solutions (notion plus faible de prédiction)
- ▶ diagnostics/corrections minimales

Application : modélisation de la régulation des acides gras (60 produits)

[Veber 2005, Complexus]

Passage à l'échelle : vérification de RegulonDB

- ▶ base de données dédiée à la **régulation transcriptionnelle** chez *E. coli* (Salgado 2006)
- ▶ contenu : unités de transcription, promoteurs, mécanismes de régulation ...
- ▶ construction **manuelle** par recherche bibliographique
- ▶ **références bibliographiques** pour chaque objet



Nombre de sommets	1258
Nombre de régulations	2526
Sommets sans successeurs	1101
avec plus de 80 successeurs	7

Jeu de données pour la réponse au stress nutritionnel

- ▶ stress nutritionnel → entrée en phase stationnaire
- ▶ jeu de données avec validations expérimentales des variations (RegulonDB)
- ▶ 42 observations sont documentées (avec références)

gene	effect	gene	effect	gene	effect	gene	effect	gene	effect
acnA	+	csiE	+	gadC	+	osmB	+	recF	+
acrA	+	cspD	+	hmp	+	osmE	+	rob	+
adhE	+	dnaN	+	hns	+	osmY	+	sdaA	-
appB	+	dppA	+	hyaA	+	otsA	+	sohB	-
appC	+	fic	+	ihfA	-	otsB	+	treA	+
appY	+	gabP	+	ihfB	-	polA	+	yeiL	+
blc	+	gadA	+	lrp	+	proP	+	yfiD	+
bolA	+	gadB	+	mpl	+	proX	+	yihI	-

La construction ne termine pas, la taille du diagramme excède la place disponible en mémoire

Explication : complexité en espace

- ▶ la taille d'un diagramme croît au pire de manière exponentielle avec le nombre de variables
- ▶ l'efficacité dépend de la taille du diagramme
- ▶ dépend de la **topologie** du graphe d'interaction
- ▶ en général, la limite se trouve entre *100 et 300 variables*

Solution : algorithme de réduction du système

[Bastide 2005, JOBIM]

Définition d'une opération de réduction du système

- ▶ la contrainte réduite possède moins de variables et moins de contraintes
- ▶ toute solution de la contrainte réduite peut-être étendue en une solution de la contrainte originale
- ▶ propriété spécifique de la règle de consistance

Propriétés :

- ▶ réduction du système en temps polynomial
- ▶ extension d'une solution en temps polynomial (à partir d'une solution au système réduit)
- ▶ calcul des invariants en temps polynomial (à partir des invariants du système réduit)

Application à la vérification de RegulonDB

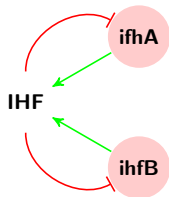
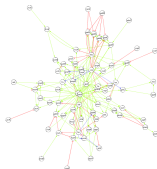
- ▶ le système original comporte 1258 variables et autant d'équations
- ▶ le système réduit n'en comporte plus que 77
- ▶ construction du diagramme en moins d'une minute



Résultat : le graphe n'est pas compatible avec le jeu de données

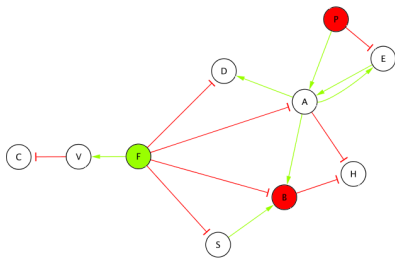
Correction (assistée) du modèle

- ▶ Plusieurs itérations ont permis d'arriver à un modèle et des données consistants
- ▶ Chaque fois, la correction est guidée par la génération d'un **diagnostic** :



- ▶ corrections : ajout de régulations, annotations erronées [Guziolowski 2007, JBPC]

Prédiction de la réponse globale

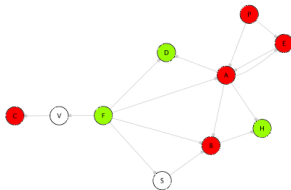


- ▶ à partir des 42 observations pour le stress nutritionnel
- ▶ prédire la réponse dans le reste du graphe
- ▶ valider la prédiction grâce à une mesure d'expression (Allen 2003, J Bacteriol)

Résultats [Guziolowski, en préparation]

- ▶ 526 prédictions dans le réseau
- ▶ dont 163 ont une variation significative sur la puce
- ▶ 80% de prédictions sont en accord
- ▶ comparable avec les résultats obtenus sur des modèles quantitatifs (Palsson 2004)

Nouvelle question biologique : reconstruction de modèle



- ▶ Précédemment, on connaissait les signes sur les arcs, et on souhaitait trouver les signes sur les sommets
- ▶ maintenant, les arcs n'ont plus de signes, et on voudrait les retrouver à partir de plusieurs mesures

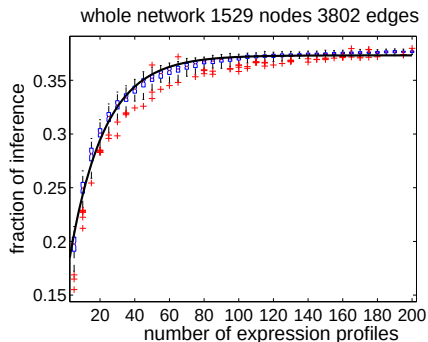
Application : exploitation conjointe des données d'expression et des données de puce, **pour déterminer l'effet d'un facteur de transcription sur ses cibles**

[Veber 2008, accepté pour publication dans BMC Bioinformatics]

Inférence de signes les réseaux de régulation

- ▶ Ce problème est abordé dans [Ideker 2005]
- ▶ Modèle probabiliste, mais également basé sur des règles de signes
- ▶ Limité à l'analyse des expériences de **délétion** de gènes

Dépendance nombre d'observations / nombre de signes inférés



- ▶ étude sur l'exemple du réseau transcriptionnel d'*E. coli*
- ▶ évaluation empirique, par simulation
- ▶ 18% des signes inférables à partir d'une seule mesure
- ▶ 30 mesures suffisent pour arriver à 30% de signes inférés
- ▶ plateau apparent vers 37%

Algorithme pour l'inférence de signes

- ▶ Ces systèmes qualitatifs comportent plusieurs dizaines de milliers de contraintes et environ 3500 variables
- ▶ Cette fois, la réduction ne simplifie pas le système
- ▶ Dans le cadre d'une collaboration avec l'université de Potsdam, nous avons proposé d'employer des techniques de résolution de contraintes booléennes (Answer Set Programming)

Answer Set Programming (Programmation par ensemble réponse)

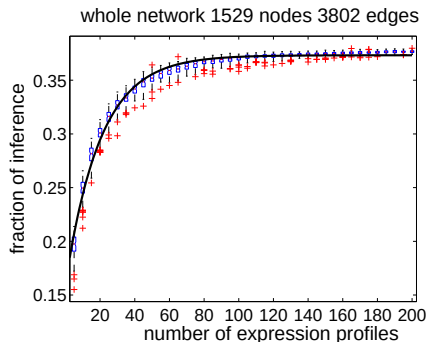
- ▶ Paradigme de programmation logique
- ▶ Peut être vu comme la confluence de deux efforts de recherche
 - ▶ fournir une “bonne” sémantique pour les programmes de logique non monotone, permettant de raisonner sur des données incomplètes,
 - ▶ les progrès considérables effectués autour de la résolution du problème SAT
- ▶ À cela s'ajoute la recherche de langages facilitant la spécification de problèmes combinatoires ou de planification

Utilisation d'ASP pour les contraintes qualitatives

Stratégie

- ▶ description du problème à l'aide du langage
 - ▶ modèles du programme logique $\leftrightarrow$ solutions de la contrainte
 - ▶ résolution par un moteur dédié (c1asp, Université de Potsdam)
 - ▶ fournit **une** solution
-
- ▶ Permet de répondre aux trois types de requêtes
 - ▶ Cette deuxième méthode nous a permis de tester la correction des deux approches
 - ▶ La résolution est particulièrement efficace : quelques minutes au plus pour les systèmes liés à l'inférence de signe

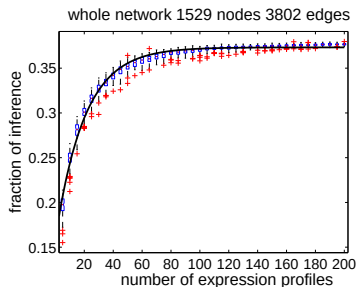
Limite de la courbe



- ▶ Calcul de la hauteur du plateau ?
- ▶ Revient à déterminer quels sont les signes inférables à partir d'un nombre arbitraire d'observations
- ▶ **Problème de Σ_2^P complet**
- ▶ ne peut être résolu en ASP "normal"
- ▶ trop de variables pour une approche directe avec les diagrammes de décision

Solution : décomposition de systèmes qualitatifs

- ▶ découpage en sous-systèmes partageant un minimum de variables
- ▶ algorithmes pour vérification/prédiction s'adaptant à cette décomposition en modules
- ▶ la recherche d'une décomposition optimale est elle-même un problème NP-complet
- ▶ mais l'optimalité n'est pas cruciale



Résultat : la limite se situe à 40,8%

Comparaison avec l'approche d'Ideker *et al*

Données expérimentales sur la levure

- ▶ 210 expériences de délétion [Hugues 2000]
- ▶ données ChIP-chip [Lee 2002]
- ▶ Graphe d'interaction : 2419 sommets et 4344 régulations
- ▶ L'approche d'Ideker *et al* produit 234 prédictions
- ▶ Notre approche produit 162 prédictions
- ▶ Intersection de 17 régulations
 - ▶ les résultats concordent
 - ▶ prédictions d'Ideker *et al* dans les parties denses du réseau
 - ▶ nos prédictions se situent principalement sur les sommets "cibles"

Conclusion

Nos contributions

- ▶ un formalisme et une méthodologie génériques pour la mise au point de **modèles étendus** en biologie moléculaire
- ▶ des outils algorithmiques adaptés à la masse de données disponible
- ▶ des applications non triviales de ces démarches sur données réelles

Points spécifiques de notre approche

- ▶ la définition d'une notion générale de consistance données/modèle
- ▶ la définition de diagnostics associés
- ▶ une approche adaptée aux cas où le ratio observations/nombre de variables est défavorable

Perspectives

Traîtement des données bruitées

- ▶ Prédiction malgré la présence d'inconsistances
- ▶ Aide à la correction/reconstruction de régulations

Autres notions de consistance

- ▶ Introduire une notion de causalité dans les équations
- ▶ Fonctions de régulations

Plans d'expérience

- ▶ Contrôle du système
- ▶ Discrimination efficace de modèles

Remerciements

IRISA

Michel Le Borgne Anne Siegel
Rumen Andonov Ovidiu Radulescu
Carito Guziolowski

INSERM

Nathalie Théret Jeremy Gruel

INRA

Sandrine Lagarrigue

Université de Potsdam

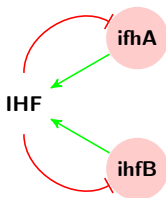
Torsten Schaub Martin Gebser
Steve Dworschak Sven Thiele

Publications

- ▶ Philippe Veber, Michel Le Borgne, Anne Siegel, Sandrine Lagarrigue, and Ovidiu Radulescu. Complex qualitative models in biology : A new approach. *Complexus*, 2(3-4) :140–151, 2004.
- ▶ Y Bastide, S Lagarrigue, ML Borgne, A Siegel, P Veber, O Radulescu, and AL Behec (2005). Une méthodologie pour l'analyse qualitative des réseaux biologiques : De la base de données à la vérification formelle JOBIM 2005, poster session
- ▶ A Siegel, O Radulescu, M Le Borgne, P Veber, J Ouy, and S Lagarrigue. Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks. *Biosystems*, 84(2) :153–174, May 2006
- ▶ Ovidiu Radulescu, Sandrine Lagarrigue, Anne Siegel, Philippe Veber, and Michel Le Borgne. Topology and static response of interaction networks in molecular biology. *J R Soc Interface*, 3(6) :185–196, Feb 2006.
- ▶ C Guziolowski, P Veber, M Le Borgne, O Radulescu, and A Siegel (2006). Checking Consistency Between Expression Data and Large Scale Regulatory Networks : A Case Study In : Réseaux d'interaction : analyse, modélisation et simulation. RIAMS'06, Lyon, France.
- ▶ P. Veber, C. Guziolowski, M. Le Borgne, O. Radulescu, A. Siegel. Inferring the role of transcription factors in regulatory networks. Accepted for publication in *BMC Bioinformatics*

Vérification : premier round

- ▶ les données ne sont pas compatibles avec le graphe
- ▶ diagnostic :

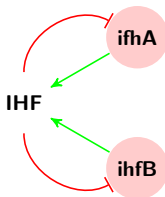


$$\begin{array}{lll} C_{IHF} : & X_{IHF} & \approx -X_{ihfA} - X_{ihfB} \\ C_{ihfA} : & X_{ihfA} & \approx X_{IHF} \\ C_{ihfB} : & X_{ihfB} & \approx X_{IHF} \end{array}$$

- ▶ le sous-graphe n'admet qu'une variation nulle (pas de boucle positive)

Vérification : premier round

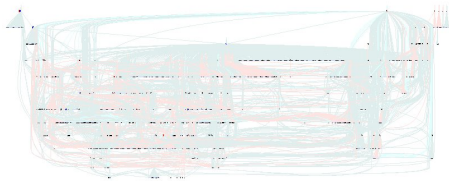
- ▶ les données ne sont pas compatibles avec le graphe
- ▶ diagnostic :



$$\begin{array}{llll} C_{IHF} : & X_{IHF} & \approx & + \\ C_{ihfA} : & - & \approx & X_{IHF} \\ C_{ihfB} : & - & \approx & X_{IHF} \end{array}$$

- ▶ le sous-graphe n'admet qu'une variation nulle (pas de boucle positive)

Correction : introduction des facteurs σ

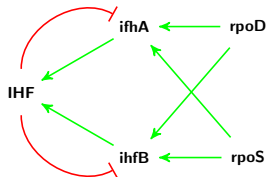


Number of nodes	1529
Number of interactions	3385
TF with more than 80 targets	14
sigma factors	7
protein complex	4

- ▶ les facteurs σ sont des protéines
- ▶ qui régulent la transcription de nombreux gènes
- ▶ facteurs globaux ($\rightarrow$ réponse synchronisée à une condition particulière)
- ▶ les facteurs σ sont spécifiques d'une condition

Protéine	Gène	Gènes induits
σ^{70}	rpoD	1047
σ^{38}	rpoS	114
σ^{54}	rpoN	100
σ^{24}	rpoE	48
σ^{32}	rpoH	29
σ^{19}	fecI	7

Correction du sous-système

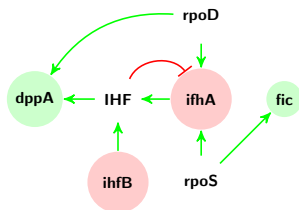


$$\begin{aligned}C_{IHF} : \quad X_{IHF} &\approx -X_{ihfA} - X_{ihfB} \\C_{ihfA} : \quad X_{ihfA} &\approx X_{IHF} + X_{rpoS} + X_{rpoD} \\C_{ihfB} : \quad X_{ihfB} &\approx X_{IHF} + X_{rpoS} + X_{rpoD}\end{aligned}$$

- ▶ les facteurs σ sont considérés comme des entrées
 - ▶ leur régulation n'est pas expliquée dans le modèle
- ▶ à présent ihfA et ihfB admettent des variations non nulles

Vérification : deuxième round

- ▶ les données et le modèle amendé sont encore **non compatibles**
- ▶ diagnostic :



$C_{IHF} :$	X_{IHF}	$\approx$	$-X_{ihfA} - X_{ihfB}$
$C_{ihfA} :$	X_{ihfA}	$\approx$	$X_{IHF} + X_{rpoS} + X_{rpoD}$
$C_{fic} :$	X_{fic}	$\approx$	X_{rpoS}
$C_{dppA} :$	X_{dppA}	$\approx$	X_{rpoD}

Annotations ?

- ▶ dans RegulonDB, *ihfA* and *ihfB* sont annotés comme **diminuant** à l'entrée en phase stationnaire
- ▶ **mais** les références associées affirment qu'ils **augmentent**

*Transcriptional control of the *himA* and the *himD/hip* genes coding for the two subunits of the integration host factor (IHF) was investigated. [...] Expression from both promoters was found to increase as the cells enter stationary phase.*

NCBI PubMed A service of the U.S. National Library of Medicine and the National Institutes of Health www.pubmed.gov

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

Display AbstractPlus Show 20 Sort By Send to

All: 1 Review: 0

1: Mol Microbiol. 1994 Dec;14(5):1021-31.

Expression of the genes coding for the Escherichia coli integration host factor are controlled by growth phase, *rpoS*, *ppGpp* and by autoregulation.

Aviv M. Giladi H. Schreiber G. Oppenheim AB. Glaser G.

Department of Cellular Biochemistry, Hebrew University-Hadassah Medical School, Jerusalem, Israel.

Transcriptional control of the *himA* and the *himD/hip* genes coding for the two subunits of the integration host factor (IHF) was investigated. The promoters for the two genes were identified by the use of primer extension and S1 analysis. Expression from both promoters was found to increase as the cells enter stationary phase. Mutation in *rpoS*, known to be induced upon entry to stationary phase, dramatically reduced the growth-phase response of the *himA* P4 promoter but had only a small effect on the induction of the *himD/hip* promoter. The increased activity of both promoters required the presence of the *relA* and *spoT* genes, suggesting that *ppGpp* plays a major role in the response to stationary phase. An artificial increase in *ppGpp* in exponentially growing cells induced a rapid increase in

Vérification : troisième round

- ▶ correction : ihfA et ihfB augmentent suite à un stress nutritionnel
- ▶ cette fois, les données (corrigées) sont compatibles avec le graphe (étendu)

Retour