Chapter 19

Résumé

L'amélioration de la technologie a vu la taille et la complexité des jeux de données augmenter considérablement dans de nombreuses disciplines scientifiques ces dernières années. Le résultat est qu'il est devenu très difficile de continuer à travailler avec des méthodes traditionnelles conçues pour des tailles de données plus restreintes. Les domaines concernés par cette inflation de données sont l'imagerie médicale, la physique, la science des matériaux, la télédétection...

L'analyse spectrale est l'une des branches principales du traitement du signal. De nombreux appareils expérimentaux produisent des signaux qui sont des sommes de sinus amorties (Résonance Magnétique Nucléaire, Imagerie par Réonance Magnétique, Spectrométrie de Masse à Résonance Cyclonique Ionique (ICR-MS)), sismomètres, etc. Mais avec les perfectionnements de ces appareils, le volume de données qu'ils produisent ne cesse de grandir. Le spectre d'un ICR-MS actuel contient typiquement 16 méga points de données et 10 000 sinus. Parmi les différentes méthodes utilisées pour analyser le spectre d'un signal discret, la plus standard est une transformée rapide de Fourier (FFT), qui permet de décomposer le signal en une somme de sinus, pour en trouver les harmoniques.

Dans cette thèse, nous nous concentrerons sur des données issues d'un spectromètre de masse à résonance cyclonique ionique (ICR-MS) sur la substance P ainsi que sur des données simulées. Dans les ICR-MS, l'analyse effectuée consiste à mesurer le rapport masse/charge de molécules ionisées, en se basant sur leur fréquence cyclotron obtenue en les faisant tourner dans un champ électromagnétique variable de plusieurs Teslas, pour trouver la composition du produit analysé. Comme dit précédemment, la technique habituellement utilisée pour l'analyse est la transformée de Fourier, bien qu'elle présente plusieurs limitations, car le bruit pose problème et le signal doit être apodisé pour gérer les oscillations induites par le démarrage et l'arrêt du signal, du fait de sa taille finie. De plus, l'appareil ne fournissant qu'une composante réelle au nombre complexe que nécessite une transformée de fourier, il est impossible de déterminer la phase des sinus trouvés.

Le principal problème est celui du bruit, dû aux imperfections des machines et qui diminue la performance d'une analyse mathématique parfaite comme une analyse de Fourier. De ce fait, les données brutes doivent être débruitées autant que possible avant de pouvoir être utilisées. En effet, une analyse mathématique (et donc déterministe et parfaite) d'un

signal bruité amènerait à une décomposition en un nombre de sinus infini, pour modéliser exactement le bruit. Pour gérer ce problème, de nombreuses approches existent, mais la taille des données produites est très grande (plusieurs millions de points) ce qui rend le débruitage coûteux en temps. En 2014, Chiron *et. al.* ont proposé un algorithme de débruitage de signaux harmoniques appelé urQRd (*uncoiled random QR denoising*) qui fonctionne bien plus rapidement qu'une approche classique SVD (*Single Value Decomposition*) et qui peut être appliquée aux très gros volumes de données produites par des ICR-MS ou autres spectroscopes à haute résolution, mais malgré la performance de l'algorithme, le temps de débruitage sur un signal comportant des millions de points se compte en semaines, sur des serveurs Xeon-Phi comportant un grand nombre de coeurs.

Le signal obtenu par ICR-MS est mesuré à intervalles temporels réguliers. La précision du spectre obtenu dépend de la méthode d'échantillonnage utilisée pour obtenir le signal. Un échantillonnage uniforme est utilisé dans des expériences uni-dimensionnelles. Cependant, dans le cas d'échantillonnages multi-dimensionnels, des méthodes non-uniformes sont préférables. Plusieurs méthodes Fourier et non-Fourier ont été imagniées pour gérer ces échantillonnages non-uniformes, mais la plupart de ces méthodes sont limitées par une importante complexité temporelle, des limitation sur la taille des échantillons, et une mauvaise résolution lorsque le rapport signal/bruit est faible.

Notre apport consistera à explorer l'apport de méthodes évolutionnaires (stratégies d'évolution, programmation génétique) pour dépasser les limitations des autres méthodes existantes dans l'état de l'art.

L'évolution artificielle se base sur la théorie darwinienne de l'évolution naturelle, qui postule que les individus les plus adaptés ont plus de chance de survivre dans leur environnement et de se reproduire pour créer une nouvelle génération d'individus. L'évolution artificielle et la programmation génétique produisent régulièrement des résultats compétitifs avec l'intelligence humaine sur de nombreux problèmes depuis le début des années 2000.

SINUS-IT ('algorithme principal développé dans le cadre de cette thèse) des valeurs réelles pour coder l'amplitude, la fréquence et la phase des sinus qui composent le signal, mais utilise une structure fonctionnelle d'algorithme génétique, car il utilise principalement un croisement monopoint et un opérateur de mutation.

La principale raison de l'utilisation d'un tel moteur évolutionnaire comparé à une stratégie d'évolution comme CMA-ES (Covariance Matrix Adaptation Evolutionary Strategy) dont les opérateurs sont conçus pour travailler sur des problèmes généraux, est que dans le cas de l'analyse harmonique, le problème est bien défini et connu, ce qui signifie qu'on peut concevoir des opérateurs de croisement et de mutation spécifiques, utilisant une compréhension profonde du problème, permettant un très bon compromis exploration / exploitation.

De plus, notre approche utilise un algorithme massivement parallèle qui tourne sur des cartes GPGPU, ce qui lui permet d'estimer directement les paramètres des sinus qui composent le signal sans utiliser de transformée de Fourier, ce qui apporte plusieurs avantages:

1. on peut directement trouver la phase (ce que ne peut trouver la transformée de Fourier dans le cas où (et c'est le cas pour l'ICR-MS) l'appareillage ne permet pas d'obtenir la partie imaginaire du nombre complexe nécessaire à la transformée de Fourier pour trouver la phase),

2. cette méthode n'étant pas déterministe, elle n'a pas besoin de de phase de débruitage

et peut travailler directement sur un signal brut,

3. aucune apodisation n'est nécessaire.

Nous avons effectué une étude comparative des méthodes par transformée de Fourier et par évolution artificielle. L'analyse se base sur des données de différentes tailles, artificielles ou réelles, avec un niveau de bruit de 100, et avec ou sans amortissement exponentiel.

Nous avons aussi utilisé différentes méthodes d'échantillonage, uniformes, non-uniformes et nous avons proposé une nouvelle méthode que nous avons appelée *Global Random Sampling* (GRS), qui consiste à utiliser de manière répétitive un petit échantillon jamais pris au même endroit, ce qui permet d'utiliser l'ensemble des points de données, mais jamais en une fois, pour diminuer le temps de calcul sans diminuer la qualité. Deux versions de GRS sont étudiées : un GRS par remplacement et un par extension.

Les résultats obtenus avec SINUS-IT sont meilleurs qu'avec une transformée de Fourier, sans pour autant nécessiter de débruitage, d'apodisation, avec en plus l'avantage d'obtenir la phase avec une bonne précision, le tout sur un échantillonnage non-uniforme, ce qui permet de diminuer le temps d'acquisition sur la machine ICR-MS dans le cas multi-dimensionnel, tout en utilisant seulement une petite portion du signal.

Les tests ont été effectués sur des pics isotopiques fins (3 sinus rapprochés d'1 part par million) ainsi que pour trouver les 6 ou 7 pics principaux de la substance P.

À titre d'exemple, nous trouvons avec notre méthode les 6 sinus principaux de la substance P avec seulement 8000 points, là où la méthode FT n'en détecte que 4 avec 64000 points. 512k points ne permettent toujours pas de trouver le 6è sinus par FT.

D'autres pistes ont aussi été explorées comme un algorithme évolutionnaire hybridé avec un algorithme à essaim particulaire inspiré de la physique quantique qui semble très prometteur.

Ce travail montre l'intérêt des approches évolutionnaires massivement parallèles pour l'analyse harmonique car elles ouvrent de nombreuses pistes de recherche, à la fois fondamentales et appliquées pour mieux exploiter les machines actuelles qui fabriquent un tel volume de données que leur limitation pratique ne sera bientôt plus physique, mais liée au temps de traitement informatique des données produites.