Algorithmique et optimisation combinatoire pour la bioinformatique
Toulouse – équipe SaAB

Contexte: la contribution de l'algorithmique et des techniques d'optimisation combinatoire à l'analyse de données de biologie moléculaire a augmenté de façon phénoménale depuis une dizaine d'années. La croissance des volumes de données de séquences, de génotypage... a nécessité le développement d'outils logiciels dédiés, basés sur des structures de données et des algorithmes de plus en plus sophistiqués et capables de faire face à des volumes de données inimaginables peu de temps avant. Le développement de tels outils nécessite de croiser des compétences pointues en algorithmique et en optimisation avec une bonne connaissance des problèmes biologiques.

Depuis plus de 10 ans, nous poursuivrons l'ambition de se maintenir sur le front de la recherche en algorithmique de l'optimisation tout en mettant à disposition des biologistes des outils logiciels capables de l'accompagner dans la résolution de problèmes génériques et de faire face à la volumétrie croissante des données à traiter. Cette ambition a porté ses fruits, nous profitons d'une bonne visibilité dans le domaine de l'algorithmique (en particulier dans le domaine des réseaux de contraintes pondérés) que du côté de nos collègues biologistes (dans l'INRA et au niveau international) qui utilisent nos outils logiciels pour traiter leurs jeux de données de grande taille.

Côté informatique, le laboratoire a fortement contribué au développement d'une famille de minorants devenue une sorte de référence, et adoptée et étendue par d'autres (problèmes de type MaxSAT pondérés par exemple, fonctions de coût globales). Ces travaux intéressent aussi les algorithmiciens des champs de Markov. Cette famille de minorants a été plus spécifiquement adaptée aux problèmes incluant des variables avec des grands domaines pouvant représenter des séquences génomiques et forme le cœur de l'outil de détection de gènes d'ARN de familles connues DARN!. Une seconde contribution, développée dans le cadre d'un projet ANR blanc, a aboutit à une famille d'algorithmes d'optimisation exacts exploitant la structure des problèmes, motivés par des problèmes de traitement de pedigree animaux de grande taille, cycliques mais avec des propriétés de structure permettant des gains exponentiels en complexité. Ces résultats sont intégrés dans l'outil logiciel toulbar2, qui accueille aussi des contributeurs internationaux. Cet outil a été décliné dans une variante permettant l'analyse de pedigree animaux (utilisé au Centre de Traitement de l'Information Génétique).

Plus indirectement, l'existence d'algorithmiciens de l'optimisation combinatoire à l'INRA permet d'injecter et d'adapter des technologies classiques dans le domaine pour résoudre très efficacement des problèmes génériques tels que la construction de cartes (génétiques ou physiques, dans le logiciel CarthaGène, qui inclut des algorithmes d'optimisation combinatoire inspiré de ceux développés pour résoudre le voyageur de commerce) ou la prédiction de gènes de protéines (EuGène inclut des algorithmes sur les graphes spécialisés pour le problème ainsi que des algorithmes d'optimisation stochastique pour l'estimation de paramètres).

L'arrivée des données de type NGS (Next Generation Sequencing), et plus généralement l'augmentation des volumes de données produits soulève de nouvelles questions. Elle nécessite de revisiter en profondeur les outils qui peuvent être dépassés par les volumes ou les types de données produits. C'est le cas pour le traitement de séquences mais aussi de données de génotypage, qui, loin d'être rendus obsolètes par les données de séquence, deviennent incontournables pour assister l'assemblage, parfois difficile, de ces données.

Motivations: le candidat sélectionné, compétent en algorithmique discrète et en optimisation combinatoire. Sera en charge de produire des méthodes originales dans le domaine de l'optimisation combinatoire permettant de maintenir notre visibilité côté méthodes mais également pour répondre aux problèmes de tailles croissantes soulevés par la biologie, Les principales questions soulevées par l'évolution des données sont d'abord liées à des soucis de taille qui génèrent des besoins en efficacité algorithmique. Les principales voies de recherche envisagées à l'heure actuelle dans le domaine des réseaux de contraintes pondérés consistent à mieux prendre en compte la spécificité des données traitées : renforcer la qualité de nos minorants, prendre en compte les symétries dans les problèmes, développer des algorithmes de filtrage dédiés (contraintes globales) pour des fonctions de coûts qui apparaissent dans les problèmes traités, étendre les méthodes de filtrage développées dans l'équipe aux modèles graphiques probabilistes discrets (très utilisés en bioinformatique pour la modélisation de données biologiques), combiner les algorithmes d'optimisation combinatoire et ceux sur la combinatoire sur les textes

Le chercheur recruté viendra renforcer et compléter les compétences existantes et sera mobilisé, pendant les premières années, sur un travail principalement méthodologique, bien qu'enrichi des questions finalisées que nous traitons en bioinformatique (sélection de marqueurs, haplotypage, cartographie, localisation de gènes d'ARN, annotation...). Il contribuera ainsi à augmenter notre masse critique qui reste insuffisante fasse aux besoins.

Compétences/origines: le chercheur pourra être issu d'un laboratoire de recherche opérationnelle, d'informatique ou d'intelligence artificielle, avec des compétences dans le domaine de l'algorithmique discrète, de la programmation par contraintes et de l'optimisation combinatoire. Des compétences supplémentaire autour du traitement algorithmique de modèles probabilistes discrets, de la programmation stochastique, de la biologie ou de la bioinformatique sont un atout supplémentaire.

Dans tous les cas, au delà de compétences méthodologiques marquées, la capacité à effectuer des développement informatiques de façon autonome et en équipe est importante. Bonne maîtrise de la langue anglaise, dynamisme, autonomie, capacité à travailler en équipe.

Partenariat: Du côté informatique, ces compétences assurent une insertion aisée tant au niveau local (avec les membres de l'équipe) qu'au niveau régional qui offre plusieurs pôles de compétences en programmation par contraintes -- avec lesquels nous collaborons (G. Verfaillie -CERT, H. Fargier, MC. Cooper, UPS) -- ou en recherche opérationnelle (P. Lopez, LAAS CNRS et N7). Au niveau national, avec les chercheurs en programmation par contraintes (dont C. Bessière - Montpellier, P. Boizumault Caen, P. Jégou, Marseille, pour ne citer que ceux qui ont été impliqués avec nous dans des propositions ANR). Au niveau international, il pourra collaborer avec les équipes avec qui nous interagissont : R. Dechter (UCLA, USA), P. Meseguer (CSIC, Barcelone), ou J. Lee (Univ. Hong Kong) par exemple.

En termes finalisés, le jeune chercheur recruté pourra bénéficier directement, pendant ses premières années, des compétences de l'équipe autour de la génétique, de la génomique (séquences protéiques, ARN) ainsi que du réseau de collaborateurs de l'équipe. Localement, au Laboratoire des Interaction Plantes-Microorganismes (SPE, sur la prédiction de gènes), au laboratoire de génétique Cellulaire (T. Faraut, B. Servin, haplotypage, sélection de marqueurs, analyse de textes) ou de la Station d'Amélioration Génétique des Animaux (A. Legarra, haplotypage).