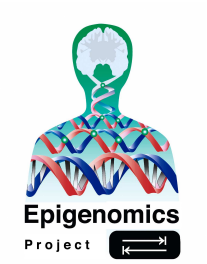
Quelques questions informatiques apportées par la génomique

Gilles Bernot

Programme Épigénomique
Maison Genopole® des Sciences de la Complexité

LaMI UMR 8042
CNRS – Université d'Évry
& Genopole®



Informatique et génomique

Traitement de séquences

— assemblage et annotation

— comparaisons de séquences

— applications (phylogénie, etc)

BD, données à haut débit, extraction de connaissances

— Entrepôts de données, BD hétérogènes

— ontologies

— fouille de données, classification

— extraction de modèles

Modélisation, simulation

— comportements & objets émergeants

— abstraction, graphes, sémantique dymanique

— multi-échelles

— observabilité, prédictivité & « retour à la paillasse »

Bioinformatique à Genopole®

– Généthon : AFM

– Centre National de Séquençage / Génoscope

– Centre National du Génotypage

– Génoplantes

– LaMI / Laboratoire de Méthodes Informatiques

– Maison des sciences de la complexité : Programme Épigénomique

– Statistique & génome

– Maladies multifactorielles

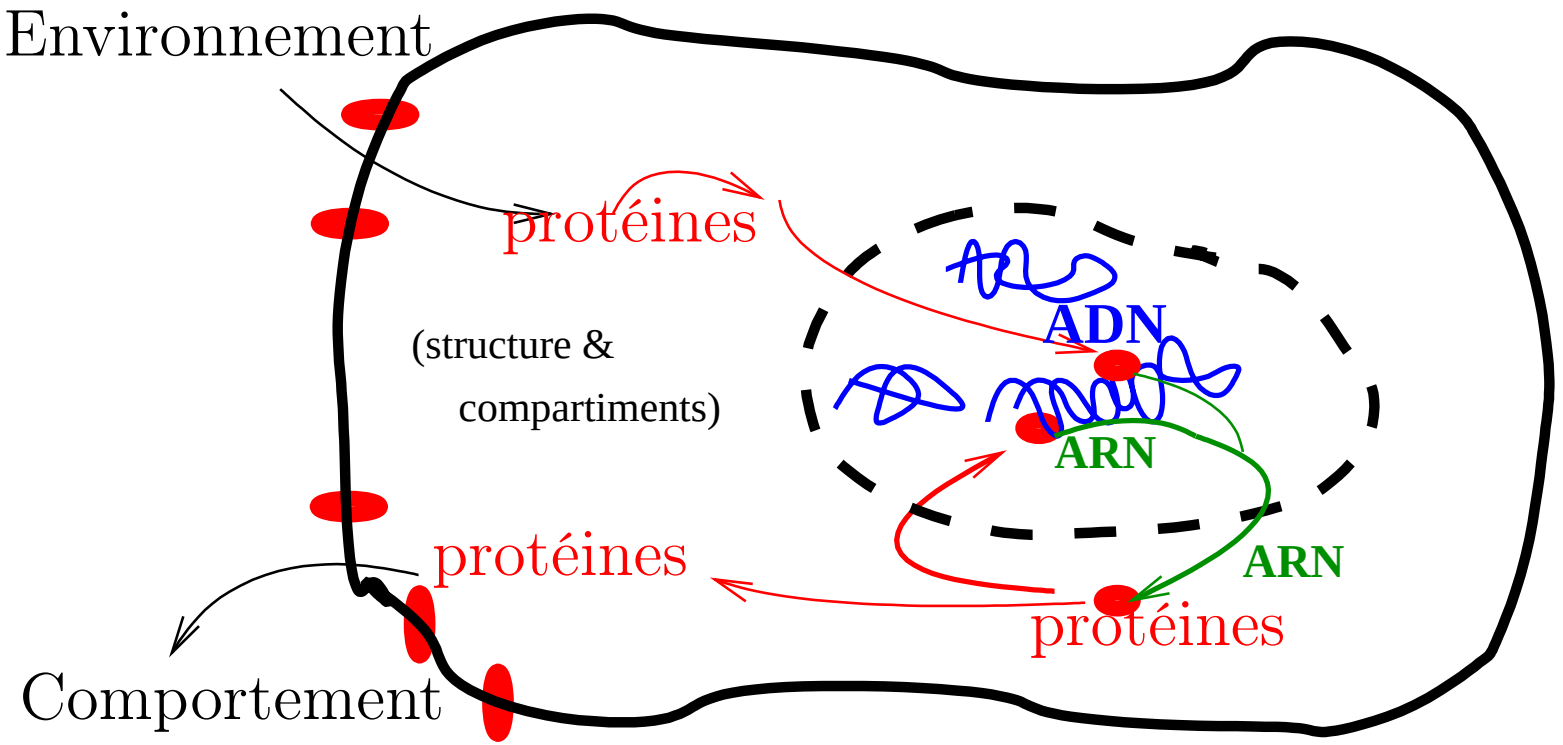
– Génome et informatique

– Annotation des génomes

– Infobiogen

– Genset-Serono, ...

Biologie moléculaire de la cellule

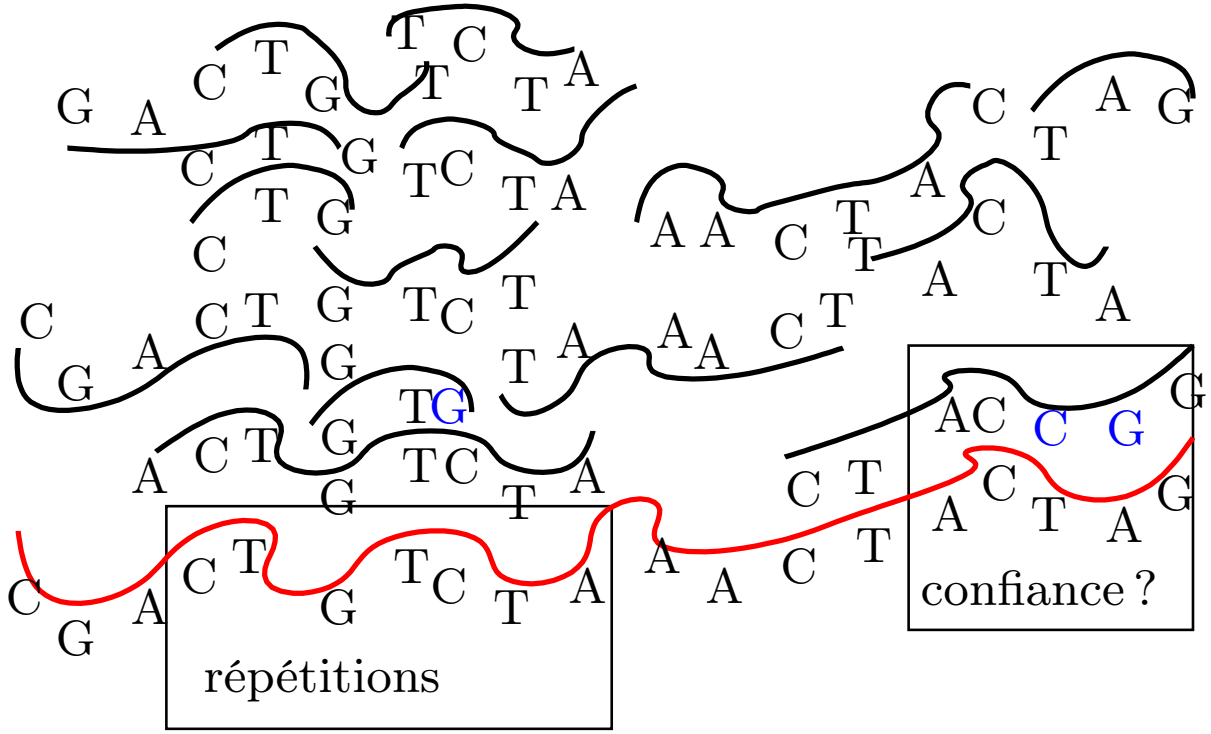


Génome – Transcriptome – Protéome –

Voies métaboliques – Réseaux de régulation –

Compartiments – Fonctions – Phénotype

Assemblage (ADN : ATGC)

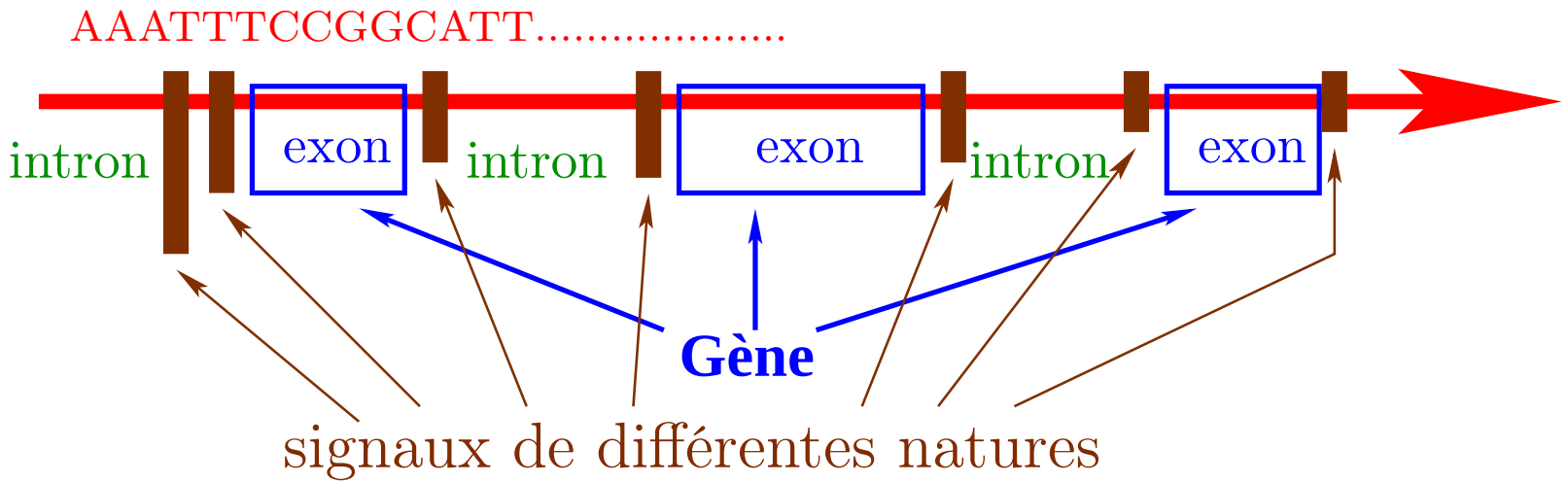


Informations entre 2 séquences : *niveau de similarité, distance approximative, orientations relatives, cartographie,* etc.

Existence d'une solution ? Unicité ? ?

Reconstruction « la plus crédible »

Annotation du génome (ADN)



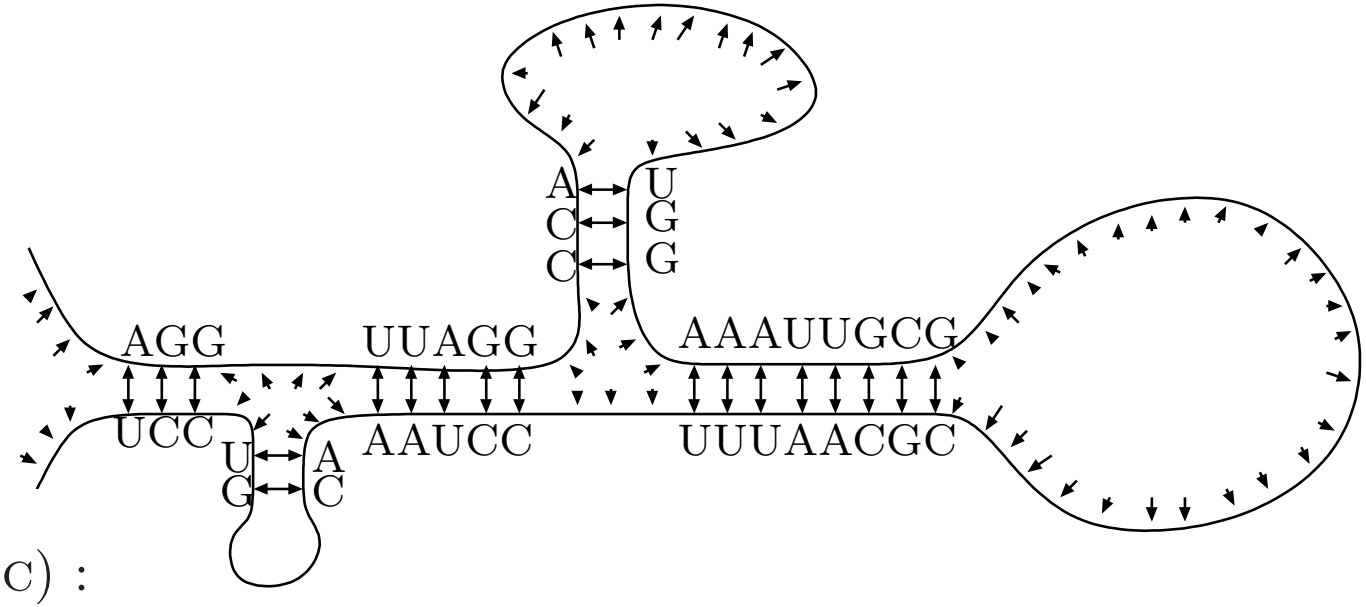
Objectif = retrouver les exons formant un gène (≈ code pour une protéine), leurs promoteurs, ...

– *ab initio* : expressions régulières approximatives, chaînes de Markov, etc.

– Par comparaisons : avec des gènes connus, par conservation inter-espèces, ...

Note : liens avec la compression de séquences.

Prédiction de Structures



ARN (AUGC) :

Chercher palindromes (A↔U et G↔C) + nœuds plus complexes

Protéines (acides aminés) : Quelques principes de mise en vrilles (hélices α) ou autres formes régulières connues (feuillets β)

BD de formes connues et recherches par similarités de séquences

Phylogénie

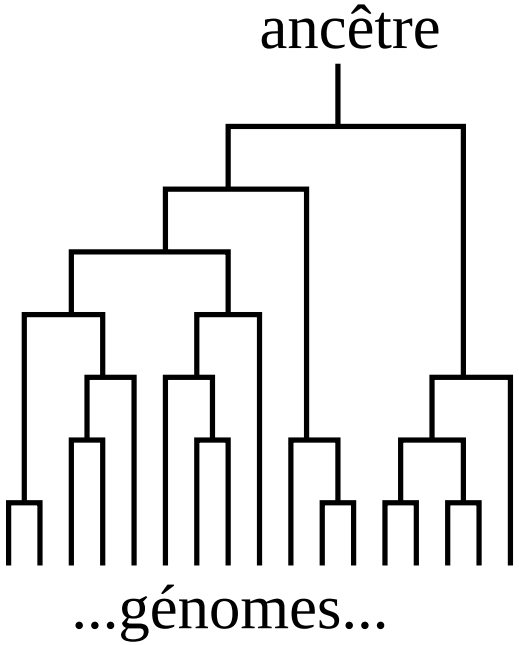
Évolutions des espèces par mutations du génome.

— inventaire des mutations élémentaires (insertion, délétion, permutation...)

— structure algébrique des mutations

— calculs de distances entre génomes

— fabrication d'un arbre phylogénique



Classification, Clustering

Divers « types d'objets » à classifier :

– gènes de diverses espèces (suites ADN)

– portions « typées » selon normes d'extraction (ORF, EST...)

– ... foison de subtiles variations ...

À la base :

– indices de similarité de séquences, alignement

– fonctions similaires (phénotype)

– représentation dans un espace vectoriel et découpage des nuages de points

Problèmes majeurs :

– l'arrivée incrémentale des données peut modifier les clusters, donc l'organisation des bases

– pas d'historique d'une base de clusters (glissements de barycentre)

– capture des rares expertises biologiques

Biopuces (= puces à ADN), méthode SAGE

Mesurent l'ARN produit par les gènes.

- on soumet une population homogène à un stress
- on mesure $\frac{\textit{population étudiée}}{\textit{population témoin}}$ à des temps successifs bien choisis,

sur un ensemble de gènes choisi

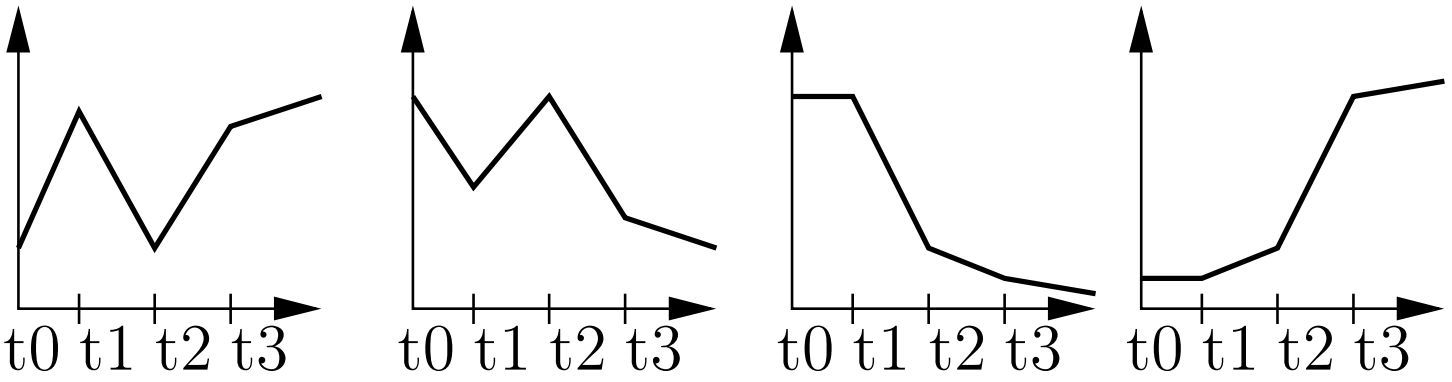
1. Acquisition des données
   (analyse d'image)
2. Base de données
3. Profils d'expression

Buts : - appréhender la dynamique des réseaux génétiques
- étudier les réactions à un stress / une maladie
- diagnostic médical

$\Longrightarrow$ croiser de nombreuses expériences

Profils d'expression

Un grand nombre de mesures par gène en parallèle :



Des conditions d'expériences précises :

- protocoles expérimentaux
- images d'origine
- « recettes » de normalisation entre images
- choix des sondes
- choix des clusters de gènes
- etc.

Modélisation en génomique

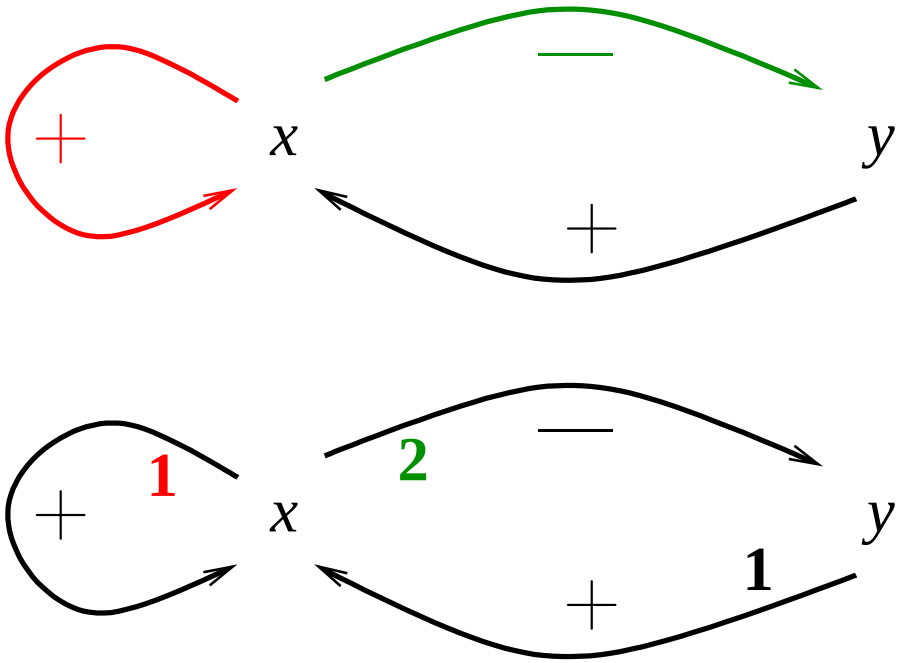
Modéliser pour :

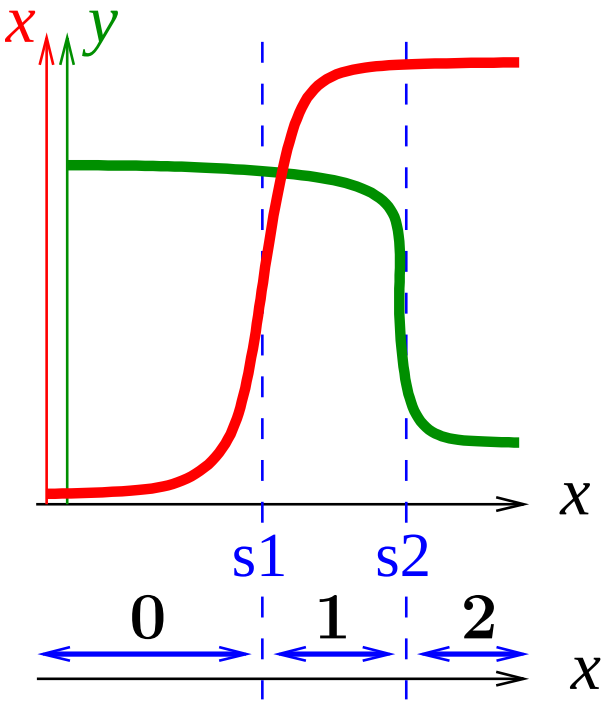
- abstraire et comprendre
- réviser des idées reçues contradictoires
- intégrer de nombreuses connaissances lacunaires
- suggérer des expériences « humides »
- minimiser leur coût et leur nombre
- effectuer des expériences « in silicio »
  impossibles *in vivo* ou *in vitro*

$\Longrightarrow$ Prédictivité...

Réseaux de régulation (René Thomas)

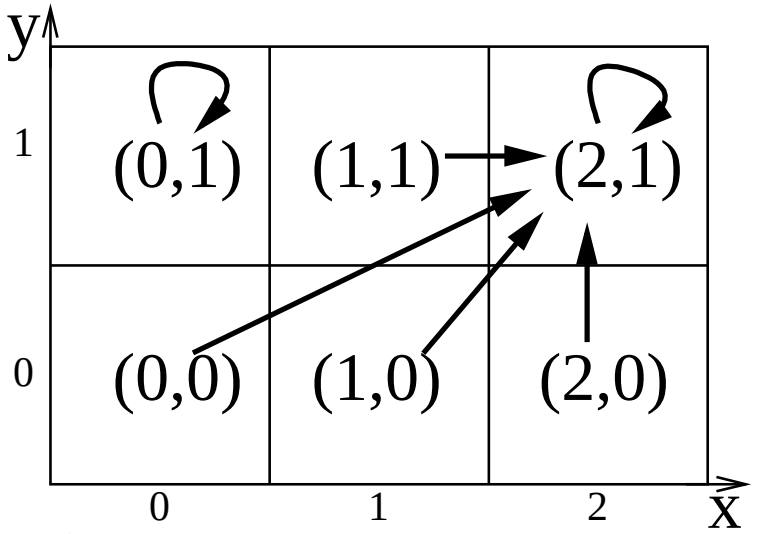
Abstraire par une variable : un gène, l'ARN
et la protéine pour laquelle il code





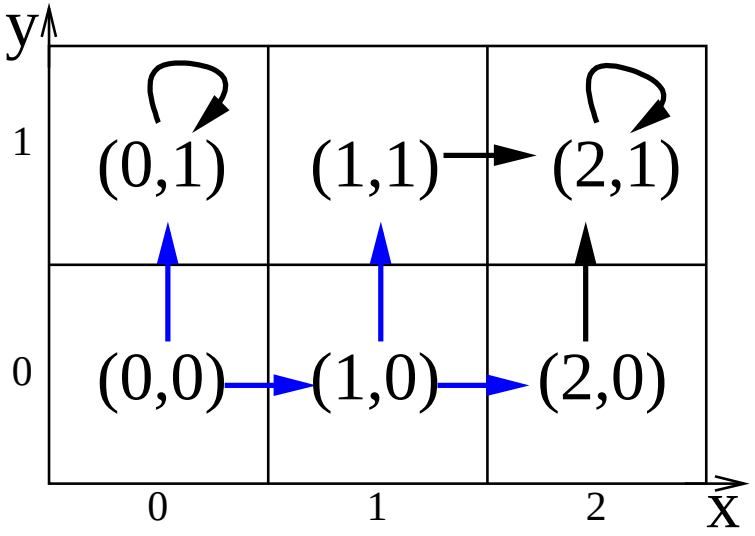
Réseau de régulation → Graphe d'états

(x,y)	*Attracteur*
(0,0)	$(K_{x,\overline{y}}, K_y)$=(2,1)
(0,1)	(K_x, K_y)=(0,1)
(1,0)	$(K_{x,x\overline{y}}, K_y)$=(2,1)
(1,1)	$(K_{x,x}, K_y)$=(2,1)
(2,0)	$(K_{x,x\overline{y}}, K_{y,x})$=(2,1)
(2,1)	$(K_{x,x}, K_{y,x})$=(2,1)



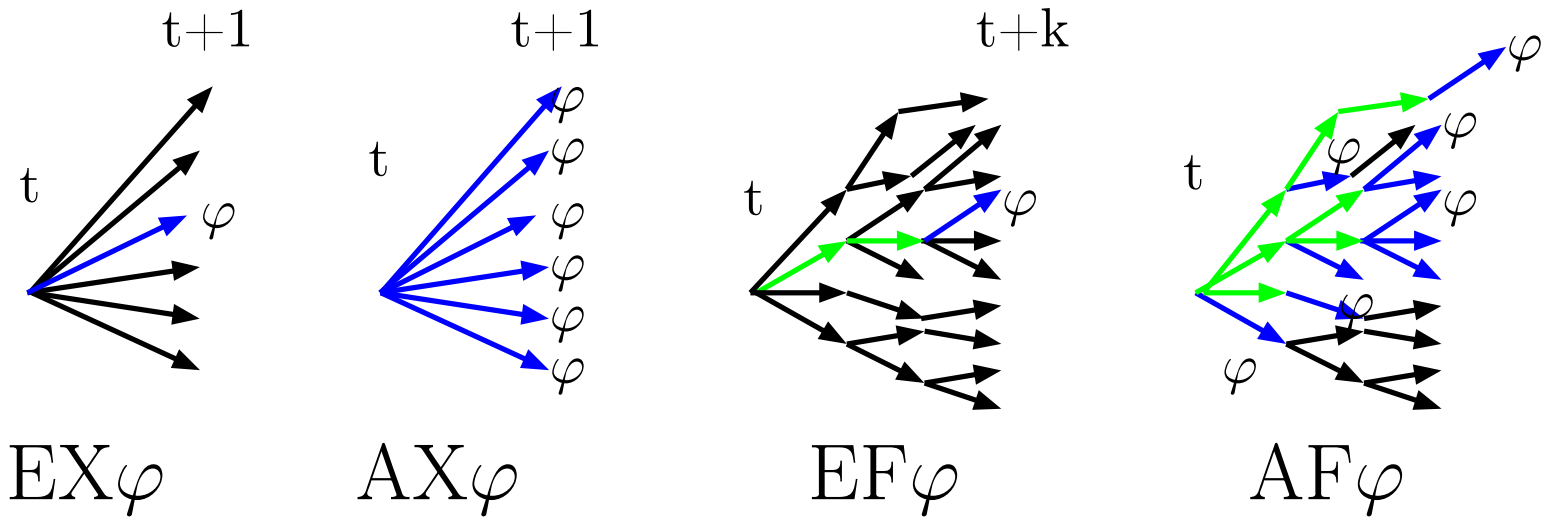
« désynchronisation » ⟶

par unités de distance Manhattan

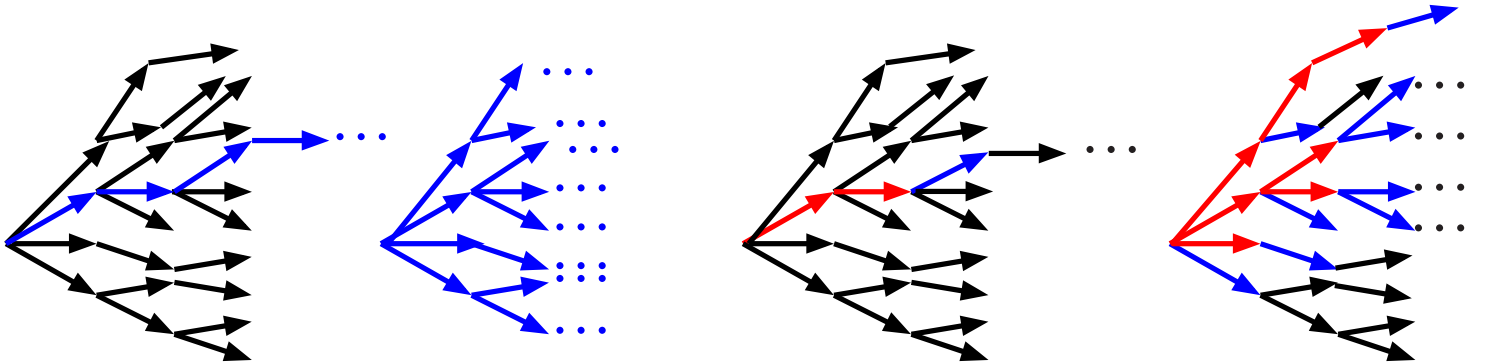


Arbre des traces, etc.

Sémantique de CTL sur les traces



$\mathrm{EX}\varphi$ $\mathrm{AX}\varphi$ $\mathrm{EF}\varphi$ $\mathrm{AF}\varphi$



$\mathrm{EG}\varphi$ $\mathrm{AG}\varphi$ $E[\psi U \varphi]$ $A[\psi U \varphi]$

Formule CTL = Lien modèles–expériences

Les formules ainsi construites sont valides ou invalides par rapport à un ensemble de traces donné partant d'un état donné

Elles peuvent être confrontées à l'ensemble des traces possibles du modèle théorique

Elles peuvent être confrontées à l'ensembles des expériences connues

Elles font donc le lien entre modèles et objets biologiques

Mucoviscidose et *P. aeruginosa*

Mucoviscidose (humain) $\Longrightarrow$ phénotype mucoïdie (P.aeruginosa)
$\Longrightarrow$ problème respiratoire (humain) & biorésistante (P.aeruginosa)

Modification de phénotype, terminologie :

modification génétique : héritable et non réversible (mutation)
modification épigénétique : héritable mais réversible
adaptation : non héritable et réversible

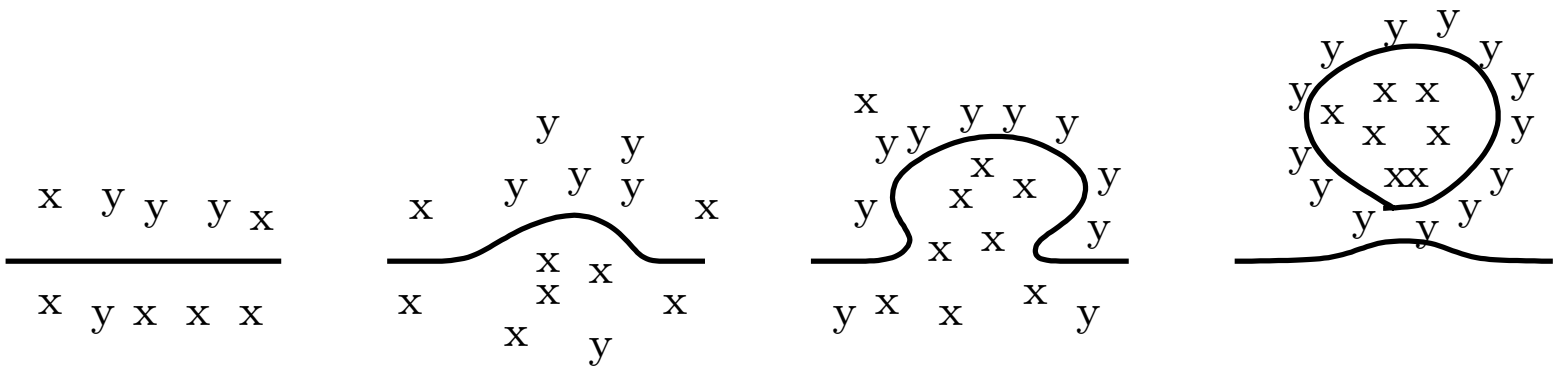
La question biologique (Janine Guespin) :

la synthèse de mucus chez la bactérie *Pseudomonas aeruginosa* est-elle une modification génétique ou épigénétique ?

FAIT : après stabilisation des populations productrices de mucus, beaucoup sont mutantes.
Malgré cela, la modélisation montre que le phénomène est épigénétique.

Déformation de membranes

et auto-recrutement de molécules :



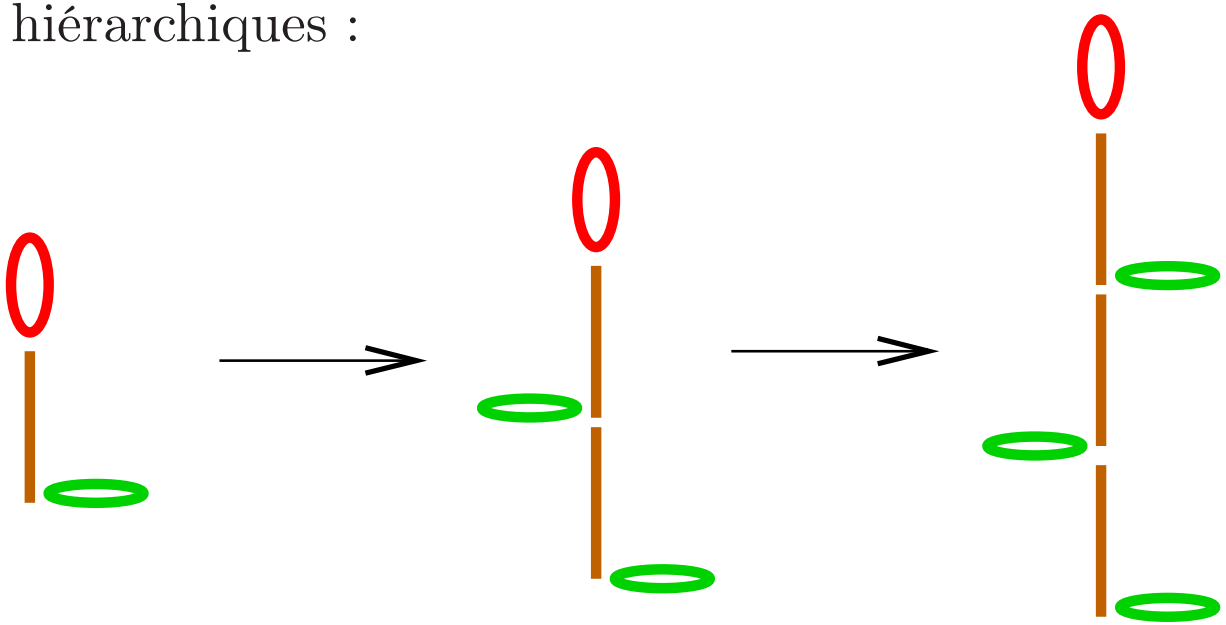
modèlisation par équations différentielles.

x et y s'accumulent par affinités

$\Longrightarrow$ « manteau » et formation de vacuole

Croissance des plantes

par descriptions hiérarchiques :



systèmes de réécriture
avec liste de patterns élémentaires et structurés.

Plus généralement : automates cellulaires, approches multi-agents, etc.

→ émergence d'hyperstructures, de phénotypes, . . .

Modélisation multi-niveaux

- Plusieurs techniques de modélisation
- Plusieurs vues d'un même objet biologique
- Plusieurs niveaux d'organisation biologique
- Plusieurs niveaux de raffinement des modèles
- Connaissances « à trous »
  $\Longrightarrow$ faire feu de tous bois
  $\Longrightarrow$ approches bottom-up ou top-down impossibles
- Nombreux phénomènes d'émergence
- Création dynamique d'objets biologiques nouveaux
- Forte limitation de « l'observabilité » biologique
- Problème des artefacts induits par la technique de modélisation

Modélisation → Simulation

- Ergonomie de la visualisation
- Rendre les modèles exécutables
- Quels résultats présenter ?
- Sous quelle forme ?
- Dangers des implicites faux, induits par l'interface
- Bien distinguer et montrer
  . les connaissances biologiques
  . les hypothèses dûes au modèle
  . les déductions « sûres »
  . les choix aléatoires lors de l'exécution
  . etc.

Conclusion : la complexité

D'une forte combinatoire de phénomènes élémentaires émergent des comportements complexes robustes

La complexité est incontournable en biologie moléculaire et l'informatique, en tant que science, peut y apporter des solutions

- les connaissances biologiques sont par essence parcellaires

- en très grand nombre

- seule l'aide de l'informatique peut tester la cohérence des extrapolations biologiques.

- importance des approches « multi-vues » et « multi-niveaux »

- mélanger de manière cohérente : connaissances, déductions, extrapolations. . .