**Application de méthodes de recherche d'isomorphismes de graphes à la comparaison des structures de protéines**

**Introduction**

La connaissance des structures de protéines aide la recherche pour la création de médicaments et de nouvelles enzymes. Ces structures sont déterminées par cristallographie aux rayons X ou spectroscopie de résonnance magnétique, qui bien qu’efficaces, sont longues et coûteuses. Depuis 20 ans, la recherche informatique s’efforce de prédire ces structures et c’est dans la compétition CASP, que, tous les deux ans, la performance des méthode de détermination informatique est évaluée. Ils ont recours à un outil de mesure des similarités entre les structures des prédictions et celles déjà déterminées. C’est parce que la mesure de la précision du modèle informatique par rapport au modèle réel permet de faire avancer la recherche dans le domaine que je m’intéresserai à ce sujet.

**Professeur encadrant**

Mr Vallaeys, professeur de mathématiques au lycée Faidherbe

**Positionnement thématique**

INFORMATIQUE (Théorie des graphes), MATHEMATIQUES (Géométrie)

**Mots-clés**

Mots-clés(français) Mots-clés(anglais)

Isomorphisme/automorphisme de graphes Graph isomorphism/automorphism

Identification de sous-graphes isomorphes Subgraph isomorphism

Repliement des protéines Protein folding

Modélisation (des protéines) (Protein) Modeling

**Bibliographie commentée**

La comparaison de graphes est un problème d'informatique classique, qui, en plus d'avoir des utilités pratiques, s'avère être un problème théoriquement très intéressant du point de vue informatique. En effet, c'est l'un des seuls problèmes où l'on ne connaît ni algorithme de résolution en temps polynomial (pour le cas général), ni de preuve qu'il est NP-complet. Une généralisation est la recherche de sous-graphes isomorphes. Il a été prouvé qu'il n'existe pas d'algorithmes de résolution en temps polynomial pour celui-ci, il appartient donc aux problèmes de classe NP-complets.

Scott Fortin fait état de l'avancée du problème dans son papier [3], étudiant le travail accompli jusqu'ici, et citant quelques unes des ses applications. Par exemple, ces algorithmes sont utilisés pour classer les structures des molécules.

Il distingue deux approches : l'une est de directement chercher l'isomorphisme entre les graphes, l'autre est d'utiliser un intermédiaire, une fonction d'étiquetage canonique.

Pour la première, il peut être utile de combiner des invariants de sommets comme le degré, le type, la distance à un sommet donnée… László Babai le souligne dans son papier [6] où il montre un algorithme utilisant l’aléatoire mais aussi la classification des sommets par leurs invariants, on dit alors que les sommets sont « colorés ». L'avantage de cette première approche est que l'algorithme peut s'arrêter avant son terme si un isomorphisme a été trouvé mais il est difficilement applicable puisque le nombre de possibilités est très élevé et que, s'il faut déterminer que les graphes ne sont justement pas isomorphes, on parcourt toutes les possibilités.

Pour l'autre, la méthode est plus applicable, c'est d'ailleurs celle utilisée en pratique. Il faut déterminer une manière efficace et canonique d'étiqueter les graphes. C'est ainsi que McKay [4] propose un algorithme efficace "nauty", utilisant ce procédé. Il prouve que son étiquetage

est le même pour toute paire de graphes isomorphes. Il est ainsi le premier à l’époque, à résoudre le problème pour des graphes non-triviaux d'une centaine de sommets

Récemment, László Babai a proposé le meilleur algorithme répondant au problème général, de complexité quasi-polynomiale [5], ce qui correspond à une avancée majeure dans le domaine.

En ce qui concerne le problème des sous-graphes isomorphes, David Eppstein [1] trouve une méthode de résolution du problème d'isomorphisme de sous-graphes pour les graphes planaires en temps linéaire. Il se base sur une technique de partition du graphe planaire en morceaux de petite taille, une méthode de type "diviser pour régner". Finalement, ces concepts peuvent être appliqués à la mesure de similarité de deux graphes [2]. Ce qui aidera à mesurer la proximité de deux protéines.

**Problématique retenue**

Il s'agit de comparer les structures de protéines lorsqu'elles ont formé leurs liaisons et se sont formées dans l'espace, et déterminer leur proximité tant sur la forme que sur la position. Un outil de comparaison permettrait de mesurer la précision des prédictions informatiques d'un modèle, par rapport à un modèle référent dont la structure a été déterminée par une méthode longue et coûteuse (par rayons X par exemple).

**Objectifs**

Je me propose:

- d'étudier des modèles simples, une suite de points sans ramifications par exemple, et déterminer leur proximité par différentes méthodes

- de me familiariser avec la structure des protéines et le format pdb,

- de proposer des moyens de comparer des protéines (graphes), en commençant par écrire un algorithme naïf de détection d'isomorphismes de graphes, améliorer en utilisant des invariances entre les sommets.

- d'étudier la méthode McKay [4]

- d'adapter ces concepts à la recherche du plus grand sous-graphe isomorphe.

- d'établir un indice de comparaison, en fonction de la ressemblance de structure et de position des atomes dans l'espace

- d'en estimer l'utilité et de comparer avec les moyens actuellement mis en œuvre, notamment à la compétition CASP (Global Distance Test)

**Bibliographie**

[1] Subgraph Isomorphism in Planar Graphs and Related Problems, David Eppstein, 1999

[2] Measuring the similarity of labeled graphs, Pierre-Antoine Champin, Christine Solnon, 2003

[3] The Graph Isomorphism Problem, Scott Fortin, 1996

[4] B. McKay. Practical graph isomorphism. Congressus Numerantium, 1981

[5] Babai László, Graph Isomorphism in Quasipolynomial Time, 2016

[6] Babai László, Monte-Carlo algorithms in graph isomorphism testing, 1979