# Détection de voies métaboliques actives par puces à ADN

Jean-Philippe Vert
Ecole des Mines de Paris
35 rue Saint-Honoré
77300 Fontainebleau
Jean-Philippe.Vert@mines.org

Pratiquement toutes les réactions chimiques qui permettent à un organisme ou une cellule de se maintenir dans un état vivant, se reproduire, communiquer ou résister à des variations dans son environnement, sont catalysées par des enzymes, c'est-à-dire des molécules (protéines) synthétisées par l'organisme lui-même. L'activité biochimique d'une cellule à un instant et dans des conditions données est donc controlée de manière très précise par la quantité d'enzymes disponibles, ce qui fournit un moyen à la cellule de l'ajuster en contrôlant la synthèse et la dégradation des enzymes. Les enzymes étant des protéines synthétisées par traduction d'ARN messager, une manière de réguler leurs concentrations consiste à réguler l'expression des gènes correspondant, c'est-à-dire la quantité des ARN correspondants.

La technique des puces à ADN fournit une estimation de la concentration d'un grand nombre d'ARN d'un génome à un instant donné. En ne regardant que les ARN qui codent des enzymes, il est donc possible d'observer la régulation des quantités d'enzyme au niveau de la transcription.

Indépendamment, un grand nombre de réactions chimiques ayant lieu dans les cellules vivantes ont été mises en évidence au cours du 20e siècle. Les réactions chimiques faisant partie du métabolisme (synthèse ou dégradation de molécules) sont généralement arrangées en voies métaboliques, c'est-à-dire en successions de plusieurs réactions permettant de passer d'un composé A à un composé B en plusieurs étapes. Par exemple, le cycle de Krebs est une voie métabolique permettant de transformer de l'acétate en bicarbonate via une dizaine de réactions chimiques catalysées par autant d'enzymes. L'ensemble des voies métaboliques connues, ainsi que les enzymes servant de catalyseurs, ont été intégrés dans des bases de données telles KEGG (Kanehisa *et al.*, 2002).

Lorsqu'une voie métabolique telle le cycle de Krebs est activée, la cellule doit s'assurer que la dizaine d'enzymes qui catalysent les réactions successives sont présentes simultanément. Il est donc vraisemblable que les gènes correspondant soient co-régulés, et que leurs profiles d'expression révèlent l'activité de la voie métabolique sous-jaçantes.

Nous présentons une méthode pour vérifier cette hypothèse et fournir un outil permettant d'observer l'activité des voies métaboliques à partir de données de puces à ADN. Etant donnée une série de mesures d'expressions de gènes par puces à ADN, chaque enzyme est caractérisée par un profil d'expression. Parallèlement, chaque enzyme catalyse une ou plusieurs réactions dans les voies métaboliques connues. Ces voies métaboliques peuvent être représentées comme des chemins sur un graphe dont les noeuds sont les enzymes, et les arètes des composés chimiques; ainsi, deux enzymes sont liées si elles catalysent deux réactions chimiques dont le produit de l'une est le substrat de l'autre. La méthode que

nous proposons vise à détecter une forme de corrélation entre les enzymes vues comme des profiles d'expression d'une part, et des noeuds d'un graphe d'autre part. Typiquement, nous voudrions détecter le fait que les profiles d'expression des gènes qui catalysent une voie métabolique (et forment donc un chemin connexe sur le graphe) sont corrélés avec l'activité de la voie métabolique (elle-même inobservable directement).

Intuitivement, notre méthode fonctionne de la manière suivante. Pour un profil candidat censé représenter l'activité d'une voie métabolique, le coefficient de corrélation entre ce candidat et le profil d'expression de chaque gène est calculé. Les nombres obtenus (entre -1 et +1) sont ensuite projetés sur le graphe de gènes représentant les voies métaboliques, et la régularité de ces nombres par rapport à la structure du graphe est étudiée. Notre intuition étant qu'un bon profil candidat sera corrélé avec les profils d'expression de plusieurs gènes proches les uns des autres sur le graphe, la régularité de la fonction de corrélation sur le graphe via un calcul de sa transformée de Fourier discrète permet de quantifier si un profil est un bon candidat ou non. Via cette quantification, il est alors possible de rechercher les "meilleurs" profils candidats, et ensuite de vérifier dans quels régions du graphe ils semblent particulièrement corrélés avec les expression des gènes pour identifier les voies métaboliques concernées.

Plus techniquement, l'algorithme que nous présentons permet de directement calculer les meilleurs profils candidats. Cet algorithme repose sur le formalisme des méthodes à noyau (Schölkopf and Smola, 2002), grâce auquel on peut montrer que les meilleurs profils candidats peuvent être obtenus par une forme généralisée d'analyse de corrélation canonique entre les gènes projetés dans deux espaces de Hilbert auto-reproduisants (Bach and Jordan, 2002), définis par un noyau de diffusion sur le graphe (Kondor and Lafferty, 2002) d'une part, et un noyau linéaire sur les profiles d'expression d'autre part. Plus de détails sont disponibles dans (Vert and Kanehisa, 2003).

Nous illustrerons la méthode par une analyse de données publiques d'expression des gènes de la levure *S. Cerevisia* lors de cycles cellulaires.

## References


Bach, F., Jordan, M. (2002) Kernel independent component analysis, *Journal or Machile Learning Research*, 3, 1-48.

Kanehisa, M., Goto, S., Kawashima, S, Nakaya, A. (2002) The KEGG databases at GenomeNet, *Nucleic Acid Research*, 30, 42-46.

Kondor, R.I., Lafferty, J. (2002) Diffusion kernels on graphs and other discrete inputs, *Proceedings of ICML 2002.*

Schölkopf, B., and Smola, A. (2002) Learning with kernels, *MIT Press.*

Vert, J.-P., Kanehisa, M. (2003) "Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA", *Advances in Neural Information Processing Systems 15, MIT Press.*