Cours n°1

1 Du bon usage de l'IA

L'Intelligence Artificielle recouvre plusieurs techniques différentes et il faut savoir choisir le bon outil pour l'usage qu'on veut en faire.

- Les techniques qui furent à la base de l'IA dans les années 80 (typiquement les systèmes experts et la preuve de théorèmes assistée par ordinateur) relèvent maintenant d'un domaine appelé *l'IA symbolique*. Ce type d'IA repose actuellement beaucoup sur la logique et est utile lorsque l'on a besoin d'une *explication* sur les affirmations faites par IA. L'IA *explicable* est devenu un sujet majeur, fortement d'actualité.
- Les techniques probabilistes sont typiquement utiles lorsqu'on souhaite faire de la *classification*, par exemple pour identifier les gènes co-exprimés en réponse à un stress, ou pour regrouper des phénotypes selon des critères donnés. Ce type d'IA repose beaucoup sur le *clustering* où les données sont traduites en nombres de sorte que l'IA navigue dans un espace de grande dimensions.
- Les techniques dites « bio-inspirées » comprennent essentiellement deux classes : celles qui s'inspirent des réseaux de neurones et celles qui s'inspirent de l'évolution par mutations successives. Les *réseaux de neurones profonds* (« deep learning ») existent depuis longtemps mais les puissances de calcul actuelles ont rendu l'IA très populaire depuis quelques années. Leur domaine d'application a d'abord été l'analyse d'images (*e.g.* en imagerie médicale). Plus récemment *les LLM* (Large Language Models) sont spécialisés dans le traitement des textes et ont rendu l'interface homme-machine utilisable par le grand public.
- L'autre classe de techniques bio-inspirées sont les *algorithmes évolutionnaires*. Ils reposent, comme pour les approches probabilistes, sur une représentation de l'ensemble des solutions possibles à un problème au moyen d'un espace de grande dimension. Il s'agit alors, partant de solutions choisies plus ou moins aléatoirement, de leur faire subir des « mutations » et de sélectionner à chaque « génération » les solutions les meilleures. Ceci jusqu'à converger vers des solutions correctes.

Comme on le voit, si les LLM sont particulièrement adaptés pour « converser » avec l'utilisateur, ce ne sont pas toujours les techniques d'IA les plus adaptées pour la biologie. D'une part, les mesures expérimentales habituelles en biologie conduisent à des données qui « vivent » dans un espace de grande dimension, donc à appliquer des techniques probabilistes ou des techniques bio-inspirées. D'autre part, les connaissances biologiques ne peuvent être enrichies que si l'on dispose d'explications rationnelles sur telle ou telle affirmation, ce qui conduit à appliquer les techniques d'IA symbolique.

Il résulte de ce survol rapide des techniques d'IA que l'essentiel pour tirer parti de l'IA est de bien *préparer les données* qu'on lui soumet en fonction du type de problème qu'on veut résoudre. Cette étape préalable est cruciale.

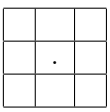
Dans tout ce cours, on va aborder la *modélisation de systèmes biologiques* car la biologie des systèmes est un pan majeur des biotechnologies, par conséquent elle est indispensable à tout ingénieur en biologie.

Les sections suivantes abordent deux techniques de modélisation où les données utilisées sont rigoureusement codifiées, donc manipulables par ordinateur. On verra que les *paramètres* d'un modèle *in silico* sont un point d'articulation majeur pour prédire les comportements d'un système biologique. L'IA, en biotechnologie, sert essentiellement à trouver les valeurs crédibles de ces paramètres.

2 Les automates cellulaires

Un *automate cellulaire* découpe l'espace considéré en « morceaux » disjoints tous identiques. Ainsi une droite sera découpée en intervalles de longueur fixée, un plan peut être découpé en carrés, et l'espace en cubes. Chaque *case* de l'espace considéré est appelé une « cellule », d'où la terminologie « automate cellulaire » (rien à voir avec la notion biologique de cellule donc).

Les *voisins* d'une cellule en 2D sont alors au nombre de 8 :



On considère par ailleurs un ensemble d'états possibles d'une case. Par exemple une case peut être coloriée et la couleur est alors l'état de la case. Ou encore une case peut être vide ou pleine, ou bien encore vivante ou morte, *etc.*

On fait évoluer dans le temps le contenu des cases avec pour convention que la transformation d'une case ne dépend *que* de l'état de ses cases voisines.

2.1 Règles de transformations en 2D et jeu de la vie

Une *règle de transformation* est de la forme $\begin{array}{|c|c|c|}\hline x_1 & x_2 & x_3 \\ \hline x_4 & \text{c} & x_5 \\ \hline x_6 & x_7 & x_8 \\ \hline\end{array} \longrightarrow \boxed{\text{c'}}$

L'exemple le plus connu d'automate cellulaire est le *jeu de la vie* où une cellule peut être vivante ou morte :
— une cellule morte possédant exactement trois voisines vivantes devient vivante (elle naît) ;
— une cellule vivante possédant deux ou trois voisines vivantes le reste
— dans tous les autres cas, elle meurt.
Certaines formes évoluent alors de manière remarquable. Par exemple un « planeur » dans le jeu de la vie est un motif ayant la forme suivante :

v	v	-
v	-	v
v	-	-

où une case vide représente une cellule morte et une case vivante contient un « v ».
En simulant à la main l'évolution d'un planeur dans un plan dont toutes les autres cellules sont mortes, on redécouvre l'intérêt de cette appellation... Le faire !
En vous connectant sur Wikipédia, consultez le jeu de la vie et entre autres le canon à planeurs.

2.2 Modélisation de feu de forêt

En pratique, les automates cellulaires sont utiles pour prédire des phénomènes de propagation (épidémies, rumeurs et autres fake-news, *etc*).

Ici, on va modéliser la propagation d'un feu de forêt avec un découpage tel qu'il y a un arbre par case, comme dans une exploitation industrielle de bois, et on découpe le temps en heures.
Un arbre pas encore atteint par le feu sera dit *sain*. À l'inverse, un arbre ayant totalement brûlé sera dit *calciné*, et il n'est donc plus en feu. Les règles de l'automate sont les suivantes :
— Un arbre sain, lorsqu'il est atteint par le feu, brûle pendant 3 heures avant de devenir calciné.
— Un arbre sain ayant au moins 2 arbres en feu dans son voisinage devient en feu.

On est donc amené à considérer 5 états possibles d'une case : sain (s), première heure de feu (f1), deuxième heure de feu (f2), troisième heure de feu (f3) et calciné (c).

Exercice : Tracez l'évolution des états successifs de l'automate dont l'état initial est le suivant :

s	s	c	c
f2	f1	s	c
f3	f1	c	s
c	s	c	s

Cette simulation visualise l'intérêt de la technique bien connue du contre-feu qui sauve les 2 derniers arbres sains.

Pour un modèle plus proche de la réalité qui puisse aider les pompiers à établir leur stratégie de contrôle du feu, il faut ajouter une direction préférentielle due au vent, améliorer les règles de passage en feu par une approche probabiliste plutôt que de simplement compter le nombre de voisins, *etc*. Ces éléments constituent les *paramètres* mentionnés dans la section précédente et ce sont eux qui peuvent être approchés par IA en regard du comportement des feux de forêt passés...

3 Les modélisations à base de règles

Dans ce type de modélisation, l'espace n'est pas découpé en cases, de sorte que les règles sont écrites d'une manière très proche des réactions chimiques.

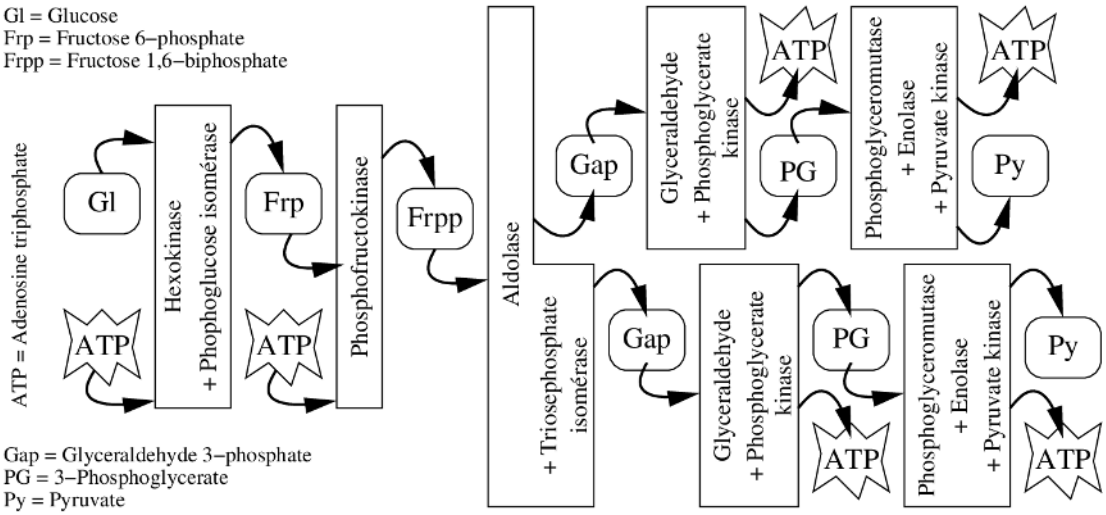
Ici on va s'intéresser à un logiciel de simulation, appelé HSIM, qui représente un espace 3D entouré d'une membrane. L'objectif de ce cours n'est pas d'apprendre à écrire les règles que comprend HSIM en particulier mais de comprendre le rôle de la simulation pour déduire les paramètres d'un système biologique complexe.

Si l'on modélise des réactions, ces paramètres sont essentiellement les vitesses de ces réactions :
— elles sont, en pratique, impossibles à déterminer *in vitro* car les conditions sont trop différentes de ce qui se passe *in vivo* ;
— cependant elles ne sont généralement pas mesurable directement *in vivo* car trop intriquées avec d'autres phénomènes ou car ces mesures perturberaient trop le fonctionnement sauvage de la cellule.

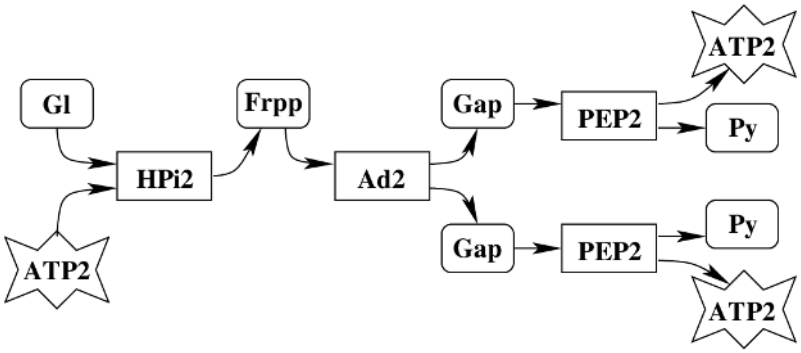
La simulation est donc un moyen alternatif peu coûteux pour prédire des vitesses qui reproduisent fidèlement les phénotypes observables : à partir d'un grand nombre de jeux de vitesses potentielles, les simulations *in silico* conduisent à des trajectoires du système plus ou moins éloignées des phénotypes observés *in vivo*. Les méthodes d'IA de la section précédente permettent alors de prédire quels sont les classes de jeux de vitesses « crédibles », c'est-à-dire dont le comportement simulé est le plus proche des phénotypes observés biologiquement.

La page web de ce cours montre quelques systèmes très simples, simulés, pour lesquels les paramètres ont été déterminés de cette façon.

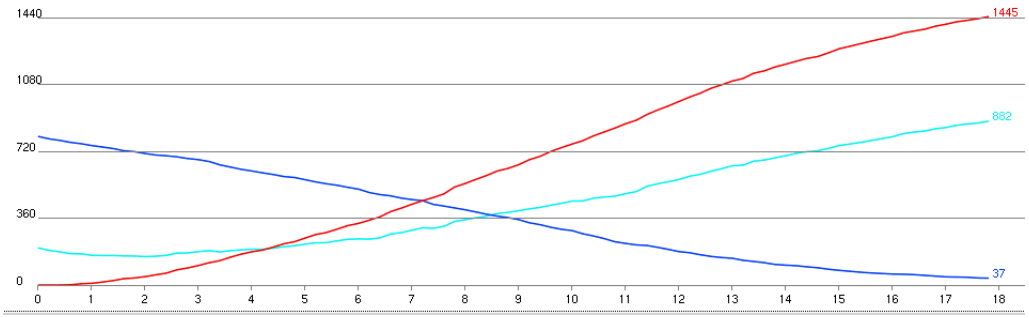
D'autres exemples plus crédibles sont bien sûr abordés par les bio-informaticiens. Pour mentionner un grand classique, *la glycolyse*, chacun sait que la cellule doit d'abord « investir » de l'ATP pour lancer la glycolyse et que la glycolyse en produit seulement plus tard.
Pour focaliser la modélisation sur ce phénomène, on simplifie la suite de réaction glycolytiques comme suit, en regroupant les réactions intermédiaires qui ne concernent pas l'ATP :



De plus, une version encore plus simplifiée regroupe les ATP par paires :

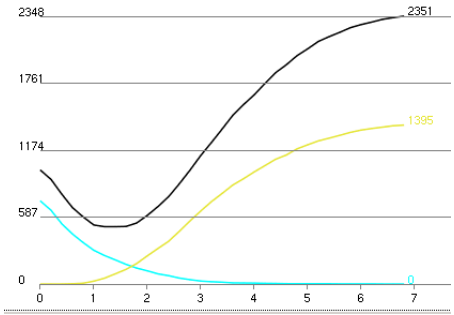


Lorsqu'on lance l'IA sur la version la plus simplifiée en cherchant à exhiber au mieux la nécessité d'investir de l'ATP avant d'en récupérer, le résultat est quelque peu décevant, puisqu'on n'observe qu'une très faible baisse d'ATP au début :



L'ATP « double » est en bleu ciel, le glucose en bleu et le pyruvate en rouge

En revanche, sur la première version, le résultat est net :



L'ATP est en noir, le glucose en bleu ciel et le pyruvate en jaune

Cela montre que le niveau de description d'un système biologique influe sensiblement la qualité des prédictions que l'on peut obtenir. L'IA ne fait pas tout. Les connaissances biologiques pertinentes sont la base de modèles crédibles, eux mêmes nécessaires pour des prédictions réussies. Cela constitue l'essentiel de la bio-informatique pour des avancées notables en biotechnologies.