Jean-Loup Faulon

La rétrosynthèse en biologie : production de molécules bioactives et dispositifs pour le diagnostic

Jean-Loup Faulon est directeur de recherche, responsable de l'équipe Bio-RetroSynth à l'Institut Micalis[1] de l'Institut national de la recherche agronomique (INRA), et professeur en biologie de synthèse au département de chimie de l'Université de Manchester (Royaume-Uni). Il est responsable du master biologie des systèmes et de synthèse de l'Université Paris-Saclay et a développé avec son équipe des méthodes de rétrosynthèse pour concevoir et mettre en œuvre des voies de biosynthèse et biodégradation dans le cadre de l'ingénierie métabolique et de l'ingénierie de biosenseurs cellulaires et acellulaires.

1 La rétrosynthèse en biologie : état de l'art

La rétrosynthèse[2] est bien connue en synthèse organique et nous allons ici en voir les applications dans le domaine de la biologie et plus particulièrement de la biologie de synthèse.

1.1. Quelques applications en biologie

C'est pour l'ingénierie métabolique que la rétrosynthèse a d'abord été appliquée en biologie. L'ingénierie métabolique utilise des souches (aussi appelées des châssis) pour y introduire des enzymes hétérologues[3] de façon à synthétiser une molécule cible. Lorsqu'on pratique la

1. www.micalis.fr/Institut-Micalis
2. Rétrosynthèse : technique qui consiste à retrouver des voies de synthèse en partant de la molécule finale. Le principe de cette méthode d'analyse s'appuie sur des ruptures des liaisons, pour constituer des molécules plus simples.
3. L'expression hétérologue est l'expression d'un gène ou d'un fragment de gène dans un organisme hôte, qui ne possède pas naturellement le gène ou son fragment.

rétrosynthèse, on part d'une molécule cible, et on applique des réactions enzymatiques de façon à remonter jusqu'aux métabolites[4] qui sont naturellement produits par la souche utilisée (***Figure 1***).

Ce même concept peut être utilisé pour pratiquer l'ingénierie de biocapteurs (ou biosenseurs). Ici, le problème est de modifier une cellule de telle sorte qu'elle soit capable de détecter une molécule, par exemple un métabolite. Une molécule est généralement détectée directement par une cellule *via* une interaction allostérique (fixation de la molécule induisant un changement de conformation spatiale de l'enzyme) avec un facteur de transcription[5] ou un riborégulateur[6]. Une fois la molécule cible détectée, on exprime un gène reporteur, par exemple un marqueur fluorescent (protéine GFP, « *Green Fluorescent Protein* »). L'ensemble facteur de transcription, riboregulateur et gène reporteur constitue un biocapteur.

Le nombre de molécules directement détectables par des facteurs de transcription ou des riborégulateurs est faible. L'idée est alors d'utiliser la rétrosynthèse

4. Métabolite : composé stable issu de la transformation biochimique d'une molécule initiale par le métabolisme.

5. Facteur de transcription : protéine nécessaire à l'initiation ou à la régulation de la transcription d'un gène dans l'ensemble du vivant. Elle interagit avec l'ADN et l'ARN polymérase.

6. Riborégulateur (ou « riboswitch ») : structure d'ARN présente sur un ARN messager (ARNm) qui peut lier directement un ligand. Très souvent, le ligand du riborégulateur est un métabolite de la réaction catalysée par la protéine codée par l'ARNm, ce qui conduit à un mécanisme de rétroaction directe. Cette fixation déclenche un effet sur l'expression du gène porté par l'ARNm en bloquant ou en activant la traduction de la protéine correspondante. L'utilisation des riborégulateurs est ainsi une des voies possibles de régulation de la traduction.

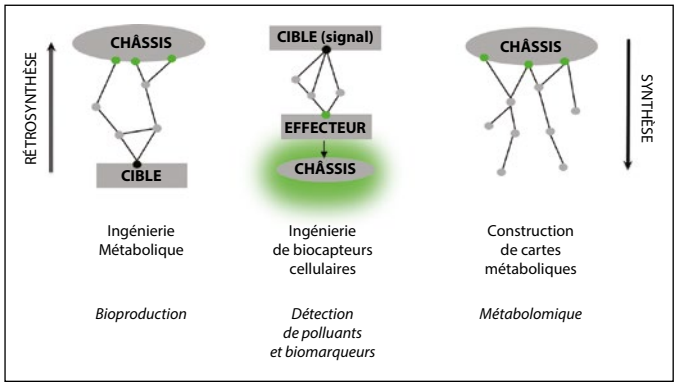


Figure 1

Différents exemples d'applications de la rétrosynthèse en biologie.

pour rechercher l'ensemble des molécules accessibles. On étend de cette façon le nombre de molécules détectables par des cellules.

La troisième application abordée dans ce chapitre est liée au fait qu'on ne connaît pas tous les métabolites des souches utilisées en biotechnologie. Les méthodes de rétrosynthèse peuvent aider à trouver de nouveaux métabolites dans nos souches et rechercher les enzymes responsables de la synthèse de ces métabolites.

Dans ce chapitre, nous passerons en revue les trois types d'applications de la ***Figure 1*** qui font toutes appel à des méthodes de rétrosynthèse.

1.2. La rétrosynthèse : un problème étudié depuis longtemps mais toujours d'actualité

La rétrosynthèse a été développée par Elias James Corey dans le cas de la synthèse organique, ce qui lui a valu le prix Nobel de chimie en 1990. Il a proposé plusieurs méthodes de mise en œuvre, notamment la technique LHASA (***Figure 2***), un logiciel développé à la fin des années 1960. L'idée est de partir d'une molécule cible, la molécule que l'on veut synthétiser, et d'appliquer des règles de déconnexion des liaisons dans cette molécule cible jusqu'à remonter à des molécules disponibles ou qu'on sait synthétiser.

Dans sa présentation lorsqu'il a reçu le prix Nobel, Elias Corey parlait d'intelligence artificielle. C'était prémonitoire car à partir de 2016, une série d'articles s'inscrivant dans cette démarche ont été publiés (***Figure 3***) : ils utilisent des méthodes d'apprentissage profond pour calculer automatiquement les règles de déconnexion dont parlait Elias Corey, et plus généralement pour proposer des règles pour les réactions utilisées en synthèse organique.

Les méthodes d'apprentissage profond sont d'autant performantes qu'elles reposent sur un grand nombre de données. Le nombre de réactions connues en chimie organique est considérable : le Chemical Abstract Service contient environ 80 millions de réactions. Dans les applications en biologie, on en est plutôt à 30 000 ou 40 000 réactions stockées dans les bases de données, l'intelligence artificielle est toutefois utilisée en particulier pour la recherche de séquences enzymatiques, comme présenté dans le paragraphe 2.2.

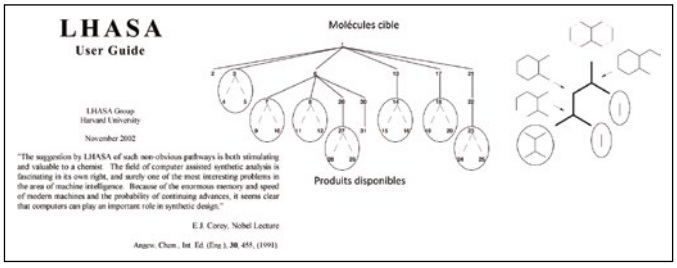


Figure 2

Modélisation du processus utilisé par le logiciel LHASA. L'objectif est de partir d'une molécule cible et de remonter progressivement vers des fragments connus en synthèse organique.

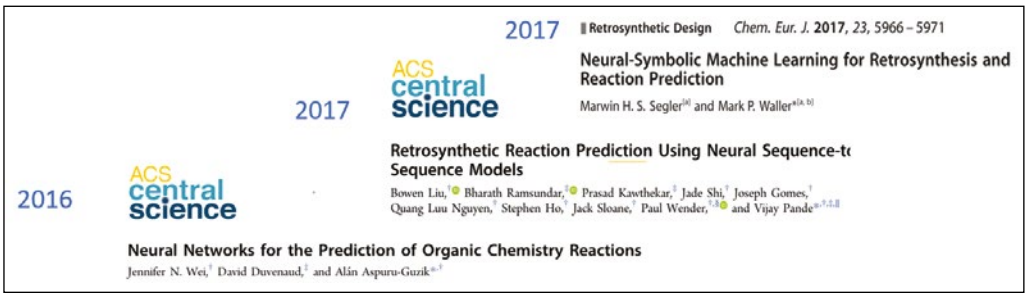


Figure 3

Série d'articles publiés entre 2016 et 2017 qui s'intéressent à l'application de la rétrosynthèse pour la chimie organique. L'utilisation de la chimie computationnelle est de plus en plus développée pour la prédiction des réactivités. Aujourd'hui, certains groupes s'attachent au développement d'une intelligence artificielle pour la rétrosynthèse.

1.3. État de l'art des méthodes de rétrosynthèse appliquées en biologie

Un certain nombre de groupes de recherche développent les méthodes de rétrosynthèse pour la biologie. Trois principales méthodes ont donné lieu à plusieurs publications avec suivi et améliorations (***Figure 4***) : il s'agit de SimPheny, développé par Sang Yup Lee en Corée, BNICE co-développé aux États-Unis et en Suisse, et le système RetroPath développé en France. Comme on le verra par la suite, ces systèmes peu ou prou utilisent le même algorithme pour coder la rétrosynthèse.

Figure 4

État de l'art des méthodes de rétrosynthèse utilisées actuellement. Cinq méthodes ont été développées. Les trois principales sont : SimPheny (Corée), BNICE (États-Unis/Suisse) et RetroPath.

2 Fonctionnement du processus de rétrosynthèse

2.1. Règles de réactions codant la spécificité/promiscuité enzymatique

Pour SimPheny et BNICE, les règles de réactions ont été générées manuellement à partir de l'ensemble des réactions et de la classification des enzymes acquises par le passé. On utilise environ cinquante règles pour SimPheny et une centaine pour BNICE. Pour le système RetroPath, les règles sont générées automatiquement à partir de bases de données.

Les règles peuvent être spécifiques, c'est-à-dire concerner la transformation d'un seul substrat en un produit, mais elles peuvent être aussi généralistes et s'appliquer sur plusieurs substrats, voire un très grand nombre. Ainsi les règles que l'on trouve dans BNICE et SimPheny sont très généralistes puisqu'elles tentent de représenter l'ensemble des réactions métaboliques par au plus une centaine de règles, alors qu'il existe entre 30 000 et 40 000 réactions dans les bases de données.

Le caractère spécifique ou généraliste des règles est à rapprocher d'une propriété des enzymes. En effet, ceux-ci peuvent être promiscuitaires[7], c'est-à-dire accepter des substrats différents de ceux pour lesquels ils ont été annotés. Cette caractéristique est illustrée dans la ***Figure 5A***, où une lyase[8] permet de catalyser la tyrosine[9] en coumarate[10].

Afin de coder une réaction enzymatique, dans un premier temps on numérote les atomes pour suivre la transformation entre substrat et produit. Les atomes qui changent leurs configurations sont marqués en rouge dans la ***Figure 5A*** ; ils constituent le centre de la réaction. Si notre réaction est spécifique, elle concerne tous les atomes de la molécule considérée ; on entoure alors l'ensemble de la molécule d'une « sphère » de diamètre infini.

Les substrats et produits inclus dans la sphère sont ensuite représentés sous forme de chaînes « SMILES » ou « SMARTS » (un système de codage de molécules et réaction largement utilisé en chimie computationnelle). Le codage sous forme de chaînes « SMILES » ou « SMARTS » permet d'utiliser des logiciels

7. Promiscuité : la capacité, pour une enzyme, de catalyser efficacement une réaction chimique distincte de celle(s) principalement catalysée(s) par cette enzyme.

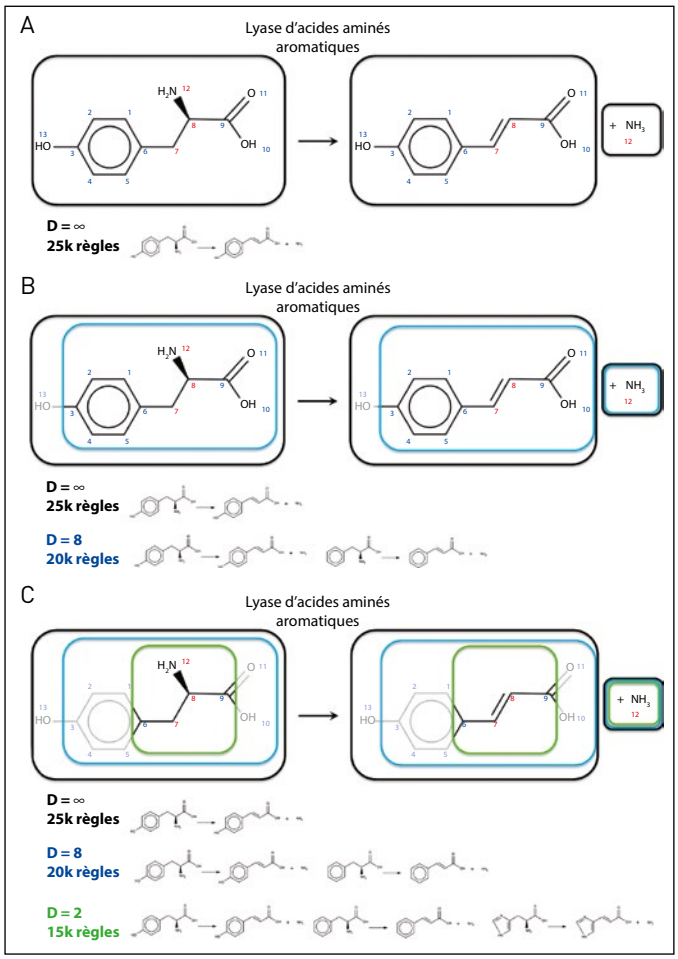
8. Lyase : enzyme capable de casser des liaisons covalentes, créant souvent de nouvelles doubles liaisons. Comme toutes les enzymes, les lyases sont des protéines qui possèdent un site actif permettant la réaction enzymatique, et un site de reconnaissance des molécules cibles, assurant la spécificité de la réaction.

9. Tyrosine : acide aminé, présent dans le corps humain. Elle participe notamment à la synthèse de l'adrénaline, la noradrénaline, la dopamine et la DOPA. Elle est aussi précurseur de la mélanine et des hormones thyroïdiennes.

10. Coumarate : enzyme appartenant à la famille des ligases. Cette enzyme est responsable de formation des liaisons carbone-soufre.

Figure 5

Modélisation du système de codage utilisé dans le cas de la transformation de la tyrosine en coumarate. Les atomes désignés en rouge vont former le centre de la réaction. A) La sphère de diamètre infini modélise la spécificité de la réaction (en noir). Le diamètre représente les atomes qui vont être pris en considération par le programme. B) Le diamètre de la sphère peut être rétréci autour des atomes désignés en rouge : le diamètre en bleu est égal huit et en vert à deux. C) La diminution du diamètre a pour effet d'augmenter la promiscuité de la réaction. À partir de la base de données MetaNetX (www.metanetx.org), il existe ~25 000 règles de réactions à diamètre infini, ~20 000 règles à diamètre 8, et ~15 000 à diamètre 2. Au total, tout diamètre confondu, le nombre de règle est d'environ 120 000.



calculant automatiquement les produits possibles à partir d'un substrat et d'une réaction. Cette procédure est particulièrement intéressante pour des sphères de petits diamètres où une règle peut accepter des substrats différents et donner lieu à un grand nombre de produits possibles.

Il existe des séquences enzymatiques qui peuvent catalyser à la fois la tyrosine et la phénylalanine[11]. Dans le cas de la phénylalanine, on va produire du trans-cinnamate[12]. On ne peut pas dans un tel cas utiliser un codage qui aurait une « sphère » de diamètre infini car la règle de réaction doit accepter des substrats différents. Dans l'exemple de la ***Figure 5B***, on réduit le diamètre à huit. Cela signifie qu'on va prendre en considération tous les atomes jusqu'à quatre liaisons du centre de la réaction. Cette règle de réaction permet

11. Phénylalanine : acide aminé, présent dans le corps humain. La phénylalanine est notamment un précurseur de l'adrénaline et de la mélanine. L'aspartame en dérive également.

12. Trans-cinnamate : sel ou ester de l'acide cinnamique. L'acide cinnamique est utilisé dans l'industrie du parfum. Il possède également des propriétés antiseptiques et antifongiques.

de passer la tyrosine et la phénylalanine (***Figure 5B***).

On peut continuer et diminuer le diamètre pour obtenir des règles de réaction encore plus promiscuitaires. Dans la réaction de catalyse de l'histidine[13], on utilise un diamètre de 2 – une règle de petit diamètre permettant d'accepter un grand nombre de substrats (***Figure 5C***).

Lorsque qu'on utilise des règles qui ont un très petit diamètre, ce qui est le cas des jeux de règles de SimPheny ou de BNICE, on a rapidement une explosion combinatoire du nombre de composés susceptibles d'être générés. Comme décrit ci-dessous, la solution implémentée dans RetroPath consiste à utiliser un diamètre variable pour limiter l'explosion combinatoire de solutions.

2.2. Recherche de séquences enzymatiques par apprentissage automatique

Dans le système RétroPath, les règles sont à diamètre variable : on cherche dans un premier temps une solution avec des règles de diamètre infini ; si aucune solution n'est trouvée, on diminue alors le diamètre de façon à explorer plus de voies enzymatiques.

RetroPath permet aussi d'effectuer une recherche automatique de séquences enzymatiques capables de catalyser les réactions produites par le programme de rétrosynthèse. Pour ce faire, on utilise des méthodes d'apprentissage automatique. Les techniques mathématiques utilisées par ces méthodes sont complexes et nous n'allons pas les exposer. Nous donnons ici simplement quelques éléments qui peuvent faire saisir leur esprit.

Les méthodes d'apprentissage automatique manipulent des objets mathématiques (ici des vecteurs). Dans notre cas, nous devons d'abord associer ces vecteurs de manière non équivoque aux séquences d'acides aminés et aux réactions chimiques : cela s'appelle le codage.

Pour coder les séquences, une technique bioinformatique est largement utilisée : elle consiste à décomposer les séquences protéiques (ici des enzymes), en k-mer[14]. On déplace une fenêtre sur la séquence et on comptabilise le nombre d'occurrences de chacun des k-mers que l'on rencontre lors du déplacement. Le résultat se représente sous forme d'un vecteur (***Figure 6***, gauche).

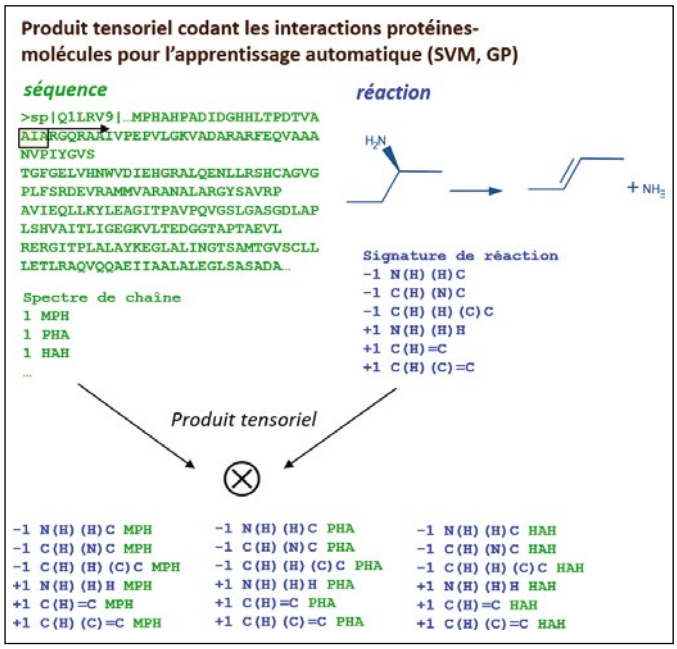
Pour coder les réactions, c'est un peu la même idée : on place une sphère sur chaque atome et on compile l'équivalent du k-mer, qui est ici l'environnement atomique de cet atome, et on comptabilise le nombre d'occurrences de ces environnements. Cela se fait sur les substrats et sur les produits. Pour calculer la « signature » d'une réaction, on soustrait à la signature des produits la

13. Histidine : acide aminé précurseur de l'histamine et de la carnosine. Il n'est pas essentiel au corps humain sauf durant l'enfance et la grossesse.

14. K-mer : sous-séquence possible, de longueur k, obtenue à partir du séquençage de l'ADN.

Figure 6

Produit tensoriel codant les interactions protéines-molécules pour l'apprentissage automatique. En entrée, des vecteurs codent des séquences (k-mers ou spectre de chaînes) et des réactions (signatures). Il s'agit ensuite de faire le produit tensoriel de ces deux vecteurs (séquence et réaction) qui représente l'ensemble des combinaisons possibles entre les signatures et les k-mers.



signature des substrats. Ainsi, la ***Figure 6*** à droite montre des nombres positifs et les nombres négatifs : les nombres négatifs sont les configurations que l'on trouve dans les substrats mais pas les produits, et les nombres positifs sont les configurations que l'on trouve dans les produits et pas les substrats. Ce travail conduit aussi à une représentation vectorielle (***Figure 6***, droite).

On est ensuite en mesure de coder le complexe réaction-séquence. Cela se fait par une méthode à noyau nommé produit tensoriel. Le produit tensoriel prend en compte toutes les combinaisons entre les k-mers des séquences et les signatures des réactions (***Figure 6***, en bas). Le résultat est un vecteur qui représente le complexe : réaction, séquence enzymatique.

On peut alors utiliser plusieurs techniques d'apprentissage automatique : machine à vecteurs de support, forêt aléatoire ou processus gaussien. Dans tous ces cas, on obtient de bons résultats pour la classification de séquences enzymatiques et réactions (***Figure 7***, gauche). Ces méthodes permettent aussi de prédire des constantes cinétiques (telle que la constante de Michaelis) pour certaines classes d'enzymes (***Figure 7***, droite).

Les méthodes que nous venons de présenter nous permettent de construire des cartes de rétrosynthèse en appliquant les règles de réactions et de prédire les séquences enzymatiques catalysant ces réactions. Nous avons ainsi tous les éléments en place pour décrire le « workflow »[15] de rétrosynthèse.

15. Workflow : processus d'automatisation des tâches permettant un enchaînement automatisé des différentes opérations et étapes de validation d'une tâche plus ou moins complexe.

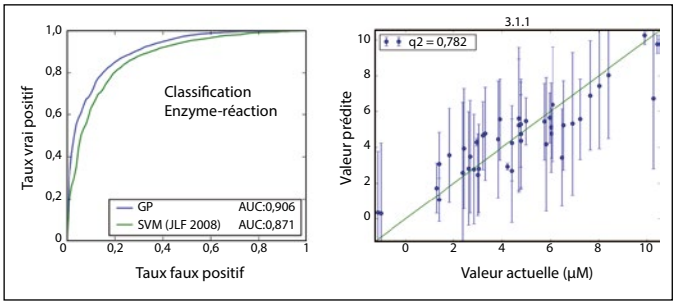


Figure 7

Grâce au produit tensoriel, il est possible de classifier les enzymes et les réactions. À gauche, une courbe ROC montrant que le taux faux positif est faible en utilisant des machines à vecteur de support (SVM) ou des processus gaussiens (GP). En utilisant des processus gaussiens, il est aussi possible de prédire des constantes cinétiques comme la constante de Michaelis dans l'équation de Michaelis-Menten (pour les enzymes de la classe EC 3.1.3, à droite).*

* Une courbe ROC (« *Receiver Operating Characteristic* ») est une courbe sensibilité/spécificité.

2.3. Rétrosynthèse : le workflow

Le paragraphe précédent présente le système de codage des règles de réactions gouvernant la rétrosynthèse. Nous décrivons ici le déroulement du processus : le workflow. Celui-ci encode un algorithme de rétrosynthèse général pouvant être utilisé avec tout type de jeux de règles telles que celle développées dans SymPhenie, BNICE ou RetroPath.

Le workflow présenté dans la ***Figure 8A*** a été développé sur la plateforme KNIME ; il est stocké dans la base de données MyExperiment.org et peut être téléchargé sur Internet[16].

En entrée du workflow, on donne la molécule que l'on désire synthétiser. Il s'agit ici d'une flavanone, la pinocembrine[17], qui est un précurseur des flavonoïdes. Les flavonoïdes sont des molécules intéressantes pour l'industrie pharmaceutique par leurs nombreuses propriétés : anti-inflammatoires, anti-oxydantes, antibactériennes, anticancéreuses... La pinocembrine est l'un des trois précurseurs de l'ensemble des flavonoïdes. En entrée du worflow, on donne également le jeu de règles : ici il s'agit d'un jeu de règles à diamètre 4 autour des centres de réactions extraites de la base de données MetaNetX[18]. On donne aussi un « sink », c'est-à-dire un ensemble de métabolites vers lequel la rétrosynthèse doit aboutir. Le « sink » est l'ensemble des métabolites du châssis

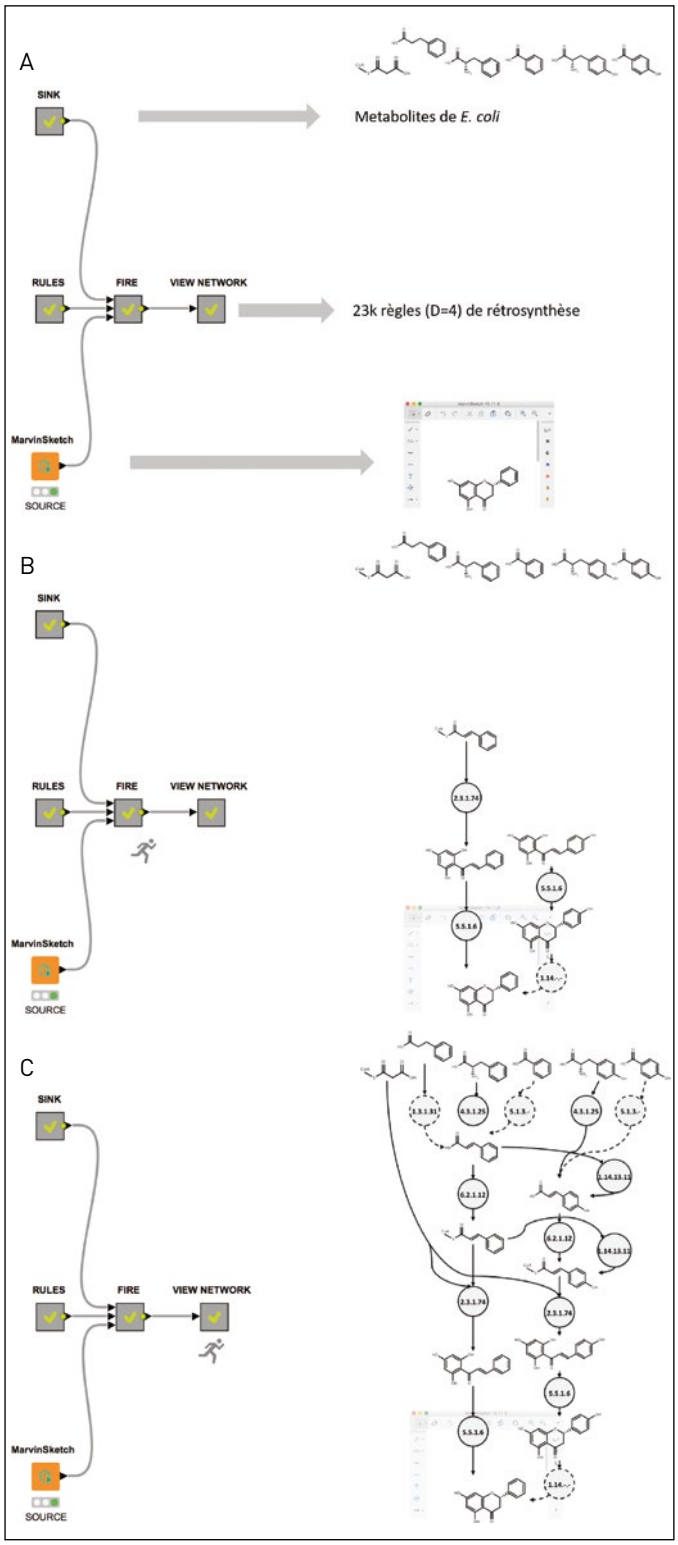
16. www.myexperiment.org/workflows/4987.html

17. Pinocembrine : antioxydant présent dans le miel ou le propolis.
18. www.metanetx.org

Figure 8

*A) Modélisation du workflow avec les différents paramètres d'entrée : les métabolites d'*Escherichia coli, *le nombre de règles de rétrosynthèse et le châssis de la réaction ; B) 2e itération ; C) 3e itération. « Source » est la molécule que l'on désire synthétiser (ici la pinocembrine) ; « sink » est l'ensemble des molécules de la souche châssis (ici* Escherichia coli*). MarvinSketch est un logiciel de ChemAxon* permettant de dessiner une molécule.*

*https://chemaxon.com/products/marvin



de production (c'est-à-dire *Escherichia coli*).

En faisant fonctionner le workflow, on applique les règles inversées de la synthèse – ce sont des règles de rétrosynthèse – sur le produit final. On remonte itérativement jusqu'au métabolite d'*Escherichia coli*. La ***Figure 8A*** montre la première itération, la ***Figure 8B*** la seconde itération, et la ***Figure 8C*** la troisième.

Ce faisant, on construit une carte de rétrosynthèse. Cette carte étant complexe, le premier problème est de répondre à la question « *combien de voies de synthèse y-a-t-il dans la carte ?* ». En effet, il y a probablement plusieurs façons de synthétiser la pinocembrine.

Le problème à résoudre est donc l'énumération des voies métaboliques dans les cartes de rétrosynthèse. Ici, il ne s'agit pas simplement de trouver un chemin dans une carte puisque lorsqu'on parle d'une réaction, cette réaction peut avoir plusieurs substrats et lorsqu'on remonte le chemin, il va falloir remonter cette propagation sur tous les substrats. On a en fait à faire à ce qu'on appelle techniquement un hyperchemin dans un hypergraphe. Pour résoudre ce problème, nous avons développé une méthode basée sur les modes élémentaires. Cette méthode, développée dans le cadre de l'étude du métabolisme en général, permet ici d'énumérer les différentes voies. Il y a onze voies différentes pour produire la pinocembrine (***Figure 9***). Le nombre de voies pouvant être élevé, on peut être amené à les classer et ne construire que les premières du classement.

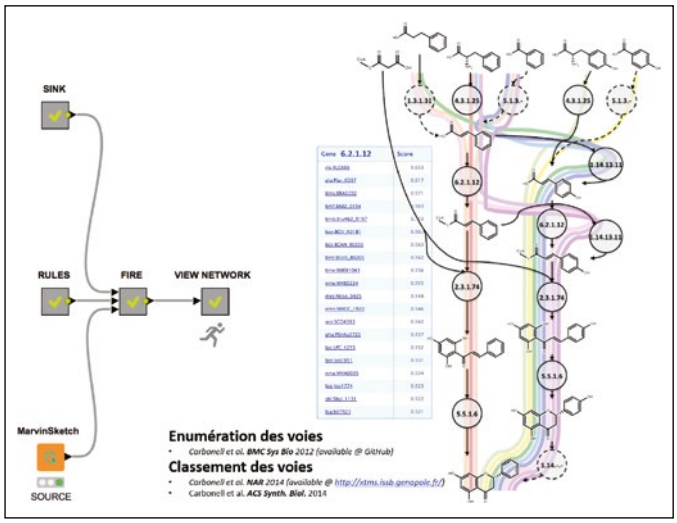
Le classement se base sur plusieurs critères, tout d'abord au niveau des gènes codant les enzymes, car les enzymes peuvent être plus ou moins efficaces en fonction de leurs séquences. Pour classer les séquences enzymatiques on utilise les méthodes d'apprentissage abordées plus haut (voir le paragraphe 2.2) ; en effet, ces méthodes permettent de calculer les scores des différentes séquences enzymatiques.

Le second critère est basé sur les flux théoriques des voies. Les différentes voies métaboliques peuvent avoir différents flux vers le produit final car elles utilisent différents cofacteurs, par exemple différentes quantités d'ATP[19], ou autres métabolites essentiels à la pousse de la souche châssis. Le troisième et dernier critère est lié à la toxicité du produit final et des produits intermédiaires de la voie. Si ceux-ci sont toxiques, alors il n'est pas recommandé de construire cette voie car même si les enzymes sont efficaces et les flux élevés, la souche châssis ne poussera pas à cause de la toxicité. À partir de ces trois critères, les voies peuvent être classées, ce qui permet de passer à l'étape suivante : la vérification expérimentale.

19. ATP : Adénosiné TriPhosphate, nucléotide formé à partir d'adénine liée à un ribose attaché à un triphosphate. Il fournit l'énergie nécessaire aux réactions chimiques du métabolisme à travers les membranes biologiques.

Figure 9

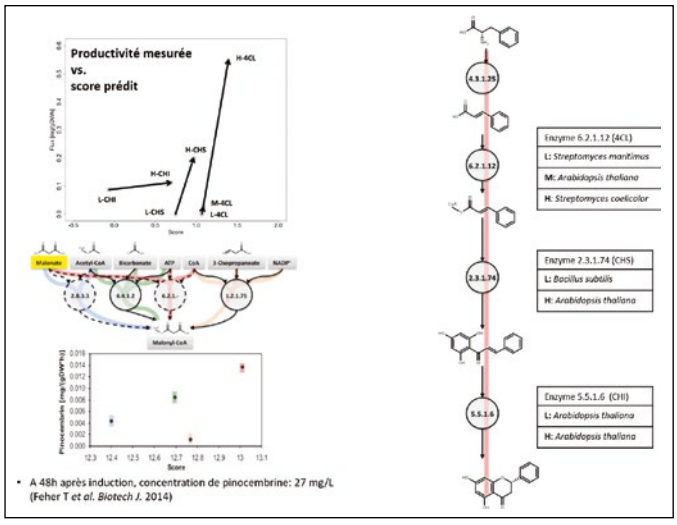
Après plusieurs itérations, des chemins de synthèse sont obtenus (ici représentés en couleur), et constituent une carte de rétrosynthèse. Chacune de ces voies (onze au total) permettrait de synthétiser la pinocembrine ; elles ne vont pas toutes être exploitées. L'énumération des voies est un problème non trivial : il faut en effet remonter la propagation pour chaque substrat, et le nombre de substrats peut s'avérer élevé.



Pour vérification expérimentale, nous avons pris systématiquement, toujours dans la voie de la pinocembrine, des séquences qui avaient un score élevé et d'autres avec un score plus bas d'après les méthodes d'apprentissage (***Figure 10***, partie droite). La ***Figure 10***, haut gauche, montre que les séquences avec un score bas donnent moins de produit final que les séquences au score plus élevé. Nous avons testé le classement des voies pour une autre molécule, le Malonyl-CoA, un cofacteur de la voie de production de la pinocembrine. On a là encore une assez bonne corrélation entre les scores retournés des classements des voies et la quantité de pinocembrine obtenue (***Figure 10***, bas gauche).

Figure 10

Une fois énumérées, les voies de synthèse vont être classées en fonction du score des séquences enzymatiques retourné par la méthode d'apprentissage. À gauche, les graphiques montrent qu'il y a une bonne corrélation entre le score prédit par les méthodes d'apprentissage et la productivité. En effet, les séquences avec un score élevé donnent une quantité de produit final plus importante que celles avec un score bas.



3 Exemples d'applications de ce dispositif de rétrosynthèse

3.1. Application à l'ingénierie métabolique

Les méthodes développées en rétrosynthèse se révèlent très précieuses pour l'ingénierie métabolique, où l'on veut construire des banques combinatoires, en variant par exemple les promoteurs[20], les origines de réplication, ou les séquences des sites de fixation du ribosome. D'une façon complètement robotisée, nous avons construit une banque de 41 plasmides[21], où nous avons varié le nombre de copies des plasmides, la force des promoteurs et l'ordre des gènes en prenant les séquences enzymatiques de la ***Figure 10*** ayant le meilleur score. Pour l'une de ces constructions, nous avons obtenu environ cinquante milligrammes par litre de pinocembrine – à comparer à l'état de l'art actuel qui est de quarante milligrammes par litre. Le châssis utilisé était là aussi *Escherichia coli* (***Figure 11***).

La ***Figure 12*** montre d'autres exemples d'applications de ces mêmes méthodes de rétrosynthèse. La ***Figure 12A*** correspond à la production de téréphtalate (TPA),

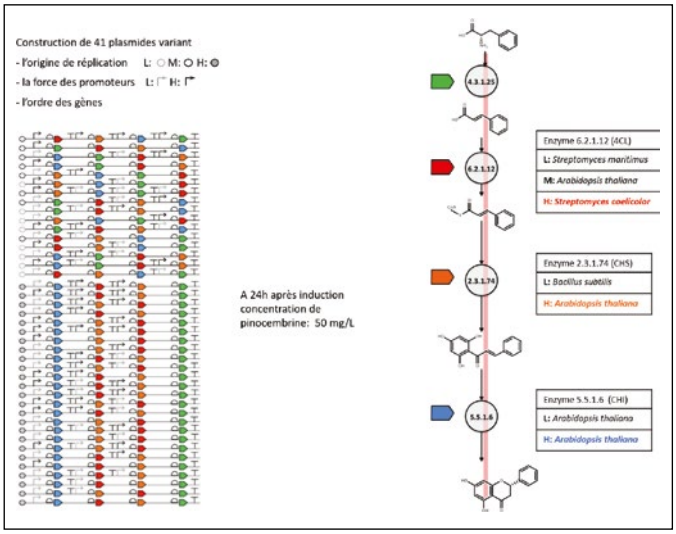


Figure 11

À gauche, modélisation de la construction de 41 plasmides en faisant varier l'origine de réplication, la force des promoteurs et l'ordre des gènes. Cette banque permet de tester toutes les combinaisons possibles ; pour faire un premier tri, on ne garde que celles qui ont un score élevé. Une des combinaisons a montré un taux de production très élevé, 50 mg/L de pinocembrine, représenté à droite.

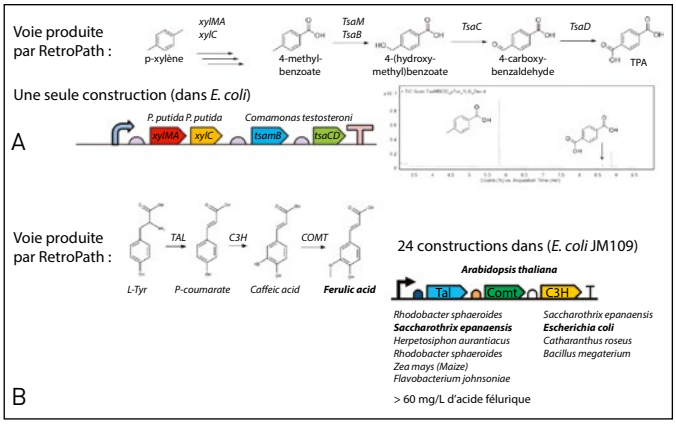


Figure 12

Exemples d'utilisation de RetroPath pour trouver des voies de synthèse. Pour le TPA, en haut, une voie de synthèse est proposée en utilisant quatre enzymes (xylMA, xylC, tsaMB, tsaCD). Cette voie a montré une bonne productivité dans le châssis Escherichia coli. *Pour l'acide férulique, une voie a été proposée avec plusieurs constructions possibles (24 au total dans* Escherichia coli*). La construction avec les trois enzymes (Tal, Comt et C3H) donne une production supérieure à 60 mg/L d'acide férulique.*

20. Promoteur : courte séquence d'ADN, généralement situé en amont du gène, qui en contrôle l'expression, notamment en régulant sa transcription.
21. Plasmide : molécule d'ADN circulaire double brin, naturelle ou modifiée artificiellement, dans le but de l'utiliser en recherche biologique.

un monomère utilisé dans le polyéthylène téréphtalate (PET) et aussi dans le kevlar®[22]. La méthode de rétrosynthèse nous donne une voie à partir du xylène, en utilisant quatre enzymes. Cette voie a effectivement été mise en œuvre dans *Escherichia coli* avec succès.

L'exemple de la ***Figure 12A*** est la production de l'acide férulique, un précurseur de la vaniline, qui est un arôme artificiel mais aussi un précurseur d'autres molécules intéressantes pour l'industrie cosmétique comme le malate de synapoyle, intervenant dans les crèmes anti-UV. Ici, nous avons une voie et plusieurs séquences enzymatiques possibles. Nous les avons toutes construites (soit 24). L'une d'entre elles a donné plus de 70 mg/L d'acide férulique (***Figure 12***).

Parmi les autres vérifications expérimentales que nous avons menées, nous avons utilisé le workflow RetroPath2.0 pour des molécules de la base de données LASER, qui référencie toutes les constructions d'ingénierie métabolique de différents groupes académiques et industriels. Nous avons donné au workflow de rétrosynthèse seulement la molécule finale et la souche utilisée. Dans 80 % des cas, RetroPath2.0 a retrouvé les voies de synthèses stockées dans la base de données LASER.

22. Kevlar® : marque déposée d'une fibre d'aramide (produite par l'entreprise Dupont de Nemours). Le kevlar® est recherché pour ses propriétés spécifiques, sa transparence aux ondes radar, son comportement linéaire et sa tolérance élevée aux chocs et à l'usure.

3.2. Application à l'ingénierie de biocapteurs

Nous présentons dans ce paragraphe des applications de la rétrosynthèse liées aux biocapteurs. Nous rappelons qu'il s'agit ici de modifier au travers de réactions enzymatiques une molécule en une autre détectable par un facteur de transcription ou un riborégulateur.

Pour ce faire, nous avons utilisé le workflow RetroPath2.0 présenté plus haut avec, comme donné d'entrée, les molécules que l'on aimerait détecter (« source ») – ici des molécules thérapeutiques (médicaments) répertoriées dans la base de données DrugBank[23], des biomarqueurs issus de la base de données HMDB[24] et des produits toxiques pour l'environnement de la base de données Tox21[25]. Dans ces trois cas, les molécules sont modifiées pour être détectées par les règles de réactions présentées plus haut (***Figure 13***).

Afin d'augmenter les chances de succès expérimentaux, nous avons choisi des diamètres infinis pour s'assurer qu'il y avait bien une séquence enzymatique capable de catalyser les réactions. Les molécules finales recherchées (« sink ») sont ici des effecteurs, c'est-à-dire des molécules détectables directement par des facteurs de transcription ou des riborégulateurs. Malheureusement cette information n'est pas disponible directement dans une base de

23. https://www.drugbank.ca/
24. http://www.hmdb.ca/
25. https://ntp.niehs.nih.gov/results/tox21/index.html

données, et tout un travail a été nécessaire pour compiler une liste de ces effecteurs au travers différentes bases de données comme RegulonDB[26], qui s'intéresse au réseau de régulation chez *Escherichia coli*. Toujours est-il qu'à partir du workflow présenté dans la ***Figure 13***, on est capable de concevoir le design d'environ mille biocapteurs.

À partir de ce travail, nous avons développé un site web[27] où l'on peut rentrer n'importe quelle molécule et obtenir les voies métaboliques qui la transforment en effecteur, ainsi que les informations sur les facteurs de transcription et les séquences enzymatiques. À partir du site http://sensipath.micalis.fr/, on dispose de toutes les informations nécessaires pour passer à la construction, ce que nous avons réalisé pour une douzaine de biosenseurs.

Dans nos constructions, nous utilisons deux plasmides : un plasmide qui va coder pour les enzymes qui vont transformer la molécule non détectable en effecteur, et un plasmide comprenant le senseur lui-même, c'est-à-dire le facteur de transcription ainsi qu'un marqueur fluorescent.

Une application expérimentale de ces techniques est illustrée sur la ***Figure 14***. Le but est de détecter l'acide hippurique, qui est un biomarqueur du cancer de la prostate et de diverses intoxications, par exemple par le toluène. L'acide hippurique, indétectable, peut se transformer en acide benzoïque qui, à

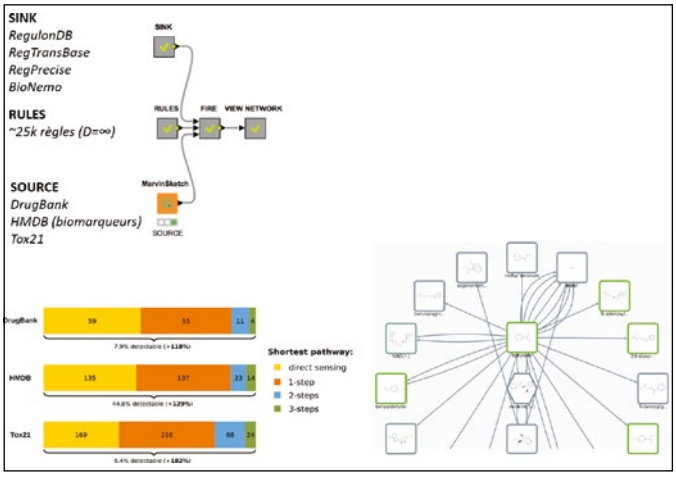


Figure 13

À gauche, la modélisation du workflow montre les différentes sources d'entrée possibles : DrugBank, HMDB, Tox21. Les pourcentages de molécules détectables sont donnés pour chacune des sources choisies. Par exemple, pour DrugBank, 7,9 % des molécules sont détectables après transformation enzymatique, ce qui constitue une augmentation de +118 % par rapport aux molécules directement détectables. Le graphe en bas à droite est un visuel extrait du site http://sensipath.micalis.fr, la molécule à détecter est ici la cocaïne et les effecteurs sont en vert (benzoate, benzaldehyde...).

son tour, active le facteur de transcription BenR permettant d'exprimer une protéine fluorescente.

L'insert de la ***Figure 14*** montre les courbes dose-réponse de d'acide hippurique (qui ne présente qu'une fluorescence négligeable) et de l'acide benzoïque (qui présente une fluorescence importante) lorsque le module métabolique est absent. Comme attendu, l'acide hippurique n'est pas détecté par le biosenseur.

La ***Figure 14*** montre qu'en présence du module métabolique, la fluorescence de l'acide hippurique suit celle de l'acide benzoïque, montrant que la transformation est complète. On vérifie donc que l'acide hippurique est bien détectable

26. http://regulondb.ccg.unam.mx/
27. http://sensipath.micalis.fr/

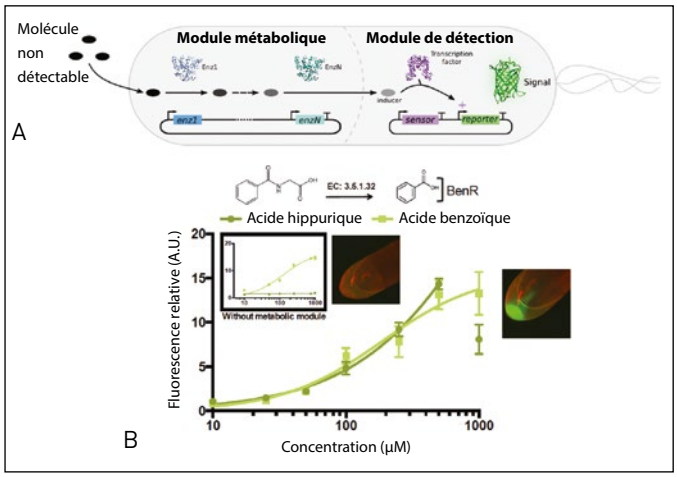


Figure 14

A) La molécule non détectable est d'abord transformée par réactions enzymatiques, ensuite le produit de ces réactions est détecté par un facteur de transcription qui à son tour permet d'exprimer une protéine fluorescente ; B) en bas, la flurorescence est donnée en fonction de la concentration en acide hippurique et en acide benzoïque. Sans module métabolique, l'acide hippurique n'est pas transformé en acide benzoïque (insert) et aucune fluorescence n'est observée pour l'acide hippurique. Lorsque du module métabolique est ajouté, les courbes relatives à l'acide benzoïque et à l'acide hippurique sont superposées, ce qui signifie que l'acide hippurique a totalement été consommé en acide benzoïque.

lorsque que le module métabolique est introduit dans la construction du biosenseur.

La ***Figure 15*** donne d'autres exemples de détection :

– le parathion : un polluant environnemental, qui est aussi un produit de dégradation du cyclosarin[28], un gaz de combat. Là encore on a une belle courbe dose/réponse (***Figure 15***), le parathion est complètement transformé en une molécule-effecteur détectable ;

– le chloronitrophénol : on a besoin de deux enzymes pour transformer ce polluant en une molécule détectable directement par un facteur de transcription. Les deux courbes sont quasiment identiques, indiquant là encore que la molécule est complètement transformée.

La ***Figure 16*** présente d'autres exemples de détection de biomarqueurs du cancer de la prostate. Dans les exemples choisis, nous avons cherché à augmenter la spécificité de nos biosenseurs, en faisant coexister plusieurs transformations où un même biomarqueur est transformé en plusieurs effecteurs. Nous avons aussi cherché à faire une détection multiplexe de différents biomarqueurs, car souvent dans les cancers, la mesure de la

28. Cyclosarin : substance chimique extrêmement toxique utilisée comme arme chimique (il entraîne la mort par asphyxie). C'est un agent organophosphoré neurotoxique dérivé du sarin.

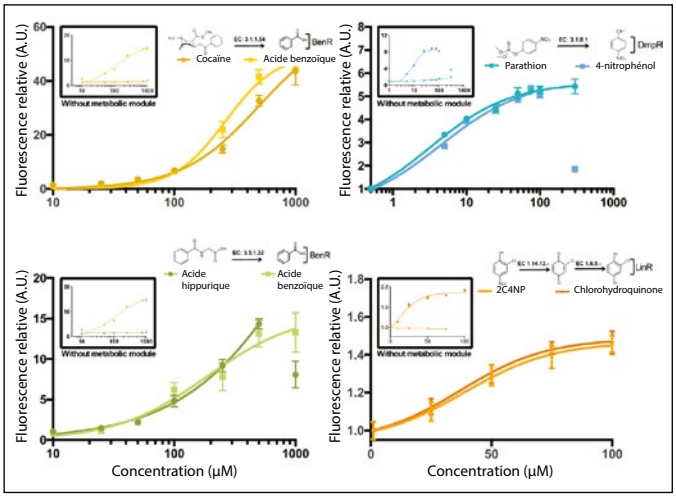
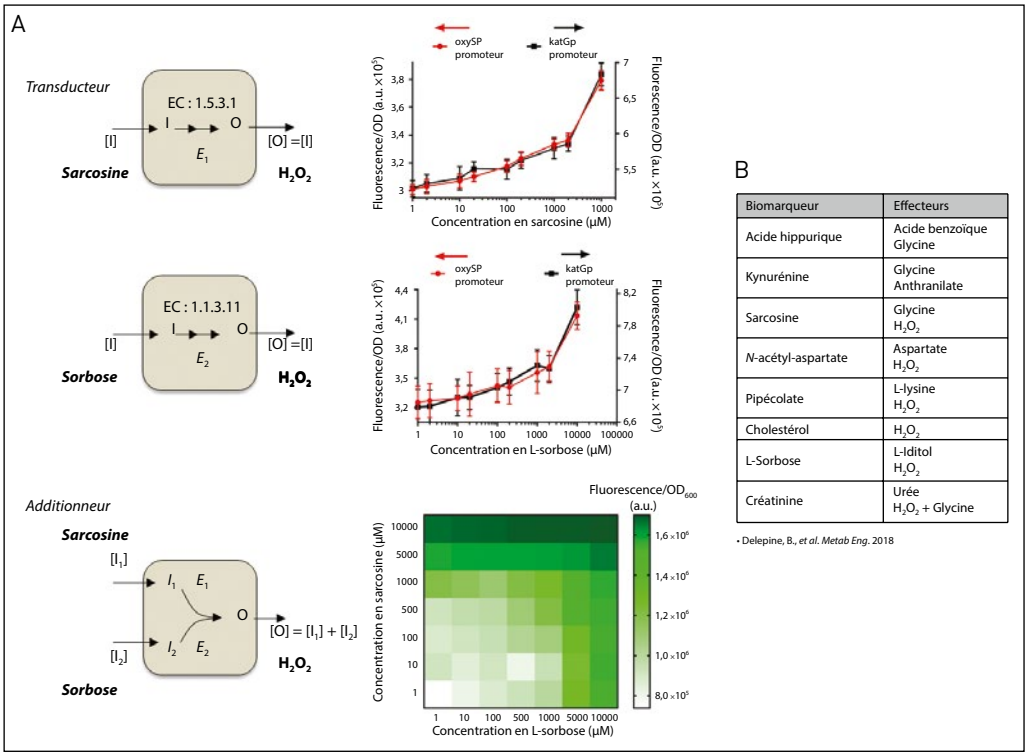


Figure 15

Plusieurs exemples de courbes dose/réponse pour la cocaïne, le parathion, l'acide hippurique et le 2C4NP. Chaque courbe montre que grâce au module métabolique, les molécules à détecter sont totalement consommées. Ainsi, ces molécules ont été transformées en molécules détectables.



B

Biomarqueur	Effecteurs
Acide hippurique	Acide benzoïque Glycine
Kynurénine	Glycine Anthranilate
Sarcosine	Glycine H_2O_2
N-acétyl-aspartate	Aspartate H_2O_2
Pipécolate	L-lysine H_2O_2
Cholestérol	H_2O_2
L-Sorbose	L-Iditol H_2O_2
Créatinine	Urée H_2O_2 + Glycine

• Delepine, B., *et al. Metab Eng.* 2018

Figure 16

A) Modélisation des dispositifs électroniques utilisés dans la biologie de synthèse : transducteur et additionneur. Dans le cadre des transducteurs, les biomarqueurs vont être transformés en effecteurs ; B) ce tableau permet de répertorier les effecteurs relatifs aux biomarqueurs. Les graphiques au centre montrent la fluorescence observée en fonction de la concentration en biomarqueur en utilisant différents facteurs de transcriptions (oxySp et katGp). Dans le cadre de l'additionneur, on va coupler deux biomarqueurs en un effecteur, ici la sarcosine et le L-Sorbose tous les deux transformés en H_2O_2. En sortie, on obtient une cartographie de la fluorescence en fonction de la concentration des deux biomarqueurs ; on peut aisément vérifier que la fluorescence est bien proportionnelle à la somme des concentrations des biomarqueurs.

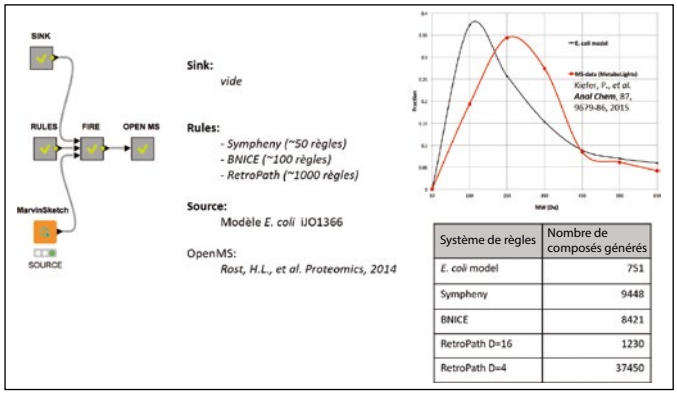
concentration d'une seule molécule n'est pas suffisante à diagnostiquer la maladie.

Pour réaliser des détections spécifiques et multiplexes, nous avons construit plusieurs dispositifs, et en particulier des transducteurs et des additionneurs (***Figure 16***, gauche). Il est à noter que nos dispositifs sont différents de ceux réalisés de façon courante en biologie de synthèse, où l'on a souvent affaire à des dispositifs génétiques-numériques, par exemple des portes logiques, où l'on doit exprimer des protéines ou des séquences d'ARN. Ici, nous parlons de dispositifs métaboliques-analogiques.

L'avantage de travailler de façon analogique au niveau du métabolisme, c'est d'éviter d'avoir à discrétiser nos entrées. En effet, un métabolique n'est jamais ON ou OFF dans une cellule mais est présente dans une certaine concentration. D'autre part, les circuits métaboliques, basés sur des transformations enzymatiques, sont beaucoup plus rapides que les circuits génériques-numériques. En effet, un circuit générique prend au minimum une demi-heure pour exprimer les protéines sur chacune de ces couches, donc l'exécution d'un circuit à plusieurs couches peut être très longue. La cinétique des circuits métaboliques-analogiques est celle des réactions enzymatiques, donc beaucoup plus rapide.

3.3. Application à la métabolomique

La dernière application présentée dans ce chapitre est une tentative de pallier à l'insuffisance de la connaissance des métabolites des souches utilisées en biotechnologie. En effet, la ***Figure 17*** montre la



Système de règles	Nombre de composés générés
E. coli model	751
Sympheny	9448
BNICE	8421
RetroPath D=16	1230
RetroPath D=4	37450

Figure 17

*Distribution des masses d'*Escherichia coli. *En noir, la courbe correspond au modèle et en rouge au résultat de spectrométrie de masse d'un extrait cellulaire d'*Escherichia coli. *Les différences entre les deux courbes s'expliquent par le manque de certains métabolites dans le modèle. En dessous, le tableau montre le nombre de composés générés par RetroPath2.0 en fonction des systèmes de règles choisis.*

distribution des masses dans un modèle d'*Escherichia coli* (courbe noire en haut à gauche). Des bases de données comme EcoCyc[29] donnent le même type de courbe. La courbe en rouge est une courbe calculée à partir du spectre de masse d'un extrait cellulaire d'une cellule d'*Escherichia coli*. On observe une différence entre la distribution expérimentale et la distribution du modèle montrant ainsi que certains métabolites sont absents du modèle.

Ne peut-on pas alors utiliser des programmes de rétrosynthèse pour trouver les molécules manquantes dans les modèles ? Effectivement, en utilisant le workflow RetroPath2.0 avec les règles de réactions de SimPheny, BNICE ou Retropath, on obtient un certain nombre de molécules qui ne sont pas initialement dans le modèle d'*Escherichia coli*.

L'intérêt d'utiliser des systèmes de workflows est de pouvoir facilement les coupler entre eux. Ainsi RetroPath2.0 peut être couplé avec le workflow OpenMS qui permet de faire de l'annotation des spectres de masse et aider à la confirmation ou non de la présence d'une molécule à partir de sa masse.

La ***Figure 18*** (haut gauche) montre le résultat obtenu par OpenMS en utilisant en entrée un spectre de masse d'un extrait cellulaire d'*Escherichia coli* téléchargé à partir de la base de données MetaboLights[30].

La ***Figure 18*** (haut droite) montre que les métabolites du modèle d'*Escherichia coli* ne couvrent que de 12 % des masses. En utilisant les règles de réactions de RetroPath d'un diamètre de 16, 23 % des masses du spectre sont couvertes (***Figure 18***, bas gauche) avec un diamètre 4, la couverture est de 60 %. Il est ainsi possible de proposer une (ou plusieurs) molécule(s) pour 60 % des masses du spectre.

Le couplage RetroPath2.0-OpenMS nous a permis de détecter la présence dans l'échantillon de l'Acétyl-Leucine[31], qui n'est pas un métabolite connus dans les modèles d'*Escherichia coli* (***Figure 19***).

Afin de confirmer la présence d'Acétyl-Leucine dans *Escherichia coli,* l'étape suivante de notre travail a été de vérifier par spectrométrie de masse que le pic observé était effectivement celui de l'Acétyl-Leucine (***Figure 20***, haut). Nous avons ensuite recherché les enzymes chez *Escherichia coli* responsables de la synthèse de l'Acétyl-Leucine. Pour ce faire, nous avons utilisé une des méthodes d'apprentissage automatique présentées au paragraphe 2.2 (c'est-à-dire un processus gaussien). L'étude a montré que les enzymes ECBD4067 et ECDB4269 étaient responsables de la synthèse d'Acétyl-Leucine (***Figure 20***). Il est bon de noter que de tels exemples d'utilisation de méthodes d'apprentissage automatique en biologie de synthèse ne sont pas si fréquents.

29. https://ecocyc.org/
30. www.ebi.ac.uk/metabolights/

31. Acétyl-Leucine : substance chimique qui est notamment utilisée comme médicament contre les vertiges.

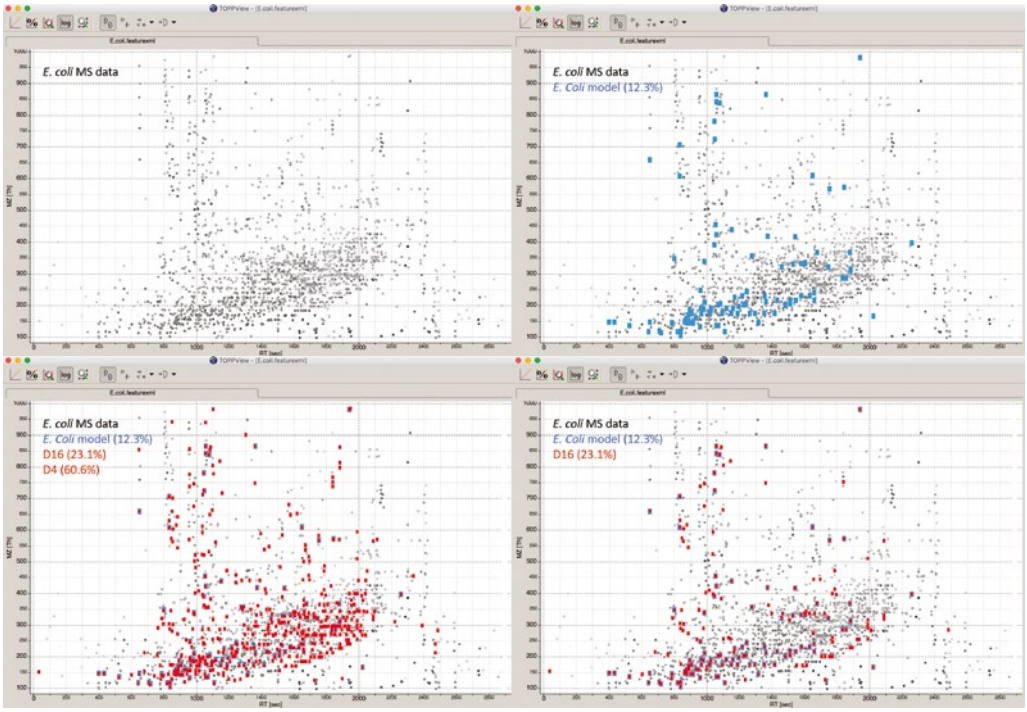
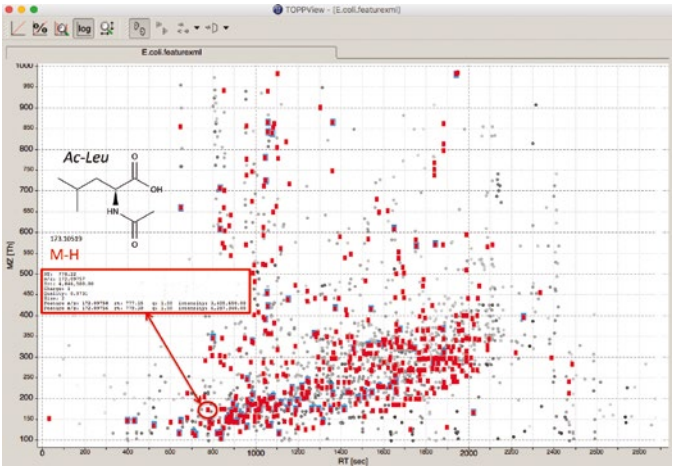


Figure 18

*Spectre de masse d'un extrait cellulaire d'*Escherichia coli *obtenu par le workflow OpenMS. Les points noirs représentent les données expérimentales. On peut ensuite corréler ce graphique avec les masses des molécules d'un modèle d'*Escherichia coli *(en bleu), où seulement 12,3 % des masses expérimentales sont couvertes. On augmente cette couverture en utilisant les masses des molécules produites par RetroPath2.0 (en rouge). En diminuant progressivement le diamètre, la partie du spectre couverte augmente : pour un diamètre de 16 : 23,1 %, et pour un diamètre de 4 : 60,6 %.*

Figure 19

Detection de l'Acétyl-Leucine (en haut à gauche) dans le spectre de masse. Les données retournées montrent que la masse trouvée correspond bien à celle de l'Acétyl-Leucine.



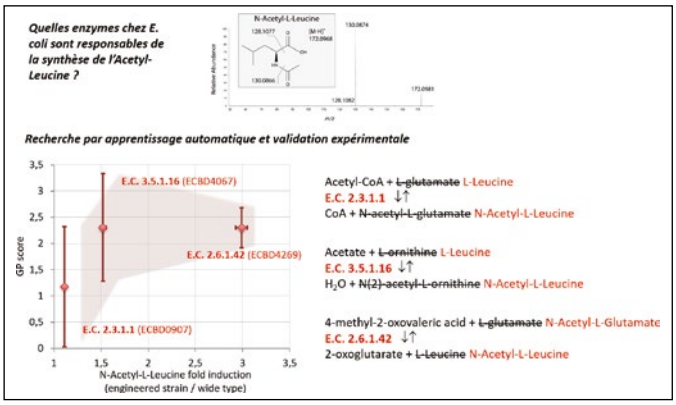


Figure 20

*Recherche par apprentissage automatique et vérification expérimentale des séquences d'*Escherichia coli *responsables de la synthèse d'Acétyl-Leucine. La recherche par apprentissage automatique retourne trois séquences possibles (ECBD0907, ECBD4067, et ECBD4269) correspondant à trois voies métaboliques différentes. Dans chaque cas, les enzymes ont été surexprimées pour mesurer l'augmentation d'Acétyl-Leucine. On a ainsi pu démontrer que les enzymes ECBD4067 et ECBD4269 étaient responsables de la production d'Acétyl-Leucine. Le spectre de masse (en haut ; couplage « MS-MS ») de l'Acétyl-Leucine permet de vérifier le produit obtenu.*

La rétrosynthèse, aujourd'hui et demain

Ce chapitre a montré que la rétrosynthèse pouvait être performante dans diverses applications biologiques. Nous avons vu explicitement trois exemples, mais il en existe d'autres : la biorémédiation[32] ou la dégradation de composés, le métabolisme alternatif, pouvant être utilisé pour fabriquer de façon plus efficace des molécules en utilisant moins d'énergie, ainsi que la combinaison des règles de réactions de synthèse chimique avec des règles de biosynthèse. On peut aussi envisager de déve-

32. Biorémédiation : technique consistant à augmenter la biodégradation ou la biotransformation, en introduisant des micro-organismes spécifiques ou en stimulant l'activité de populations microbiennes, par apport de nutriments et par ajustement des conditions de milieu.

lopper des approches de criblage de l'espace chimique consistant à faire évoluer des populations de molécules au moyen de réactions chimiques et/ou enzymatiques.

Nous avons aussi montré que la rétrosynthèse pouvait être codée sous forme de workflows scientifiques, très simples d'utilisation. Des workflows, encore en développement, permettent déjà de piloter des robots destinés à être utilisés dans les processus d'ingénierie. Ainsi, à terme, le processus d'ingénierie de souches pourrait être complètement piloté et contrôlé par un système automatique basé sur la technologie des workflows.

Finalement, nous avons montré que l'apprentissage automatique était utile dans les méthodes de rétrosynthèse telles qu'appliquées en biologie, pour la recherche des séquences enzymatiques, mais aussi pour les prédictions de toxicité et le classement des voies de synthèses. L'étape suivante est le développement d'un apprentissage actif où le cycle construction, mesure, apprentissage est itéré une première fois, puis relancé avec les mesures obtenues à la première itération. Finalement, pourquoi ne pas utiliser ces méthodes d'apprentissage pour faire directement de l'ingénierie de génome, pour rechercher par exemple automatiquement les niveaux d'expression de gènes permettant l'optimisation d'une souche de bioproduction ?