UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 427 :
INFORMATIQUE PARIS SUD

Laboratoire : Laboratoire de Recherche en Informatique

Synthèse En Français

THÈSE DE DOCTORAT

INFORMATIQUE

par

Bo YANG

Analyses bioinformatiques et classements consensus pour les données biologiques à haut débit

Date de soutenance : 30/09/2014

Composition du jury :

Directeur de thèse :	Alain DENISE	Co-directeur de thèse
Co-directeur de thèse :	Xiang-Dong Fu	Co-directeur de thèse
Rapporteurs :	Daowen WANG	Examinateur
Examinateurs :	Sarah COHEN-BOULAKIA	Examinatrice
	Stéphane VIALETTE	Examinateur
	Min Wu	Examinateur

Résumé

Nous avons été dans une ère de données biologiques grande. Différents types de données est en pleine explosion, comme des séquences de gènes et génomes, la structure de l'ARN et des protéines, métaboliques et voies de signalisation, l'expression des gènes, de l'interaction et de la réglementation, et les maladies. En raison de l'océan de données, il est pensé pour être de plus en plus important de résoudre des questions biologiques en utilisant des approches de bioinformatique à l'ère post-génomique.

La bioinformatique est un domaine scientifique interdisciplinaire de sciences informatiques et de la biologie. En vue du calcul, il inclut des algorithmes, développement de logiciels et la construction de bases de données. Dans un autre point de vue, il est question biologiques entraînés, qui comprend une analyse de séquence pour l'alignement de séquence, le gène et la prédiction de promoteur, et la découverte de motif, l'analyse de la structure de l'ARN et de la structure des protéines prévision, la classification et la comparaison, et l'analyse de fonction.

Cette thèse porte sur l'analyse bioinformatique et algorithmes développement du classement de consensus pour les données à haut débit biologique.

Dans la biologie et de la génétique moléculaire, la plupart des gènes eucaryotes sont constitués de blocs de séquences séparées l'une de l'autre par des blocs de séquence de codage de codage non. Les régions de séquences codantes sont appelés exon et les séquences intermédiaires sont appelées introns. Épissage de l'ARN est la modification de la transcription pré-ARNm naissante dans laquelle les introns et les exons sont retirées sont reliés.

Pré-épissage de l'ARNm a lieu dans le mécanisme de l'ARN multi-composant connu sous le nom spliceosome, qui est assemblé d'une manière par étapes par l'ajout séquentiel de U1, U2, et de petites particules de ribonucléoprotéines nucléaires U4 / U6 / U5 de la pré- ARNm. U1 définit le site d'épissage en grande partie par des interactions d'appariement de bases, tandis que la 3 U2 reconnaît fonctionnelle »5 site

d'épissage fonctionnels, ce qui implique également un appariement de bases avec la séquence de ramification. Parce que le BPS (branche du site du point) est assez dégénéré dans les cellules eucaryotes supérieurs, l'ajout de U2 snRNP nécessite plusieurs facteurs auxiliaires, la plus importante étant l'hétérodimère U2AF composé d'un 65 kDa et 35 kDa sous-unité. De nombreuses expériences biochimiques sur modèle de pré-ARNm ont établi la séquence spécifique à la liaison de U2AF65 au tractus polypyrimidine (Py-voies) situé immédiatement en aval du BPS et le contact direct de U2AF35 avec le dinucléotide AG, qui définit ensemble fonctionnels des sites 3 'd'épissage. De définition de la 5 fonctionnel 'et 3' des sites d'épissage en U1 et U2 snRNP et suivant une série d'étapes dépendant de l'ATP, le complexe tri-snRNP U4 / U6 / U5 se joint à la pré-épissage initiale pour la transformer en la spliceosome matures.

Bien que le rôle vital de l'hétérodimère U2AF dans la définition 3 'des sites d'épissage a été largement apprécié, il a été difficile de savoir si cela est nécessaire pour la reconnaissance de tous les 3 fonctionnels »des sites d'épissage, en particulier dans les cellules de mammifères. Dans la levure bourgeonnante, Mud2 a été caractérisé comme l'orthologue de U2AF65, mais Mud2 est un gène non essentiel, probablement en raison de BPS très invariants dans cet organisme eucaryote inférieur. De même, dans la levure à fission, une fraction significative de gènes contenant des introns semblent manquer Py-voies typique, et en effet, plusieurs introns de U2AF indépendant ont été rapportés. Chez les mammifères, la présence de niveaux élevés de facteurs activateurs d'épissage, tels que les protéines SR, semble être capable de contourner l'exigence de U2AF pour initier ensemble de spliceosome. En outre, les génomes de mammifères codent également pour plusieurs gènes ayant des fonctions liées à la fois U2AF65 et U2AF35. Par conséquent, l'exigence fonctionnelle pour U2AF peut être contourné par de multiples mécanismes, ce qui soulève une question d'ordre général en ce qui concerne le degré de l'implication de la U2AF65 / 35 hétérodimère dans la définition du site d'épissage 3 'dans les génomes de mammifères. Cette question fondamentale est restée sans suite malgré la disponibilité des données

d'interaction U2AF65-ARN du génome entier.

Deuxièmement, la spécificité de liaison de l'ARN de U2AF65 a été bien caractérisé au niveau biochimique. Introns qui contiennent une forte Py-voies sont en mesure de supporter l'ensemble splicéosome de manière indépendante AG, et U2AF65 semble être suffisante pour soutenir l'épissage de ces introns AG indépendants, au moins in vitro. Cependant, la sous-unité U2AF35 est responsable de contacter directement le dinucléotide AG sur les sites 3 'd'épissage fonctionnels typiques et ce partenariat est renforcé par U2AF65 dépendant contrôle de la stabilité de U2AF35. Fonctionnant comme un hétérodimère, U2AF65 / 35 est pensé pour fournir une forte discrimination à l'égard riche en pyrimidine exonique ainsi que des séquences introniques qui ne font pas partie des sites 3 'd'épissage fonctionnels dans les génomes de mammifères. Des protéines de liaison d'ARN spécifiques, tels que DEK et hnRNP A1, ont été impliqués dans l'amélioration de la spécificité de liaison d'ARN dans les génomes de mammifères. Cependant, il reste à démontrer si directement l'hétérodimère U2AF lie en effet préférentiellement au Py-voies suivies par le dinucléotide AG de l'analyse du génome entier.

Troisièmement, outre le rôle essentiel de U2AF dans l'épissage constitutif, à la fois U2AF65 et U2AF35 ont été impliqués dans l'épissage réglementé. En théorie, les sites d'épissage alternatifs sont faibles en général, et par conséquent, la liaison sous-optimale peuvent les rendre particulièrement sensibles aux niveaux de U2AF, qui peut en outre être soumis à une telle PTB, TIA-1 / TIAR, et plus récemment, de hnRNP C. Bien que ces mécanismes semblent expliquer facilement U2AF dépendant inclusion de l'exon, il a été en grande partie inconnu pourquoi et comment l'épuisement de U2AF pourrait aussi induire un grand nombre d'événements d'inclusion d'exon in vivo. Engineered U2AF contraignant sur l'exon a été récemment montré pour inhiber l'inclusion de l'exon, mais il a été difficile de savoir comment largement ce mécanisme est utilisé pour réguler l'épissage alternatif de gènes endogènes.

Dernier point, mais non des moindres, de multiples mutations dans les deux

U2AF65 et U2AF35 ont été signalés à associer à une myélodysplasie (MDS) et des troubles sanguins connexes. Cependant, il est difficile de savoir comment ces mutations peuvent affecter le fonctionnement normal de U2AF dans l'épissage réglementé, qui souligne en outre l'importance de la compréhension mécaniste du rôle de réglementation de U2AF dans les cellules de mammifères.

Compte tenu de cette longue série de questions mécanistes qui restent à régler, nous avons entrepris l'analyse du génome entier des interactions U2AF-ARN dans le génome humain. En définissant le paysage génomique de U2AF liaison et l'exigence fonctionnelle à la fois U2AF65 et U2AF35 dans l'épissage réglementé, nous offrons une série de points de vue mécaniste dans la fonction de U2AF dans les états normaux et pathologiques.

Nous avons utilisé la technologie de séquençage à haut débit. Il est puissant et efficace. Les scientifiques biologiques développer un grand nombre de méthodes de base sur les technologies de séquençage à haut débit, pour obtenir un aperçu de l'expression et la régulation de molécule biologique à grande échelle.

Dans cette étude, mes collègues ont utilisé le séquençage d'immunoprécipitation de réticulation pour mapper l'interaction avec l'ARN de U2AF65 dans le génome humain. Il commence par la réticulation in-vivo des complexes ARN-protéine à l'aide de lumière UV. Ensuite, les cellules sont lysées. Et la protéine d'intérêt est isolé par immunoprécipitation. Après digestion protéinase, linker ligation, RT-PCR, nous avons pu le séquençage à haut débit la bibliothèque construite. En utilisant l'ARN-Seq nous avons pu obtenir tous les résultats de l'épissage alternatif dans une large échelle du génome. En comparant le changement de rapport de raccordement entre la cellule avant et après l'effet de choc de U2AF65, on peut illustrer la fonction de régulation dans U2AF65 épissage alternatif.

Obtenir les données de séquençage brut, nous avons pu vérifier la qualité de séquençage, cartographier le lit de génome. Pour CLIP-seq, la recherche indique que la reticulation peut induire une ou plusieurs délétion de nucléotide dans le lit au

niveau du site d'interaction ARN-protéine. Donc, nous avons également fait l'analyse CIMS. Pic but appelant à savoir site immobilier liaison avec lit beaucoup plus contraignante que noise.Distribution et motif nous aident à comprendre la structure et la fonction de liaison. Combinez les cas de régulation de l'épissage alternatif avec un profil de liaison U2AF65, nous pourrions déduire le mécanisme de régulation.

Conformément à la spécificité de liaison biochimiquement définies de U2AF, analyse de motif montré séquences hautement pyrimidiques enrichi sur les sites de liaison de U2AF65 mappé. Ces données démontré haute fidélité résultats de la cartographie d'interactions U2AF65-ARN dans le génome humain

Nous avons développé une approche de quartier maximale d'estimer le pourcentage de 3 sites 'd'épissage qui pourraient être liés directement par U2AF65. Nous avons d'abord trié tous les gènes en fonction de la densité de tag moyenne par site 3 'd'épissage dans chaque gène. et ensuite divisé ces gènes en groupes consécutifs, chacun constitué de 50 gènes. Cela nous a permis de calculer la couverture de annotées des sites 3 'd'épissage par U2AF65 avec une déviation standard dans tous les groupes. Nous avons ensuite déterminé le pourcentage de couverture de la 3 'des sites d'épissage lorsque la densité de la balise par site 3' d'épissage est augmentée progressivement. Nous avons observé que la couverture a atteint la saturation à ~ 88% avec l'augmentation des niveaux de liaison en 3 'des sites d'épissage U2AF65, ce qui indique l'existence de ~ 12% d'introns U2AF65 indépendante dans le génome humain. Nous avons ensuite demandé si ces U2AF65 non liés les sites 3 'd'épissage sont dérivé de consensus contraignant U2AF65. A cet effet, Nous avons ensuite calculé la moyenne 3 'site d'épissage score de U2AF65 non lié 3' des sites d'épissage. Nous avons détecté décès progressive de la moyenne 3 'site d'épissage score U2AF65 3'ss non lié. Ces données indiquent que, parmi les gènes qui montrent vaste U2AF contraignant, le manque de U2AF obligatoire dans les introns autres résultats probables d'une mauvaise consensus dans leurs sites 3 'd'épissage. Par conséquent, couplée avec l'analyse maximale de quartier, nos données suggèrent qu'une fraction significative (~12%) des fonctionnelles des sites 3 'd'épissage peut en effet représenter

les U2AF indépendants.

Après avoir appelé pic, nous avons trouvé une fraction significative de U2AF65 est capable de se lier dans l'intron en plus fonctionnels porcheries 3 'd'épissage. Nous voulons demander si il ya un autre site d'épissage dans l'intron autour des sites de liaison. Donc nous ségrégation U2AF65 événements de liaison sur les sites d'épissage non-3 'en deux classes. Le premier contient les exons de leurre potentiel ou pseudo exons, et l'autre n'a pas de preuve évidente pour tous les signaux d'épissage. Nous avons trouvé que la liaison U2AF65 à 3 sites fonctionnels »d'épissage sont fortement associés à une AG dinucléotide en aval; sa liaison près de leurres et pseudo exons montre moins, mais toujours significative, lien vers une AG en aval; et les événements dans d'autres endroits introniques liaison U2AF65 restant présentent aucun enrichissement sélectif avec un dinucléotide AG aval. Ces données suggèrent que, une fraction significative de U2AF65 est encore capable de se lier à d'autres endroits pré-ARNm outre fonctionnels porcheries 3 'd'épissage. Ces événements de liaison de U2AF65 peuvent interférer avec la définition fonctionnelle de l'os adjacent bonne 3 'des sites d'épissage en tant que mécanisme pour moduler la sélection des sites d'épissage alternatif ou refléter un rôle de U2AF65 dans d'autres étapes du métabolisme de l'ARN.

Grâce à la bio-informatique de l'ensemble du génome analyse des interactions U2AF-ARN et évènements d'épissage alternatifs régulés, nous déclarons que metagene et analyse de mini-gène suggère que des événements de liaison introniques amont interfèrent avec l'aval site immédiat d'épissage 3 'associée soit à l'exon alternatif pour provoquer le saut d'exon ou exon constitutif concurrence pour induire l'inclusion de l'exon alternatif.

Nous construisons de plus en place un système U2AF65 de notation pour la prédiction de sa base de sites cibles sur les données de séquençage à haut débit en utilisant une méthode d'apprentissage de la machine entropie maximale, et les scores sur les cas et réglementés bas sont compatibles avec notre modèle de régulation.

Ces résultats révèlent la fonction génomique et mécanisme de régulation de U2AF, ce qui nous facilite la compréhension de ces maladies associées.

Avec le développement croissant des technologies à haut débit, de très grandes quantités de données sont produites et stockées dans des bases de données publiques pour les rendre accessibles à la communauté scientifique, par exemple, Gene Expression Omnibus (GEO) qui est un établissement public fonctionnels à haut débit de génomique de séquençage données référentiel. Du biologique Big Data, de grandes quantités de gènes listes d'expression, la réglementation, l'interaction, la corrélation pourraient être extraites des résultats d'exploration de données, tels que la cellule exprime microARN, gènes gène régulé, interaction protéine-protéine, les gènes liés à la maladie, ou simplement gène association de text mining. Face à ces sortes de listes, il est très difficile de les exploiter si elles ne sont pas classés. Toutefois, le classement de données biologiques sur une même requête sont toujours très différent entre les différentes méthodes de traitement, des algorithmes ou des ensembles de données, en particulier pour les données biologiques pour la plupart avec le bruit, flou, biais et erreurs. Sur la base de toutes ces questions, la façon d'obtenir un résultat classement convincible de données biologique devient une tâche importante dans l'ère post-génomique.

Au lieu de développer de nouvelles méthodes de classement, Cohen Boulakia-et ses collègues ont proposé de générer un classement de consensus pour mettre en évidence les points communs d'un ensemble de classements tout en minimisant leurs désaccords pour lutter contre le bruit et l'erreur pour les données biologiques. Cette idée avait déjà été utilisé pour combiner les résultats de données de puces à ADN, microARN cible algorithmes de prédiction, site des méthodes de prédiction de liaison de ligand comparaison, et ainsi de suite.

Il a été aussi beaucoup d'intérêt pour ce problème dans la communauté de l'informatique au cours des dernières années qui se pose lors de la construction méta-moteurs de recherche pour la recherche Web, où l'on veut combiner les classements obtenus par différents algorithmes dans un classement représentant. Par

exemple, Dwork combine les classements des moteurs de recherche individuels pour obtenir un classement plus robustes qui ne sont pas sensibles aux diverses lacunes et les biais des différents moteurs de recherche (par exemple, "stage rémunéré" et "inclusion payante" parmi les moteurs de recherche).

Le processus de génération d'un classement de consensus est basé sur le concept d'agrégation classement, originaire de la théorie du choix social, l'apprentissage machine, et l'informatique théorique, définie sur le classement: Compte tenu m classements de n éléments et une fonction de distance, le problème d'agrégation classement est à trouver un classement de tous les éléments qui est le plus proche des m classements donnés.

Il pourrait facilement être pensé à une sorte de méthode d'agrégation classement, où l'ordre de chaque élément est déterminé en prenant la moyenne simple des positions de celui-ci à partir de différents classements. Cette méthode a d'abord été proposé par Borda comme un système de vote pour les élections à la fin du dix-huitième siècle. Condorcet a proposé une méthode plus raisonnable de vote à la majorité par paires connu comme critère de Condorcet, qui permet A à un rang plus élevé que B si la majorité vote pour A sur B en comparaison par paire, même si la moyenne des positions de A est après B.

Obéissant à Condorcet critère étendu, Kemeny proposé l'agrégation optimale Kemeny pour déterminer le meilleur classement global basé sur la distance Kendall-tau qui compte le nombre de désaccords entre les paires ordres d'éléments.

Cependant, Kemeny agrégation optimale est malheureusement un défi de calcul, parce que le problème est NP-difficile, même pour seulement quatre classements. Puisque le problème est important dans une variété de domaines, de nombreux chercheurs à travers ces champs ont convergé sur la recherche de bonnes pratiques, des algorithmes pour sa solution. Il ya des formulations qui conduisent à exiger des algorithmes, bien sûr sans polynômes en cours d'exécution des garanties de temps. Il ya aussi un grand nombre d'algorithmes heuristiques et approximation.

Comme décrit précédemment, Borda vient de la théorie du choix social. Il est la méthode "de position", qui trie les éléments par ordre décroissant en fonction de leur position moyenne dans tous les classements d'entrée. Evidemment, ceci est un algorithme heuristique, qui ne sont pas mis au point pour résoudre le problème de la médiane. Cependant, il pourrait donner une bonne solution très rapidement.

MEDRank a été conçu pour un environnement de base de données où, afin de fournir rapidement une réponse, il faut avoir aussi peu que possible les accès à chaque dossier de chaque classement. Afin de construire le consensus, tous les classements de R sont lus en parallèle, élément par élément. Ayant m classements et un seuil de tr, $0 < tr \leq 1$, dès qu'un élément a été lu $tr \times m$ classements, il est ajouté à la fin du consensus dans un nouveau seau. De toute évidence, l'algorithme fonctionne également en $O(nm)$. Dans l'étude de Fagin et ses collègues, le seuil de défaut considéré par les auteurs est $tr = 0.5$. De cette façon, l'algorithme est juste le tri à la valeur médiane de l'ensemble des positions de chaque élément dans le classement. Comme décrit ci-dessus, si cela est juste la solution optimale en fonction de la distance de footrule Spearman.

L'algorithme de Fagin et al. Est une sorte d'amélioration de la Rank MED, basé sur l'intuition que si deux éléments i et j ai rangs médians très proches, les points i et j doit être mis dans le même seau dans le classement de sortie. Donc, il est aussi appelé l'algorithme médian d'agrégation. Il commence à partir du résultat de la commande de rang médian d'éléments, puis des éléments de groupes avec près médiane se classe en même seau pour minimiser la somme de toutes les seaux coût basée sur la programmation dynamique. A chaque étape de la programmation dynamique de la solution est construite à partir de la meilleure des sous-solutions. La variante Fagin Grand choisit le sous-solution de premier rang rencontré au cours de Fagin Petit utilise la dernière. Leurs noms proviennent de ce expérimentalement, il a été remarqué que Fagin petite tendance à faire plus petite prise de Fagin Grand. Ils courent dans le temps $O(nm + n^2)$. Il a été prouvé que cet algorithme est un facteur

constant rapprochement des deux classements complets et classements partiels. Pour le classement complet, l'algorithme d'agrégation médiane donne un classement complet quasi-optimale, avec un facteur de rapprochement des deux.

BioConsert a été proposée par Cohen Boulakia-et ses collègues. Il fonctionne de manière itérative en essayant de déplacer un élément à un autre seau ou un nouveau godet à partir d'une entrée de classement pour réduire la somme de la distance de Kendall-tau qui améliore l'étape de classement par l'étape d'entrée. Si aucun des éléments sont modifiés de leurs seaux, alors l'algorithme se termine. Contrairement aux deux précédentes un, cette heuristique est un algorithme de tout temps, comme le classement d'entrée est itérativement amélioré et interrompre l'algorithme à tout moment se retourner un résultat correct. Cette heuristique peut être mis en oeuvre avec une complexité en temps $O(n^3m)$. BioConsert est une sorte d'algorithme de recherche locale. A chaque étape, l'algorithme de BioConsert est seulement à la recherche d'un meilleur voisin. Donc, il tombe dans une meilleure solution locale, ce qui peut être le meilleur mondial parfois.

Il ya quelques autres algorithmes pour le problème d'agrégation classement. Dwork et al. introduit un algorithme basé sur la chaîne de Markov. Qin et ses collègues ont développé un algorithme basé sur la posibilité. En outre, les tentatives de combinaisons de plusieurs algorithme pour donner un meilleur résultat ont également été signalés. Par exemple, la combinaison de KwikSort et Pick-A-Perm pourraient obtenir un algorithme d'approximation facteur 11/7, qui est un peu mieux que l'algorithme KwikSort (2-facteur algorithme d'approximation).

Sauf les algorithmes ci-dessus, un groupe d'algorithmes sont considérés comme des algorithmes de pivot nommés très potentiels. En commun, ils génèrent de manière récursive une solution en choisissant un des éléments que pivot et ordonnant à tous les autres éléments par rapport au pivot selon certains critères. Il divise le problème en plus petits et conquiert séparément, et utilise la propriété transitive qui est juste dans la plupart des situations, en particulier pour le classement avec bon accord. Ainsi, les

algorithmes de pivot sont toujours rapide dans le temps et pas mal de précision.

Ailon, Charikar et Newman a développé le premier algorithme de pivot. Il a été nommé KwikSort, principalement parce que l'algorithme ressemble à un type de algorithme de tri. KwikSort algorithme génère récursivement une solution en choisissant un aléa comme "pivot" et tous les autres éléments de la commande par rapport à l'élément de pivot. De cette manière, la relation de position entre les éléments dans les ensembles de part et d'autre de l'articulation ne doit pas être pris en compte: l'ensemble des éléments sur le côté gauche sont avant tout de l'élément sur le côté droit. Bien que l'algorithme de KwikSort utilise la propriété transitive qui est généralement vrai pour les éléments, mais pas prend les conflits dans les classements en compte. Ainsi, certains plusieurs algorithmes ont été développés pour tenter de résoudre ce problème, en changeant la méthode d'affectation ou méthode pivot de prélèvement.

La programmation linéaire en nombres entiers (ILP) pour le problème de l'agrégation classement est également NP-difficile. Mais comme nous le savons, la programmation linéaire (LP) détente sans contrainte de l'intégralité peut être résolu en temps polynomial. Basé sur le pivot et le système de programmation linéaire, Ailon et ses collègues ont proposé un autre algorithme, LP-KwikSort. L'idée principale de l'algorithme est en train de changer l'affectation des autres éléments de façon à ce que, après nous choisissons un pivot j, nous devrions utiliser la valeur de la solution LP (p_{ij} et p_{ji}) pour décider où mettre tous les autres éléments, au lieu de décider avidement. Ailon et ses collègues ont montré que cet algorithme est un algorithme 4/3 approximation pour les classements sans liens, qui est meilleur que l'algorithme KwikSort.

Un autre algorithme de pivot modifié appelé DerandLP-Pivot, a été proposé par Van Zuylen et collègues. Il est un algorithme de pivot déterministe, au lieu de pivot choisissant au hasard, et cet algorithme donne directement sur les conflits dans les classements. Dans chaque appel récursif, nous choisissons le pivot k qui minimise le

rapport coût-gagnant. De cette manière, le choix de k coûts aussi peu que possible, et génère le plus possible. En outre, cet algorithme améliore également la méthode d'affectation de tous les autres éléments sur la base de la solution de programmation linéaire sans contrainte intégralité.

Ici, nous proposons un nouvel algorithme de pivot, appelé Conformément à pivot. Elle est basée sur un nouveau procédé de récolte de pivotement et l'attribution de tous les autres éléments. Nous pensons que cet algorithme est plus approprié pour la propriété transitive des données de problème d'agrégation classement.

Tout d'abord, nous définissons un score cohérente pour l'élément i comme la somme des coûts entre l'élément et tous les autres éléments. Ce score reflète la certitude de position de l'élément dans le classement. L'élément avec une plus petite partition uniforme est plus stable. Comme un repère bien connu dans la ville pour les autres bâtiments, les relations de position sont claires, l'élément avec la plus petite partition cohérente pourrait aussi être un marqueur de positionner tous les autres éléments. Avec l'intuition ci-dessus, nous proposons que l'élément avec la plus petite partition cohérente devrait être choisi comme le pivot.

Continuer à utiliser le principe, nous attribuons tous les autres éléments pas au hasard, mais dans un ordre de la partition cohérente de petite à grande. En ce qui concerne la méthode d'affectation de tous les autres éléments, nous utilisons pas directement la relation de position entre l'élément et le pivot, au lieu d'utiliser une fonction de coût que la position ayant la plus petite est choisie coût.

Tant la table de poids (W) et meilleure table de relations de position (X) entre deux éléments peuvent être calculés simultanément dans un temps de $O(n^2m)$. Les processus de tri des éléments, le choix d'un pivot et l'affectation de toutes les autres sont beaucoup plus rapides. Ainsi, la complexité du temps de cet algorithme est $O(n^2m)$, la même que KwikSort, et plus vite que DerandLP-pivot.

Les expériences montrent que l'algorithme Conformément à pivot est un

algorithme efficace pour les données réelles à la fois en termes de précision et de temps de fonctionnement. Il est beaucoup plus rapide que la BioConsert, LP-KwikSort, algorithmes DerandLP-pivot, et effectue presque aussi bien que le BioConsert pour des données réelles.

Tous les algorithmes ont des avantages avec des lacunes. La pensée de la combinaison de plusieurs algorithmes pour améliorer les performances est une bonne idée. Le Conformément à pivot suivie par la meilleure recherche de l'algorithme BioConsert dans une gamme locale, peut-être un bon algorithme combiné pour le problème d'agrégation classement.